OLAF HÜBLER
JOACHIM FROHN

Modern Econometric Analysis

Surveys
on Recent
Developments

Springer

Modern Econometric Analysis

Surveys on Recent Developments

Olaf Hübler · Joachim Frohn
(Editors)

Modern Econometric Analysis

Surveys on Recent Developments

With 8 Figures and 11 Tables

Springer

Professor Dr. Olaf Hübler
University of Hannover
Empirical Economic Research
and Econometrics
Königsworther Platz 1
30167 Hannover
Germany
huebler@ewifo.uni-hannover.de

Professor Dr. Joachim Frohn
University of Bielefeld
Universitätsstraße 25
33615 Bielefeld
Germany
jfrohn@wiwi.uni-bielefeld.de

First published in
Allgemeines Statistisches Archiv, Vol. 90, Issue 1, 2006

ISBN-10 3-540-32692-8 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-32692-2 Springer Berlin Heidelberg New York

Cataloging-in-Publication Data
Library of Congress Control Number: 2006923354

Hardcover-Design: Erich Kirchner, Heidelberg

SPIN 11680970 42/3153-5 4 3 2 1 0 – Printed on acid-free paper

Preface

The importance of empirical economics and econometric methods has greatly increased during the last 20 years due to the availability of better data and the improved performance of computers. In an information-driven society such as ours we need quickly to obtain complete and convincing statistical results. This is only possible if the appropriate econometric methods are applied.

Traditional econometric analysis concentrates on classical methods which are far from suitable for handling actual economic problems. They can only be used as a starting point for students to learn basic econometrics and as a reference point for more advanced methods. Modern Econometrics tries to develop new approaches from an economic perspective. A consequence is that we have less of a unified econometric theory than in former times. Specific branches which require specific methods have been established. Nowadays, nobody has complete knowledge of every area of econometrics. If someone is interested to learn more about a field, relatively unknown to them, they will require support.

This volume is intended to be both an introduction to new econometric methods and a reference for professional econometricians and graduate students. The book provides concise surveys have been carefully selected of many relevant new developments in econometrics. Nevertheless, the contributions are selective. We have asked leading German econometricians to write papers about their specific research fields with emphasis on their own area of interest. Nearly all of them have responded with enthusiasm and we are very pleased with the outcome. Hopefully, the reader will share this positive impression. The papers cover the methods used in simultaneous equation models over modern time series, the use of duration and panel data analysis, microeconometrics, and specific data problems. The reader will find discussion of the benefits and pitfalls as well as the statistical properties of the methods presented. The authors emphasize the advantages and disadvantages. They outline the progress which has been made is made in the last few years and the problems which remain unsolved. It is not our intention to demonstrate the workability of the methods by specific applications. For this purpose the reader should consult other publications. Nevertheless, in some chapters advices can be found. All of the contributions in this book are also published in the Journal of the German Statistical Society (Allgemeines Statistisches Archiv), Volume 90 (2006), Issue 1.

We would like to thank all the authors who have produced a superb series of papers and the referees who have contributed to a lot of improvements. The support of the German Statistical Society, particularly of Karl Mosler, the president of the society, is gratefully acknowledged. Last but not least we thank the Springer Pub-

lishers and especially Werner A. Müller for their cooperation, and their prompt and professional editorial assistance.

Olaf Hübler, Hannover
Joachim Frohn, Bielefeld March 2006

Contents

1 Developments and New Dimensions in Econometrics

Olaf Hübler[1] and Joachim Frohn[2]

[1] Institute of Empirical Economic Research, University of Hannover
huebler@mbox.iqw.uni-hannover.de
[2] Faculty of Economics, University of Bielefeld
jfrohn@wiwi.uni-bielefeld.de

Summary. This book presents 14 papers with surveys on the development and new topics in econometrics. The articles aim to demonstrate how German econometricians see the discipline from their specific view. They briefly describe the main strands and emphasize some recent methods.

1.1 Introduction

75 years ago on the 30th of December in 1930 the Econometric Society was founded and two years later the Society has decided to establish its own journal 'Econometrica'. This was the birth of econometrics, but the roots lie in the mathematical and statistical economics of the nineteenth century. The 1930s were determined by the enthusiasm of an international group of young people. National style and theoretical allegiance seemed to matter less than their common methodological programme. Ragnar Frisch (1970, p. 152), the first editor of Econometrica, describes this feeling when he talks about the First European Meeting of the Econometric Society in 1931 in Lausanne: 'We, the Lausanne people, were indeed so enthusiastic all of us about the new venture, and so eager to give and take, that we had hardly time to eat when we sat together at lunch or at dinner with all our notes floating around on the table to the despair of the waiters'. The first years can definitely be characterized by mathematical models and methods to describe macroeconomic problems and a lack of suitable data sets. Since this time we have observed a rapid and sustainable development. Was it a success story or a chain of frustrations? Both is true.

On the one hand the enthusiasm has vanished, the belief that large econometric models can completely explain the economic development has given way to a more skeptical view. The strategy of the Cowles Commission was criticized and alterna-

tive approaches were presented as 'general to specific modelling' by Hendry (1980), 'sensitivity analysis' by Leamer (1985) or 'reduced form vector autoregressive models' by Sims (1980). Forecasts are not always satisfactory. Many researchers believe that it is not really possible to estimate structural models. Summers (1991), for example, is convinced that applied econometrics is ineffective, not more than explorative data analysis can be done by econometrics. Nevertheless, many econometricians still have the objective to estimate structural models. Some of the following contributions demonstrate the progress in this field, especially the first and the fifth paper.

On the other hand the data situation has significantly improved. Not only are aggregated yearly time series available, but quarterly, monthly, weekly and even daily or tick-by-tick data can be used. Furthermore, large cross-sectional, panel, duration and multilevel data are the basis of many applied econometric studies. Computers give us the chance to process such large information. This was and is a good motivation to develop estimation methods and testing procedures for these specific data constellations.

In particular progress in time series analysis has been made, since this field is considered in the perspective of econometric regression models. Nowadays, unit roots and co-integration investigations are standard in single equations models and vector autoregressive models. This means non-stationary in contrast to traditional time series are feasible; the bridge is made between economic theory of equilibria and the econometricians, whose models concentrated on the short-run dynamics, and poor performance of simultaneous macroeconometric models has improved. Recent developments tackle the problem for structural multivariate models, but seasonal models are also considered and structural shifts are in discussion. Econometricians also try to benefit from traditional multivariate statistical methods, like factor analysis and combine these with modern time series methods.

While in the 1930's and over a period of 40 years macroeconometric models with many equations have dominated our discipline, in the 1970's the situation changed. The beginning of this new phase is characterized by the new interest in microeconomic questions and the availability of large cross-sectional data sets, but in applied econometrics the old methods were still used. A little later a revolution of new approaches started in many fields and is still continuing. First, the range of estimation methods extends enormously. Besides the traditional least squares, maximum likelihood methods and methods of moments we find several further methods in the literature: instrumental variable methods, generalized methods of moments, quasi maximum likelihood, quantile estimators, extremum estimator, generalized estimating equations, simulation based estimators. Non- and semi-parametric procedures are developed. Second, researchers pay more attention to testing. Especially the spectrum of specifications tests is extended. Incorrect omission of regressors, inclusion of other variables that are proxies for the omitted variables and a false functional form of the econometric relationship are leading sources of biased estimations. Third, nonlinear approaches become more relevant. Specific microeconomic data lead to binary, multinomial and ordered probit or logit, count data, duration and tobit models. All of them have a specific nonlinear character and the problem of unobserved heterogeneity has to be solved. Fourth, econometricians analyze specific data problems in more detail. These include measurement errors, anonymization of

data and missing values. Fifth, evaluation methods of policy intervention and the measurement of treatment effects are more important in economically difficult periods. In Germany, this phase started some years after the German reunification. The main problem is to compare two situations where only one is observable.

1.2 Contributions

This book contains 14 contributions. Most of the new developments are more or less intensively reviewed.

The first paper, presented by Pu Chen and Joachim Frohn, is in some sense a rehabilitation of large scale structural macroeconometric models. The authors emphasize the relevance for empirical economics. Apart from traditional aspects like identification and estimation of simultaneous approaches the paper integrates non-stationary variables and co-integration. In detail, statistical inference in large systems is discussed. The paper argues that until now the modern statistical inference is not fully integrated into the simultaneous structural approach. Chen and Frohn describe the state of the art of large scale simultaneous econometric models, identify the major unsolved problems and suggest a combined data-theory strategy to look for the specification of the models.

The second paper also focusses on empirical macroeconomic models. Jörg Breitung and Sandra Eickmeier discuss large dimensional dynamic factor models which have recently become popular in this field, especially in macroeconomic policy analysis. The basic idea is to reduce the dimensionality. The large number of available macroeconomic variables is combined with factors that should be interpretable. The paper gives a short survey on factor models and considers different procedures for determining the number of factors. The main new aspect is the introduction of dynamic factor models and procedures to estimate the innovations (error terms) of the factors where VAR models are used. Furthermore, the authors present an overview of economic applications and add their own empirical contribution to co-movements in Europe.

Jürgen Wolters and Uwe Hassler address problems of unit root testing, which nowadays is the starting point of most empirical studies of time series. First, the most popular approach, the Dickey-Fuller and the augmented Dickey-Fuller test including the distribution of the test statistics is presented. Additionally, the authors describe the selection of the lag length, the treatment of the deterministic part in the augmented Dickey-Fuller regression and problems with the variation of time span and frequency of observations. Most important is the overview to unit root testing under structural shifts. Consequences of ignoring breaks, suggestions of corrections under situations with breaks, smooth and multiple breaks are the content of this section.

Co-integration in single equation models within an autoregressive distributed lag framework is the topic of the next contribution. Uwe Hassler and Jürgen Wolters start their paper with different representations of autoregressive distributed lag model including the error correction model. Inference to a co-integration vector and co-integration testing follows using error correction models. Several approaches and

asymptotic propositions are discussed. A Monte Carlo study based on a generated bivariate process demonstrates finite sample properties. The main simulation results are the following: In most cases the t-type co-integration test is just as powerful as the F-type one. Co-integration tests for conditional error correction models are more powerful than those for unconditional ones.

Helmut Lütkepohl's article gives a review of the structural vector autoregressive models for co-integrated variables. This contribution extends the analysis of the last one and is related to the first and the second paper. VAR models which explicitly take into account the co-integration structure of the variables are considered. Impulses, innovations and shocks are shown in a response function. The estimation of vector error corrections models and of the impulse response function are at the centre of the article. The coefficients of the reduced form and those of the structural form are determined. The author does not only consider just identified but also overidentified models. Identifying restrictions are then also required for the impulses and their responses. Co-integration properties can help to determine these restrictions.

While the first five contributions concentrate on models with aggregated time series, the rest emphasizes approaches and problems of microeconometric models, although some aspects are also relevant for macroeconometric models. We have to distinguish between several types of data, specific data problems and some practical issues. Apart from cross section data, applied econometrics uses multilevel and panel data, duration data and high frequency data. For each type of data specific problems exist and therefore the literature has developed separated models and estimation methods. The next three chapters consider approaches for panel, duration and high frequency data.

Some aspects of the frequency of observations were already discussed by Wolters and Hassler. However, high frequency or tick-by-tick data are only now available and recently the relevant methods were developed. A special issue of the Journal of the German Statistical Society (Vol. 86, 2002) has discussed these problems in more detail where applications of financial transactions were at the centre of the analysis. Now Helmut Herwartz presents a new survey. He focusses his article on new methods of strongly disaggregated time series where the time span of new observations is not equidistant. Apart from market macrostructure modelling, high frequency data have recently attracted large interest in econometrics as a mean to estimate conditional volatility. As in Lütkepohl's contribution vector error correction models are used. Furthermore, Herwartz considers parameter estimation with incomplete samples and realized volatilities.

Duration analysis is widely used in social sciences, labor and health economics. Several surveys and text books exist on this topic. It is not the intention of Bernd Fitzenberger and Ralf Wilke's paper to add a further one in this general context. Instead, they restrict their considerations to the quantile regressions for exploring the distribution of duration data. Recently, the interest in linear and nonlinear quantile regressions has increased enormously, as these methods are robust and flexible. The contribution compares quantile regression to standard duration models, considers the issue of right censoring and the estimation of hazard rates in quantile models. A brief application with German data on unemployment duration for younger workers

demonstrates the workability of the methods. The authors show that the conventional proportional hazard model is rejected empirically as the estimated quantile regression coefficients change sign across the quantiles.

Meanwhile panel data analysis is standard in applied econometrics of micro data. Most applications are restricted to linear models where pooled, fixed or random effects estimators are employed. Olaf Hübler focusses his contribution on multilevel and nonlinear panel data models where the latter are separated between specific and unspecific nonlinearities. He starts with parametric linear models under alternative error term structures. Multilevel approaches follow. The parametric nonlinear panel data analysis is based on typical microeconometric models and concentrates on logit and probit. Until now, non- and semi-parametric models are not so widespread in applications and several open problems exist. The paper fills some missing links. The presentation emphasizes fixed effects models where new estimation methods are necessary.

Göran Kauermann presents a more general survey on non- and semi-parametric models and their estimation. He discusses smoothing procedures including the Nadaraya-Watson approach and spline smoothing. Within non- and semi-parametric models the author focusses on generalized additive and varying coefficient models. Starting with Hastie and Tibshirani's suggestions the paper extends the standard additive models by combining them with McCullagh and Nelder's ideas of generalized linear models. Some further approaches and model diagnostics supplement the presentation. As an application, data of more than a thousand new founded firms are used to determine the success rate. The study compares the results of parametric and nonparametric estimates.

Microeconometric models are the subject of Gerd Ronning's paper. Apart from a short review to the principles of probit, logit, count data, duration and tobit models, the special emphasis is laid on problems to make micro data anonymous in these models. This aspect is of special interest as many micro data sets are not publicly available due to confidentiality. Recently, researchers have developed methods to anonymize these data in order to minimize the risk of disclosure and to save statistical properties of the original data. Good progress is made for quantitative variables while for qualitative and censored data the problems are not completely solved. The paper discusses the case of post randomization for binary variables in a probit model.

Another field, the ordered response models, is widely neglected in microeconometrics. Commonly, continuous and binary or multinomial endogenous variables are considered. In their article Stefan Boes and Rainer Winkelmann give a brief survey of several strands and review standard ordered response models. They then discuss more general approaches. The authors relax three assumptions of the standard models: single index, constant threshold and distributional assumptions. The extension includes random coefficients, finite mixture and sequential models. Furthermore, they illustrate the methods by an analysis of the relationship between income and happiness using data from the German Socio-Economic Panel. The study shows that the results of standard and generalized ordered models differ substantially.

Hans Schneeweiss and Thomas Augustin's paper is concerned with recent advances in measurement error methods. Of course, measurement errors are not restricted to

microeconomic data, but have more practical relevance for this type of data. Three elements are constitutional for measurement error models: the true unobservable regression model, the measurement model and the distribution of the latent regressor. The main problem is the estimation apart from the identification of the coefficient of the latent variable. The authors provide a systematic treatment of this topic. Naive, corrected score, and simulation-extrapolation estimators are the subject of this section. In addition, structural estimation methods are presented, which use information given in the distribution of latent variable. Finally, the authors report on efficiency comparisons, extend the methods to survival analysis and discuss misclassification problems of categorical variables.

In the last 20 years the development of new evaluation methods has made enormous progress. The analysis concentrates on microeconometric estimates of treatment effects. Marco Caliendo and Reinhard Hujer give an overview of these developments. They present the most important estimation methods, identifying assumptions and compare the different approaches. The paper is restricted to non-experimental evaluation methods and distinguishes between matching regression methods which use observed information for selection. For unobservable variables the authors split their consideration into difference-in-differences, instrumental variables and selection estimators. Apart from these standard methods, the paper also presents dynamic evaluation concepts including sequential matching procedures, considerations to duration models and matching with time-varying treatment indicators.

A further data problem in applied econometrics is item non-response. What are the consequences and what can we do in this situation? Susanne Rässler and Regina Riphahn review the literature on this topic and demonstrate the prevalence of item non-response in the German Socio-Economic Panel. They report on determinants and effects. At the centre of the article is the way with which item non-response can be dealt. The authors present four approaches: casewise deletion of observations, weighting, imputation and model-based procedures. Emphasis is on the imputation techniques where multiple imputation is allowed. Simulation studies illustrate the implications of alternative imputation procedures. In the conclusion applied researchers find recommendations to handle problems with item non-response.

References

FRISCH, R. (1970). Econometrics in the world today. In *Induction, Growth and Trade: Essays in Honour of Sir Roy Harrod,* (W. A. Eltis, M. F. G. Scott, J. N. Wolfe, eds.), Clarendon Press, Oxford.

HENDRY, D. F. (1980). Econometrics – Alchemy or Science. *Economica* **47** 387-406.

LEAMER, E. E. (1985). Sensitivity analyses would help. *American Economic Review* **75** 308-313.

SIMS, C. A. (1980). Macroeconomics and reality. *Econometrica* **48** 1-47.

SUMMERS, L. H. (1991). The scientific illusion in empirical macroeconomics. *Scandinavian Journal of Economics* **93** 129-148.

2 On the Specification and Estimation of Large Scale Simultaneous Structural Models

Pu Chen[1] and Joachim Frohn[2]

[1] Faculty of Economics, University of Bielefeld
pchen@wiwi.uni-bielefeld.de
[2] Faculty of Economics, University of Bielefeld
jfrohn@wiwi.uni-bielefeld.de

Summary. This paper surveys the state of the art of the analysis and application of large scale structural simultaneous econometric models (SSEM). First, the importance of such models in empirical economics and especially for economic policy analysis is emphasized. We then focus on the methodological issues in the application of these models like questions about identification, nonstationarity of variables, adequate estimation of the parameters, and the inclusion of identities.

In the light of the latest development in econometrics, we identify the main unsolved problems in this area, recommend a combined data-theory-driven procedure for the specification of such models, and give suggestions how one could overcome some of the indicated problems.

2.1 Introduction

Simultaneity and structure are two key concepts in econometrics that help the econometricians to look at a statistical model from an economic point of view and to go beyond the analysis of statistical parameters to estimating and analyzing economic structural relations.

Simultaneous structural econometric models (SSEMs) - the implementation of these concepts - became the most important research object of econometrics in its early

time. The Cowles Commission methodology became the main paradigm of empirical research in macroeconomics. Simultaneous structural models, especially large scale simultaneous structural models, became very popular in the 1960s and the 1970s. During the 1970s, however, the scientific community became sceptic against these models due to its methodological deficits. Lucas (1976) questioned the constancy of parameters of SSEM in case of policy changes and thus the suitability of the SSEM for policy simulation. Naylor *et al.* (1972) and Nelson (1972) doubted the predictive ability of SSEMs in comparison with alternative simple time series models. Granger and Newbold (1976) pointed out the improper treatment of the time series properties in SSEMs. Sims (1980) and Sims (1982) finally criticized the 'incredible' identification restrictions in the specification of SSEMs[1]. Since then simultaneous structural models stepped back from the forefront of econometric research.

Despite of this serious and justified criticism towards simultaneous structural models, large scale SSEMs are still widely used among practitioners and policy consultants. In Germany, for instance, large scale econometric models were applied in the late 1990s at Deutsche Bundesbank, RWI, DIW, HWWA, Ifo-Institute and IWH, and some models are still applied today. For more information about these models see 'www.macromodels.de'.

Why are SSEMs still in use? There are several reasons, and all have to do with the 'economic' appeal of these models: (1) Policy makers often want to find out the impact of certain policies in different areas of economic life. An SSEM seems to be capable of taking into account various related aspects of the policy, while its scientific competitors VAR and VECM are much too small to answer the question asked. (2) Furthermore, *structural* models are more revealing of the manner in which an economy is operating[2], contrary to *reduced* form models.

The main purpose of this paper is to survey relevant new developments in econometrics concerning the methodology of SSEMs and to identify the leftover problems.

Applying an approach by Spanos (1990) (see also Chen, 2001)[3] we reinterpret simultaneous structural equations as an economic theory motivated representation of a general statistical model that describes the probability law of the observed data. In this way we provide a coherent framework for integrating the concept of statistical adequacy in the concept of simultaneity and structural modeling.

The paper is organized as follows: In Section 2 we summarize the main methodological deficits of the simultaneous structural approach from the perspective of statistical inadequacy. In Section 3 we survey the relevant new developments in econometrics and address the open problems[4]. Then we provide a general statistical framework to encompass simultaneous structural models. We also discuss the issues of statistical inference in simultaneous structural models. In Section 4 we conclude with an outlook for further research.

[1] See Granger (1990) for an overview of the methodological debate.
[2] See Dhrymes (1993) for more discussions.
[3] Similar ideas can be found in Hendry and Mizon (1990) and Hendry and Mizon (1993).
[4] See Frohn (1999) for more discussion.

2.2 SSEMs - the State of the Art

2.2.1 Modeling Procedures of SSEMs

An SSEM is denoted as

$$B'y_t + \Gamma' x_t = \epsilon_t, \qquad [\epsilon_t | X_t] \sim NI(0, \Sigma), \qquad t \in T, \tag{2.1}$$

where y_t and x_t are vectors of jointly dependent and predetermined variables, respectively. The parameter matrices B' (g×g) and Γ' (g×k) are subject to usual a priori restrictions which represent the structural information based on economic considerations as well as behavioral hypotheses on the economy.

The Cowles Commission methodology that is applied to construct such SSEMS consists mainly of the following steps[5]:

- The dichotomy between endogenous and exogenous variables is decided beforehand based on the purpose of the modeling and economic reasoning without referencing to statistical properties of the data.
- The specification of regression equations is carried out equation by equation by using the implication of economic theoretical hypotheses and/or economic intuitions and by combining these with some statistical tests (typically t-test and Durbin-Watson-test).
- The estimated model is subject to extensive simulations for possible modifications. The simulation model consists of the behavior equations (2.1) and identities.

Most structural simultaneous macroeconometric models contain long persistent variables such as price, wage rate, GDP, interest or other I(1) variables. In some models error-correction mechanisms are implemented on an ad hoc basis, i. e. the rank of cointegration is not determined statistically and also the specific introduction of error terms in an equation is ad hoc. In some models there is no explicit treatment of I(1) variables, even if the residuals show significant autocorrelations. There is rarely an explicit documentation of the estimation of the unconstrained reduced form. Many of these models would fail to pass the overidentification test (see for instance Blanchard, 2000). We will now discuss these aspects in the next subsection.

2.2.2 Statistical Adequacy of SSEMs

Spanos (1990) studies the problem of SSEMs and calls it the problem of statistical adequacy of SSEMs. He suggests to free the reduced form by interpreting it as a statistical model in the context of which the structural form can be considered. In particular he proposes to start the modeling of an SSEM from the unconstrained reduced form

$$y_t = \Pi' x_t + u_t, \qquad [u_t | X_t] \sim NI(0, \Omega), \qquad t \in T, \tag{2.2}$$

[5] See Charemza (1997) for an interesting summary of the Cowles Commission methodology.

with the following underlying assumptions:

1. The density $D(y_t|X_t)$ is normal.
2. $E(y_t|X_t) = \Pi' x_t$ is linear in x_t.
3. $Cov(y_t|X_t = x_t) = E(u_t u_t'|X_t = x_t) = \Omega$ is homoscedastic.
4. The statistical parameters of interest: $\theta = (\Pi, \Omega)$, where $\Pi = \Sigma_{22}^{-1}\Sigma_{21}$ and $\Omega = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$, are time invariant (with $\Sigma_{11} = cov(y_t)$, $\Sigma_{12} = cov(y_t, X_t)$ and $\Sigma_{22} = cov(X_t)$).
5. $(y_1, y_2, ...y_T)$ is an independent sample sequentially drawn from $D(y_t|X_t; \theta), t = 1, 2, ..., T$.

If the sample information indicates a departure from the assumption of temporal independence of the residuals, Spanos (1990) suggests to use a dynamic unconstrained reduced form as the general statistical model:

$$y_t = \Pi_0' x_t + \sum_{i=1}^{L} [\Pi_{1i}' y_{t-i} + \Pi_{2i}' x_{t-i}] + u_t, \quad [u_t|X_t, Y_-, X_-] \sim NI(0, \Omega). \tag{2.3}$$

This approach can be used as basis for a concrete modeling procedure to come to a statistically adequate SSEM[6].

2.2.3 Statistical Adequacy with Respect to the Criticisms Towards SSEMs

Applying this approach to SSEMs the structural information according to economic considerations is no longer a priori for statistical inference. It becomes a set of statistically testable hypotheses, namely as restrictions on the parameters of the unconstrained reduced form[7]. Now we focus on the question, how this new approach to SSEMs can help to cope with the above mentioned criticisms towards the traditional Cowles Commission methodology.

First of all Sims' (1980) critique on the 'incredible' identification conditions has, obviously, no relevance to this approach to SSEMs: Structural restrictions are no longer a priori. By contrast they will be tested within the unconstrained reduced form. If tests do not reject these restrictions, their credibility is guaranteed by the sample information. Furthermore, after the test for statistical adequacy, the SSEM will encompass the conditional VAR parsimoniously[8]. Moreover, it helps to solve the problem of too many insignificant parameters in unconstrained VARs and makes the statistical inference more effectively, giving a more precise description of the data.

As far as the forecast performance of SSEMs is concerned, (2.3) certainly encompasses every univariate ARMA model of the relevant variables, which is already

[6] see Chen (2001) for details.
[7] For detailed discussion see Chen (2002).
[8] For details see Dhaene (1997) and Chen and Hsiao (2004c).

pointed out in Zellner (1979) and Zellner and Palm (1974). Also, the concerns of Granger and Newbold (1976) can be taken into account in the determination of (2.2). Time series aspects of the sample information can be integrated in the specification of (2.3). The problem of nonstationarity of relevant variables will be considered later in this paper.

Lucas' critique (1976) concerns the invariance of the parameters with respect to policy changes. In this context an SSEM is immune against Lucas' critique if the structural parameters are super-exogenous[9] with respect to the parameters of the marginal process on which the SSEM is conditioned. Statistical adequacy is only a necessary condition for immunity against Lucas' critique but not sufficient. However, the super-exogeneity can be formulated as a property of the data generating process (DGP); hence sample information can be used to test the immunity against Lucas' critique. In this sense immunity against Lucas' critique forms a further statistical hypothesis on the properties of the DGP. At this point it is clear that whether an SSEM can be used for policy simulations depends crucially on the behavior of the economy from which we sample the data, but not on the 'philosophy' that was used to construct the model.

2.2.4 Limits of Statistical Adequacy

In studying the statistical adequacy of a structural model, we examine the possible misspecification in the framework of a general statistical model. Given that we cannot find a most general statistical model, we have to make a decision how general we would like our statistical model to be. Spanos (1990) uses the Haavelmo distribution family as the general statistical model under which the statistical adequacy is investigated. To take the nonstationarity of variables into account, we will use unconstrained VARs as the most general statistical model.

2.3 Statistical Adequacy of SSEMs with I(1) Variables

2.3.1 A Classification of SSEMs

In order to identify the problems of SSEMs concerning the statistical adequacy, we first classify the most often used SSEMs according to the following five basic features of their constituting elements: the functional form of the equations, i. e. whether they are linear or nonlinear; the stationarity of the endogenous and the exogenous variables, i. e. whether they are I(0) or I(1) variables; the distribution of the disturbance, i. e. whether they are normal or not; and the presence of identities.

If we take only those models into account that have linear behavioral equations with normal disturbances, we get the following six different types of models[10]. In Table

[9] See Engle *et al.* (1983) for the definition of super exogeneity.

[10] Models with I(1) exogenous variables and I(0) endogenous variable are rarely used in practice. Therefore they are not considered here.

1 we now list the most relevant statistical problems for these six models and potential solutions offered in the literature. The unsolved problems are symbolized by '?'.

Table 2.1: **Solved and Open Problems in SSEMs**

Typ	End.-/Exog.Var.	Identities	Identification	Estimation	Simulation	Notice
I	I(0)/I(0)	No	Cl[a]	Cl	Cl	
II	I(0)/I(0)	Yes	Brown[b]	Malinvaud[c]	Cl	
III	I(1)/I(0)	No	Johansen	Johansen[d]	Hendry[e]	SSM[f]
IV	I(1)/I(0)	Yes	?	?	?	
V	I(1)/I(1)	No	Hsiao[g]	Hsiao[h]	Stock[i]	SSM
VI	I(1),I(0)/ I(1),I(0)	Yes	?	?	?	SSM

[a]Cl symbolizes the classic textbooks of econometrics, such as Theil (1971), Judge *et al.* (1985), and Schmidt (1976)
[b]See Brown (1983) and Brown (1985).
[c]See Malinvaud (1980) and Chow (1983) for details.
[d]See Johansen (1995).
[e]See Hendry (1995).
[f]SSM refers to small scale models.
[g]See Hsiao (1997), Breitung (1995), Bierens (1997).
[h]See Hsiao (1997), Habro *et al.* (1998), and Johansen (1992) for details.
[i]See Stock and Watson (1988)

As for model Type I, i. e. models without identities and no consideration of nonstationarity of variables, the classic textbooks on econometrics, such as Theil (1971), Greene (1993), Judge *et al.* (1985), and Amemiya (1985) present the standard treatment. As we stated in the introduction the most practically relevant SSEMs are not of this type.

For models of Type II, Malinvaud (1980) and Chow (1983) provide a treatment of SSEMs with linear identities. Brown (1983) and Brown (1985) consider the identification problems of SSEMs with nonlinear identities. Hausman (1975), Chen and Hsiao (2004a) and Chen *et al.* (2005) consider the issues concerning the influence of identities on the estimation of SSEMs.

The modern time series econometrics textbooks such as Johansen (1995), Hamilton (1994) and Hendry (1995) deal with models of Type III, where all I(1) variables are taken as endogenous, and only deterministic variables like a constant or a trend are considered exogenous. Besides an intensive treatment of the methodological issues of empirical research, Hendry (1995) emphasizes the use of VECM to conduct a dynamic empirical analysis. Johansen (1992), Habro *et al.* (1998) and Johansen (1995) consider conditional processes of a cointegrated system as a specific structural model, where the structural information is formulated as the restrictions on the parameters in the cointegration space and on the parameters of the short run dynamics, respectively.

Several solutions for diverse special cases in Table 2 exist in the literature. Engle *et al.* (1983) deal with the concept of exogeneity in econometric models in general and for SSEMs in particular. For models of Type V, Hsiao (1997) and Breitung (1999) consider the problem of identification and estimation of parameters. Hsiao

(1997) emphasizes the difference between the role of a priori information for the SSEM approach on one side and the role of availability of data for the VAR and VECM approach on the other, where he implicitly assumes weak exogeneity of the exogenous variables. Breitung (1995) shows that there always exists an SSEM representation for a VECM, and that the traditional estimation methods used for stationary SSEMs, such as 2SLS and 3SLS, are still valid. As far as we know, however, there is not yet an econometric solution for the most relevant cases, i.e. the SSEMs of Type VI, especially for large scale SSEMs of this type.

The analysis of the statistical adequacy of SSEMs of Type I and Type II can be done in two steps: first application of misspecification tests of the unconstrained reduced form and then the overidentification tests. Models of Type III are usually specified as vector error correction models. Hence they are already formulated as a general statistical model. The issue of statistical adequacy is to conduct misspecification tests of the unconstrained VAR. Because this type of models is constructed using a data driven approach from the very beginning, the more relevant issue is here to work out the structural information contained in these VECMs.

Models of Type V are generally a difficult statistical problem, because the most often used test statistics are nonstandard and depend on nuisance parameters that describe the marginal process of the exogenous variables[11]. This problem may be solved by using nonparametric techniques[12]. The issues of identities, especially the nonlinear identities, in models of Type IV to Type VI are not yet thoroughly discussed in the literature. One problem with nonlinear identities is that they may contradict the general linear structure among the variables which is an essential assumption of the linear models discussed here. Furthermore they may be in conflict with assumptions on the disturbance.

As far as the statistical analysis of large scale systems is concerned, two specific problems will normally arise. Although the Johansen procedure provides a standard procedure to run a multivariate cointegration analysis, it is, unfortunately, only applicable in small systems due to its data intensive specification. The determination of cointegration rank and the estimation of cointegration relations for large scale systems is still an unsolved problem in econometrics.

A further problem concerning the scale of the system is the valid reduction of the number of free parameters in the system in case of undersampling. A general data-driven procedure to reduce the dimension of a model is not yet available, even not for small scale models[13].

2.3.2 SSEMs with I(1) Variables

The Basic Idea. An SSEM will be a sound description of the economic phenomenon, if it integrates, on one hand, the sample information of the data and an intuitive interpretation of the model structure on the other. This implies that the SSEM under investigation must parsimoniously encompass the general statistical

[11] See Habro *et al.* (1998) for details.
[12] See Chen and Hsiao (2004b), Choi (2003) and Bierens (1997).
[13] See Hendry and Krolzig (2001) and Winker and Maringer (2004).

model that can be used to summarize the sample information[14]. In this context the statistical adequacy of an SSEM with I(1) variables can be tested in the framework of a general statistical model.

The General Statistical Model. For a concrete economic policy issue an economist may be interested in modeling the dependence of one group of variables y_t on another group of variables x_t as well as on the past values of all these variables: Z_{t-1}[15]. The dimensions of y_t and x_t are G_y and G_x respectively.

We call $(y_t', x_t')'$ the variables of primary interest. Their components may be I(1) and I(0). We denote the I(1)- and the I(0)- components by subindexes 1 and 2 respectively: $(y_t', x_t')' = (y_{1t}', y_{2t}', x_{1t}', x_{2t}')'$. To give a general representation of the DGP for $(y_t', x_t')'$, we transform the variables of primary interest $(y_t', x_t')'$ by summing up the I(0) components to I(1) variables z_t which are now called relevant variables z_t. In this way we define the relation between the variables of primary interest and the relevant variables as follows:

$$z_t = \begin{pmatrix} z_{yt} \\ z_{xt} \end{pmatrix} = \begin{pmatrix} y_{1t} \\ \sum_{\tau=1}^{t} y_{2\tau} \\ x_{1t} \\ \sum_{\tau=1}^{t} x_{2\tau} \end{pmatrix} = \begin{pmatrix} y_{1t} \\ \tilde{y}_{2t} \\ x_{1t} \\ \tilde{x}_{2t} \end{pmatrix}. \tag{2.4}$$

Generally we assume that the relevant variable z_t has a VAR(p) presentation:

Assumption 3.1: *The relevant $n \times 1$ I(1)-vector z_t has a VAR representation*

$$z_t = A_1 z_{t-1} + A_2 z_{t-2} + ... + A_p z_{t-p} + \epsilon_t \tag{2.5}$$

with ϵ_t a white noise process.

Because the essential relations among I(1) variables are the cointegration relations, we reformulate (2.5) in vector error correction form according to Granger's representation theorem[16]:

$$\Delta z_t = \Pi^{(0)} z_{t-1} + \Pi^{(1)} \Delta z_{t-1} + ... + \Pi^{(p-1)} \Delta z_{t-p+1} + \epsilon_t\,, \tag{2.6}$$

where $\Pi^{(0)} = \beta\alpha'$. β and α are $n \times h$ matrices with $h < n$ and $rank(\beta\alpha') = h$ which is the cointegration rank.

$$E(\epsilon\epsilon') = \begin{pmatrix} \Sigma_{yy} & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{xx} \end{pmatrix}.$$

[14] For the definition of encompassing see Dhaene (1997).
[15] $Z_{t-1} = (y_{t-1}', x_{t-1}', ...y_{t-p}', x_{t-p}')'$.
[16] See Hamilton (1994, p. 582) for details.

The existence of h cointegration relations implies restrictions on the VAR(p) parameters in (2.5):

$$rank(\sum_{i=1}^{P} A_i - I) = h. \tag{2.7}$$

Structural and Reduced Forms of SSEMs. An SSEM motivated by economic theory has the following structural form (for simplicity of the presentation but without loss of generality we limit the lag-length to 3):

$$By_t+\Gamma_1 x_t+\Gamma_2 y_{t-1}+\Gamma_3 x_{t-1}+\Gamma_4 y_{t-2}+\Gamma_5 x_{t-2}+\Gamma_6 y_{t-3}+\Gamma_7 x_{t-3} = u_t\,. \tag{2.8}$$

The structural relations are formulated in the following 'a priori' identification restrictions[17]:

$$rank((B,\Gamma_1,\Gamma_2,\Gamma_3,\Gamma_4,\Gamma_5,\Gamma_6,\Gamma_7)\Psi_i) = G_y \qquad \text{for } i = 1,2,...,G_y \tag{2.9}$$

where Ψ_i is a known matrix.

Solving (2.8) for y_t, we get:

$$\begin{aligned} y_t \;=\; & -B^{-1}\Gamma_1 x_t - B^{-1}\Gamma_2 y_{t-1} - B^{-1}\Gamma_3 x_{t-1} - B^{-1}\Gamma_4 y_{t-2} \\ & -B^{-1}\Gamma_5 x_{t-2} - B^{-1}\Gamma_6 y_{t-3} - B^{-1}\Gamma_7 x_{t-3} + B^{-1}u_t\,. \end{aligned} \tag{2.10}$$

The reduced form of the SSEM is

$$\begin{aligned} y_t = \Pi_{y1}x_t + \Pi_{y2}y_{t-1} + \Pi_{y3}x_{t-1} + \Pi_{y4}y_{t-2} + \Pi_{y5}x_{t-2} \\ +\Pi_{y6}y_{t-3} + \Pi_{y7}x_{t-3} + B^{-1}u_t\,. \end{aligned} \tag{2.11}$$

Now we identify the conditions under which the general statistical model (2.6) has an SSEM representation like (2.8) and (2.9). We call (2.8) the SFSSEM(y,x) presentation, (2.10) the DRSSEM(y,x) representation, and (2.11) the RFSSEM(y,x) representation respectively[18].

[17] See Judge *et al.* (1985, p. 577) for details of identification.

[18] SFSSEM, DRSSEM and RFSSEM refer to structural form, derived reduced form and reduced form of the SSEM, respectively.

The Condition for a Valid SSEM Representation. An SFSSEM(y,x) representation like (2.8) with (2.9) is a valid representation of the DGP (2.6), if the restrictions on the parameter of the DGP can be reformulated as (2.8) and (2.9)[19]. Because SFSSMM(x,y) describes only the conditional process of y_t given $x_t, z_{t-1}, ...z_{t-p}$, we factorize (2.6) accordingly $\begin{pmatrix} I_{G_y} & C \\ 0 & I_{G_x} \end{pmatrix}$:

$$\Delta z_{yt} = -C\Delta z_{xt} + (I, C)\beta\alpha' z_{t-1} + (I, C)\Pi^{(1)}\Delta z_{t-1} + ... \\ +(I, C)\Pi^{(p-1)}\Delta z_{t-p+1} + \epsilon_{y|x}\,, \tag{2.12}$$

$$\Delta z_{xt} = (0, I)\beta\alpha' z_{t-1} + (0, I)\Pi^{(1)}\Delta z_{t-1} + ... \\ +(0, I)\Pi^{(p-1)}\Delta z_{t-p+1} + \epsilon_x \tag{2.13}$$

with $C = -\Sigma_{yx}\Sigma_{xx}^{-1}$. The condition for an efficient statistical inference on the parameters of the conditional process without consulting the total process is that x_t should be weakly exogenous with respect to the parameters of the conditional process[20].

When cointegration relations are present in the system, according to Johansen (1992) the necessary and sufficient condition for weak exogeneity of x_t with respect to the parameter of the conditional process is

$$\beta_x = 0, \tag{2.14}$$

where $(\beta_y, \beta_x) = \beta$ is the partition of β according to the corresponding variables of y_t and x_t. Under this condition the conditional process can be reformulated:

$$\begin{aligned} \Delta z_{xt} &= (0, I)\begin{pmatrix} \beta_y \\ 0 \end{pmatrix}\alpha' z_{t-1} + (0, I)\Pi^{(1)}\Delta z_{t-1} + ... \\ &\quad +(0, I)\Pi^{(p-1)}\Delta z_{t-p+1} + \epsilon_x \\ &= (0, I)\Pi^{(1)}\Delta z_{t-1} + ... + (0, I)\Pi^{(p-1)}\Delta z_{t-p+1} + \epsilon_x\,. \end{aligned} \tag{2.15}$$

Because α does not appear in the marginal process, the modeling of the marginal process (2.13) will provide no information for the inference on the parameters of (2.12). Under the assumption of weak exogeneity of x_t the restriction on the DGP due to cointegration is

[19] More precisely: An SSEM(x,y) representation is valid, only if it represents the conditional process of the DGP, and if the conditioning variables are weakly exogenous for the parameters of the conditional process. See Hendry (1995, p. 350).

[20] See Engle *et al.* (1983) for detailed discussion.

$$rank(\beta_y \alpha') = h\,. \tag{2.16}$$

To get the SFSSEM(y,x) we transform (2.12) in terms of y_t und x_t:

$$\begin{pmatrix} y_{1t} - y_{1t-1} \\ y_{21} \end{pmatrix} = -C \begin{pmatrix} x_{1t} - x_{1t-1} \\ x_{2t} \end{pmatrix} + \beta_y \alpha' \begin{pmatrix} y_{1t-1} \\ \tilde{y}_{2t-1} \\ x_{1t-1} \\ \tilde{x}_{2t-1} \end{pmatrix}$$

$$+\Pi^*_{1y} \begin{pmatrix} y_{1t-1} - y_{1t-2} \\ y_{2t-1} \\ x_{1t-1} - x_{1t-2} \\ x_{2t-1} \end{pmatrix} \quad +\Pi^*_{2y} \begin{pmatrix} y_{1t-2} - y_{1t-3} \\ y_{2t-2} \\ x_{1t-2} - x_{1t-3} \\ x_{2t-2} \end{pmatrix} + \epsilon_{y|x}\,, \tag{2.17}$$

where $\Pi^*_{1y} = (I, C)\Pi^{(1)}$ and $\Pi^*_{2y} = (I, C)\Pi^{(2)}$.

Because $\tilde{y}_{2t-1} = \sum_{\tau=1}^{t-1} y_{2\tau}$ and $\tilde{x}_{2t-1} = \sum_{\tau=1}^{t-1} x_{2\tau}$ do not appear in the SFSSMM(y,x), a cointegration system with $\tilde{y}_{2t-1}$ and $\tilde{x}_{2t-1}$ will not have a SFSSMM(y,x) representation. Therefore a necessary condition for (2.6) to have an SFSSMM(y,x) representation is

$$\alpha' = (\alpha'_{y1}, 0, \alpha'_{x1}, 0)\,. \tag{2.18}$$

The cointegration restriction is now

$$rank(\beta_y(\alpha'_{y1}, \alpha'_{x1})) = h\,. \tag{2.19}$$

Under the conditions in (2.14) and (2.18) the DGP (2.6) has a Cointegration Constrained Conditional Process representation (CCCP(y,x)).

Because there are usually some exclusion conditions on some lag components in the unconstrained reduced form of an SSEM - RFSSEM(y,x), the CCCP(y,x) representation can be taken as RFSSEM(y,x) only if these exclusion restrictions are satisfied. If this is the case, the DGP has a cointegration constrained reduced form representation (CCRFSSEM(y,x)).

It is generally known that[21] an overidentified SSEM places some rank conditions on the unconstrained reduced form. Therefore the DGP (2.6) will have a Constrained Structural Form representation (CCSFSSEM(y,x)) only if the parameters of the DGP satisfy these overidentification restrictions.

Summing up the discussion above, a cointegration system (2.6) will have a CCSFSSEM(y,x) if the following conditions are satisfied:

- z_{xt} is weakly exogenous with respect to the parameters in (2.12).

[21] See e. g. Frohn (1995) and Schmidt (1976).

- The I(1)-components in SSEM (y_{1t}, x_{1t}) span the cointegration space of the DGP.
- The DGP satisfies the exclusion restrictions.
- The DGP satisfies the overidentification restrictions, if there is any.

It is worth to notice that the presence of cointegration relations places restrictions on the conditional process and henceforth also on the SFSSEM(y,x) representation. These restrictions make sure that the SSEM represents a cointegration system. They become unbinding if the dimension of y_t equals the dimension of the cointegration space $G_y = h$. This is due to the fact that under $G_y = h$ the condition (2.16) will be satisfied for almost all estimated values.

In addition, we would like to note that the set of conditions listed above is the condition under which a cointegration system such as (2.6) with h cointegration relations has an (overidentified) CCSFSSEM(y,x) representation (2.8) and (2.9). The main concern here is to find out whether the DGP has this SSEM representation as suggested by the relevant economic theory, but not to find 'a' SSEM representation for the DGP[22].

2.3.3 The Role of Economic Theory in SSEMs

An SSEM is parsimonious if it places a large number of restrictions on the general representation of the DGP, such as exclusion and overidentification restrictions. Economic theory and behavioral hypotheses can be used to formulate such restrictions. They provide an intuitive economic interpretation for the SSEM. It should be stressed that contrary to the text book simultaneous equations approach the structural information is not the starting point of the statistical analysis, but just a statistical hypothesis which will be tested based on sample information. The structural information will not become the structure of the SSEM unless it is justified by the sample information via some statistical tests.

2.4 Statistical Inference of Large Scale SSEMs

The target of the statistical inference of large scale SSEMs is to investigate the statistical adequacy of SSEMs in question. It focuses on the tests of the four conditions for a valid CCSFSSEM(y,x) representation and the misspecification of CCSFSSEM(y,x) against a possible departure from the basic assumptions of the general DGP.

Because Johansen procedure is not applicable for large systems, we would like to propose the following strategy to cope with this problem:

The basic idea follows the two step strategy of Engle and Granger (1987): In Step 1 we try to identify the cointegration rank and the cointegration relations of the

[22] According to Breitung (1995) a VECM has always a SSEM presentation.

large scale system by subsampling. In Step 2 we test the four conditions for the valid CCSFSSEM(y,x) representation. At the end we carry out misspecification tests based on the CCRFSSEM(y,x).

Testing the Cointegration Rank in Large Systems. In the literature there are now some parametric and nonparametric procedures that can be used to test the cointegration rank in large systems. Applying the subsampling approach[23], we are able to determine the cointegration rank and the cointegration relations, i. e. we can get estimate $\hat{\alpha}$. Our investigation[24] can indicate that this approach - contrary to the Johansen procedure - can deal with large scale SSEMs (up to about 50 equations).

2.4.1 Test of Exclusion Restrictions

Here we test the restrictions for a valid CCSFSSEM(y,x) representation. We substitute the estimated $\hat{\alpha}' z_{t-1}$ by w_{t-1} and rewrite (2.6) as follows:

$$\Delta z_t = \beta w_{t-1} + \Pi^{(1)} \Delta z_{t-1} + ... + \Pi^{(p)} \Delta z_{t-p+1} + \epsilon_t \,. \tag{2.20}$$

Because the I(1) variable z_{t-1} is replaced by the stationary variable w_{t-1}, (2.20) contains only stationary variables. Hence we could apply the conventional methods to (2.20) if we had enough data.

The way out of this data-dilemma can only be found in the data themselves. It is well known that empirically unconstrained VAR-Models are characterized by large numbers of insignificant parameters. This means that the unconstrained VAR models are overparameterized for economic data, which implies that the number of free parameters of the data generating VAR is significantly lower. In other words: the unconstrained VAR parameters are subject to a large number of restrictions. If we knew these restrictions we could estimate and test the VAR with less data.

Now the task is to formulate the 'correct' restrictions. A necessary and sufficient condition for restrictions to be 'correct' is that under these restrictions the residuals ϵ_t are white noise. This property can be used to test the correctness of the exclusion restrictions, i. e. we test whether under the hypothetic restriction the estimated residuals are still white noise. This can be done in the system or equation by equation [25].

We denote the VAR parameters under exclusion restrictions by $\Pi^{(i)*}$, $i = 1, 2, ..., p$, and we have a VECM under the exclusion restrictions:

$$\Delta z_t = \beta w_{t-1} + \Pi^{(1)*} \Delta z_{t-1} + ... + \Pi^{(p-1)*} \Delta z_{t-p+1} + \epsilon_t \,. \tag{2.21}$$

Now a test of white noise can be applied to the residuals in (2.21) to verify the correctness of the exclusion restrictions.

[23] See Chen and Hsiao (2004b) for details.
[24] See Chen and Hsiao (2004b) for details.
[25] For the determination of the lag length for VAR Model (2.6) we apply the same logic.

2.4.2 Test of Sufficient Cointegration

After we have identified the correct exclusion restrictions a reduced rank regression procedure can be applied under these restrictions. To find out whether $\tilde{x}_{2t}$ and $\tilde{y}_{2t}$ enter the cointegration space we can apply the standard likelihood ratio test according to Johansen (1995).

Test of Weak Exogeneity. To test weak exogeneity of x_t we can run an F-test in (2.22) for the hypothesis: $H_0 : \beta_x = 0$ v.s. $H_1 : \beta_x \neq 0$:

$$\begin{pmatrix} \Delta z_{yt} \\ \Delta z_{xt} \end{pmatrix} = \begin{pmatrix} \beta_y \\ \beta_x \end{pmatrix} w_{t-1} + \Pi^{(1)*} \Delta z_{t-1} + ... \Pi^{(p)*} \Delta z_{t-p+1} + \epsilon_t \,. \quad (2.22)$$

If the null is not rejected, we say that the CCRFSSEM(y,x) representation is justified.

2.4.3 Test of Overidentification

After these tests the CCRFSSEM(y,x) can be formulated as follows:

$$\Delta z_{yt} = -C \Delta z_{xt} + \beta_y^* w_{t-1}^* + \Pi^{*(1)} \Delta z_{t-1} + \Pi^{*(2)} \Delta z_{t-2} + \epsilon_{y|x} \,. \quad (2.23)$$

The parameters in the CCRFSSEM(y,x) representation satisfy already the cointegration restrictions. It is now to test whether the overidentification restrictions implied by the structural form are valid. An LR test can be applied for this.

2.4.4 An Integrated Modeling Procedure

We can summarize the foregoing discussion in the following modeling procedure:

1. Description of the economic phenomenon in question and formulation of the target of modeling.
2. Determination of the variables of primary interest (y_t, x_t) and the relevant variables (z_t).
3. Study of the time series properties of the variables y_{1t}, x_{1t}, y_{2t} and x_{2t}.
4. Formulation of a general statistical model (2.6).
5. Estimation of the cointegration rank h by using subsampling procedures and determination of the error term.
6. Formulation of the economic theoretical hypotheses and determination of the conditioning variable x_t.
7. Estimation of (2.22) with OLS and testing the residuals for white noise properties. If this hypothesis is rejected, go back to Step 6 and reconsider the economic hypothesis.
8. Other misspecification tests of the CCRFSSEM(y,x).
9. Test of overidentification.
10. Summary of all restrictions on the general statistical model.
11. Estimation of (2.8) subject to the restrictions (2.9).

2.5 Concluding Remarks

In this paper we emphasize the statistical adequacy of SSEMs with respect to general properties of the observed economic data. We summarize the critiques on the traditional Cowles Commission methodology as a critique of not implementing the principle of statistical adequacy in this methodology. A multi-step testing procedure is developed to check the statistical adequacy of an SSEM.

Principally, a more general statistical model can usually accommodate more properties of data and henceforth is more adequate to describe a set of data. On the other hand, a more general model will be less efficient in conducting statistical inference, because potential a priori information is not used. A more parsimonious model may be more efficient, if it incorporates a priori information.

A useful empirical model has to meet both the requirement of statistical adequacy and the requirement of parsimony. It is the art of econometrics to synthesize these two conflicting principles in an empirical model.

Simultaneous structural econometric models provide a natural framework to convey economic considerations into statistical models and to represent structural information. The testing procedures developed in this paper can be used to check the statistical adequacy of a large scale SSEM, but it can not be used to formulate structural hypotheses. The special difficulty in modeling large scale SSEMs is that we cannot estimate the unconstrained general statistical model. Consequently, we may not obtain any hint on potential structural information from the unconstrained general model.

Therefore, creative conjecture of structural hypotheses is of great importance in construction of large scale SSEMs. The constructive process of modeling starts with a conjecture of a structural model or behavioral equations in which the economic theory and the understanding of empirical economic phenomena are combined. An iterative process of obtaining evidence, revising the framework, and reinterpreting the evidence, will be essential for the construction of SSEMs.

References

AMEMIYA, T. (1985). *Advanced Econometrics.* Harvard University Press, Cambridge.

BIERENS, H. (1997). Nonparametric cointegration analysis. *Journal of Econometrics* **77** 379–404.

BLANCHARD, O. (2000). What do we know about macroeconomics that Fisher and Wicksell did not? NBER Working Paper 7550.

BREITUNG, J. (1995). A simultaneous equation approach to cointegrated systems. Sonderforschungbereich 373 at Humboldt University, No. 45.

BROWN, B. W. (1983). The identification problem in systems nonlinear in the variables. *Econometrica* **51** 175–196.

BROWN, B. W. (1985). The identification problem in simultaneous equation models with identities. *International Economic Review* **26** 45–66.

CHAREMZA, W. (1997). *New Directions in Econometric Practice: General to Specific Modeling, Cointegration, and Vector Autoregression.* 2nd ed., Elgar, Cheltenham.

CHEN, P. (2001). Structural models: A critical review of the modeling procedure. Bielefeld University, Faculty of Economics, Pfingsttagung, DStatG.

CHEN, P. (2002). Structural models: A model selection approach. Bielefeld University, Faculty of Economics.

CHEN, P., FROHN, J., LEMKE, W. (2005). Linear and nonlinear identities in simultaneous structural econometric models. *Applied Economics Quarterly*, (to appear).

CHEN, P., HSIAO, C. (2004a). Statistical modeling and complexity. Discussion Paper, Bielefeld University, Faculty of Economics.

CHEN, P., HSIAO, C. (2004b). Testing cointegration rank in large systems. Bielefeld University, Faculty of Economics.

CHEN, P., HSIAO, C. (2004c). Weak exogeneity in simultaneous equations models. Discussion Paper, Bielefeld University, Faculty of Economics.

CHOI, I. (2003). Subsampling vector autoregressive tests for linear constraints. Department of Economics, Hongkong University of Science and Technology.

CHOW, G. C. (1983). *Econometrics.* McGraw-Hill, New York.

DEUTSCHE BUNDESBANK (2000). Macro-Econometric Multi-Country Model: MEMMOD

DHAENE, G. (1997). *Encompassing, Formulation, Properties and Testing.* Springer, Berlin.

DHRYMES, P. J. (1993). *Topics in Advanced Econometrics.* Springer, Berlin.

ENGLE, R., HENDRY, D. F., RICHARD, J.-F. (1983). Exogeneity. *Econometrica* **51** 277–304.

ENGLE, R. F., GRANGER, W. J. (1987). Cointegration and error correction: Representation, estimation and testing. *Econometrica* **55** 251–276.

FROHN, J. (1995). *Grundausbildung in Ökonometrie.* 2nd ed., de Gruyter, Berlin.

FROHN, J. (1998). Zum Nutzen struktureller makroökonometrischer Modelle. *Ifo Studien* **44** 161–177.

FROHN, J. (1999). Structural macroeconometric models versus vectorautoregressive models. *Proceedings of the ISI World Congress,* Helsinki.

GRANGER, J. (1990). *Modeling Economic Series.* Clarendon, Oxford.

GRANGER, J., NEWBOLD, P. (1976). Spurious regression in econometrics. *Journal of Econometrics* **2** 111–120.

GREENE, W. (1993). *Econometric Analysis.* 2nd ed., Macmillan, New York.

HABRO, I., JOHANSEN, S., NEILSEN, B., RAHBEK, A. (1998). Asymptotic inference on cointegrating rank in partial systems. *Journal of Business & Economic Statistics* **16** 388–399.

HAMILTON, J. D. (1994). *Time Series Analysis.* Princeton University Press, Princeton.

HAUSMAN, J. A. (1975). An instrumental variable approach to full information maximum likelihood estimators for linear and nonlinear econometric models. *Econometrica* **43** 723–753.

HENDRY, D. (1995). *Dynamic Econometrics.* Oxford University Press, Oxford.

HENDRY, D. F., KROLZIG, H. M. (2001). New development in automatic general to specific modeling. In *Econometrics and Philosophy of Economics* (B. P., Stigum, ed.), Princeton University Press, Princeton.

HENDRY, D. F., MIZON, G. E. (1990). Procrustean econometrics, or stretching and squeezing of data. Reprinted in Granger (1990), p. 121–136.

HENDRY, D. F., MIZON, G. E. (1993). Evaluating dynamic econometric models by encompassing the VAR. In *Models, Methods and Applications of Econometrics* (Phillips ed.), p. 272–300.

HSIAO, C. (1997). Cointegration and dynamic simultaneous equations models. *Econometrica* **65** 647–670.

JOHANSEN, S. (1992). Testing weak exogeneity and the order of cointegration in U. K. money demand data. *Journal of Policy Modeling* **14** 313–334.

JOHANSEN, S. (1995). *Likelihood-Based Inference in Cointegrated Vector Autoregressive Models.* Oxford University Press, Oxford.

JUDGE, G., GRIFFITHS, W., HILL, R. C., LÜTKEPOHL, H., LEE, T. (1985). *The Theory and Practice of Econometrics.* 2nd ed., John Wiley & Sons, New York.

KOOPMANS, T. C., HOOD, W. C. (1953). The Estimation of Simultaneous Linear Economic Relationships. In *Studies in Econometric Method* (W. C. Hood, T. C. Koopmans (eds.), Chapter 6. Cowles Commission Monograph No. 14, Wiley, New York.

LUCAS, R. E. (1976). Econometric policy evaluation: A critique. *Carnegie-Rochester Conference Series on Public Policy* **1** 19–46.

MADDALA, G. (1988). *Introduction to Econometrics.* Macmillan, New York.

MALINVAUD, E. (1980). *Statistical Methods of Econometrics.* 3rd ed., North Holland, Amsterdam.

NAYLOR, T. H., SEAKS, T., WICHERN, D. (1972). Box-Jenkins methods: An alternative to econometric models. *International Statistical Review* **40** 123–137.

NELSON, C. R. (1972). The prediction performance of the FRB-MIT-PENN model of US economy. *American Economic Review* **62** 902–917.

SCHMIDT, P. (1976). *Econometrics.* Marcel Dekker, New York.

SIMS, C. (1980). Macroeconomics and reality. *Econometrica* **48** 1–48.

SIMS, C. (1982). Policy analysis with econometric models. *Brookings Papers on Economic Activity* **1** 107–152.

SPANOS, A. (1990). The simultaneous-equation model revisited. *Journal of Econometrics* **44** 87–105.

STOCK J. H., WATSON M. W. (1988). Testing for common trends. *Journal of the American Statistical Association* **83** 1097–1107.

THEIL, H. (1971). *Principles of Econometrics.* Wiley, New York.

WINKER, P., MARINGER, D. (2004). Optimal lag structural selection in VEC-models. Conference of Computational Economics 2004, Amsterdam.

ZELLNER, A. (1979). Statistical analysis of econometric models. *Journal of the American Statistical Association* **74** 628–643.

ZELLNER, A., PALM, F. (1974). Time series analysis and simultaneous equations econometrics. *Journal of Econometrics* **2** 17–54.

3 Dynamic Factor Models

Jörg Breitung[1] and Sandra Eickmeier[2]

[1] Institut für Ökonometrie, Universität Bonn
breitung@uni-bonn.de
[2] Forschungszentrum, Deutsche Bundesbank
sandra.eickmeier@bundesbank.de

Summary. Factor models can cope with many variables without running into scarce degrees of freedom problems often faced in a regression-based analysis. In this article we review recent work on dynamic factor models that have become popular in macroeconomic policy analysis and forecasting. By means of an empirical application we demonstrate that these models turn out to be useful in investigating macroeconomic problems.

3.1 Introduction

In recent years, large-dimensional dynamic factor models have become popular in empirical macroeconomics. They are more advantageous than other methods in various respects. Factor models can cope with many variables without running into scarce degrees of freedom problems often faced in regression-based analyses. Researchers and policy makers nowadays have more data at a more disaggregated level at their disposal than ever before. Once collected, the data can be processed easily and rapidly owing to the now wide-spread use of high-capacity computers. Exploiting a lot of information can lead to more precise forecasts and macroeconomic analyses. The use of many variables further reflects a central bank's practice of 'looking at everything' as emphasized, for example, by Bernanke and Boivin (2003). A second advantage of factor models is that idiosyncratic movements which possibly include measurement error and local shocks can be eliminated. This yields a more reliable signal for policy makers and prevents them from reacting to idiosyncratic movements. In addition, the estimation of common factors or common shocks is of intrinsic interest in some applications. A third important advantage is that factor modellers can remain agnostic about the structure of the economy and do not need to rely on overly tight assumptions as is sometimes the case in structural models. It also represents an advantage over structural VAR models where the researcher has to take a stance on the variables to include which, in turn, determine

the outcome, and where the number of variables determine the number of shocks.

In this article we review recent work on dynamic factor models and illustrate the concepts with an empirical example. In Section 2 the traditional factor model is considered and the approximate factor model is outlined in Section 3. Different test procedures for determining the number of factors are discussed in Section 4. The dynamic factor model is considered in Section 5. Section 6 gives an overview of recent empirical work based on dynamic factor models and Section 7 presents the results of estimating a large-scale dynamic factor model for a large set of macroeconomic variables from European Monetary Union (EMU) member countries and Central and Eastern European Countries (CEECs). Finally, Section 8 concludes.

3.2 The Strict Factor Model

In an r-factor model each element of the vector $y_t = [y_{1t}, \ldots, y_{Nt}]'$ can be represented as

$$\begin{aligned} y_{it} &= \lambda_{i1} f_{1t} + \cdots + \lambda_{ir} f_{rt} + u_{it} \ , \quad t = 1, \ldots, T \\ &= \lambda_{i\bullet}' f_t + u_{it} \ , \end{aligned}$$

where $\lambda_{i\bullet}' = [\lambda_{i1}, \ldots, \lambda_{ir}]$ and $f_t = [f_{1t}, \ldots, f_{rt}]'$. The vector $u_t = [u_{1t}, \ldots, u_{Nt}]'$ comprises N idiosyncratic components and f_t is a vector of r common factors.

In matrix notation the model is written as

$$\begin{aligned} y_t &= \Lambda f_t + u_t \\ Y &= F\Lambda' + U, \end{aligned}$$

where $\Lambda = [\lambda_{1\bullet}, \ldots, \lambda_{N\bullet}]'$, $Y = [y_1, \ldots, y_T]'$, $F = [f_1, \ldots, f_T]'$ and $U = [u_1, \ldots, u_T]'$.

For the *strict factor model* it is assumed that u_t is a vector of mutually uncorrelated errors with $E(u_t) = 0$ and $E(u_t u_t') = \Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_N^2)$. For the vector of common factors we assume $E(f_t) = 0$ and $E(f_t f_t') = \Omega$.[1] Furthermore, $E(f_t u_t') = 0$. >From these assumptions it follows that[2]

$$\Psi = E(y_t y_t') = \Lambda\Omega\Lambda' + \Sigma.$$

The loading matrix Λ can be estimated by minimizing the residual sum of squares:

$$\sum_{t=1}^{T} (y_t - Bf_t)'(y_t - Bf_t) \tag{3.1}$$

[1] That is we assume that $E(y_t) = 0$. In practice, the means of the variables are subtracted to obtain a vector of mean zero variables.

[2] In many applications the correlation matrix is used instead of the covariance matrix of y_t. This standardization affects the properties of the principal component estimator, whereas the ML estimator is invariant with respect to a standardization of the variables.

subject to the constraint $B'B = I_r$. Differentiating (3.1) with respect to B and F yields the first order condition $(\mu I_N - S)\hat{\beta}_k = 0$ for $k = 1, \ldots, r$, where $S = T^{-1}\sum_{t=1}^{T} y_t y_t'$ and $\hat{\beta}_i$ is the i'th column of $\widehat{B}$, the matrix that minimizes the criterion function (3.1). Thus, the columns of $\widehat{B}$ result as the eigenvectors of the r largest eigenvalues of the matrix S. The matrix $\widehat{B}$ is the *Principal Components* (PC) estimator of Λ.

To analyse the properties of the PC estimator it is instructive to rewrite the PC estimator as an instrumental variable (IV) estimator. The PC estimator can be shown to solve the following moment condition:

$$\sum_{t=1}^{T} \widehat{B}' y_t \hat{u}_t' = 0, \tag{3.2}$$

where $\hat{u}_t = \widehat{B}_\perp' y_t$ and $\widehat{B}_\perp$ is an $N \times (N-r)$ orthogonal complement of $\widehat{B}$ such that $\widehat{B}_\perp' \widehat{B} = 0$. Specifically,

$$\widehat{B}_\perp = I_N - \widehat{B}(\widehat{B}'\widehat{B})^{-1}\widehat{B}' = I_N - \widehat{B}\widehat{B}',$$

where we have used the fact that $\widehat{B}'\widehat{B} = I_r$. Therefore, the moment condition can be written as $\sum_{t=1}^{N} \hat{f}_t \hat{u}_t'$, where $\hat{u}_t = y_t - \widehat{B}'\hat{f}_t$ and $\hat{f}_t = \widehat{B}' y_t$. Since the components of $\hat{f}_t$ are linear combinations of y_t, the instruments are correlated with $\hat{u}_t$, in general. Therefore, the PC estimator is inconsistent for fixed N and $T \to \infty$ unless $\Sigma = \sigma^2 I$.[3]

An alternative representation that will give rise to a new class of IV estimators is given by choosing a different orthogonal complement $\widehat{B}_\perp$. Let $\Lambda = [\Lambda_1', \Lambda_2']'$ such that Λ_1 and Λ_2 are $(N-r) \times r$ and $r \times r$ submatrices, respectively. The matrix $U = [u_1, \ldots, u_N]'$ is partitioned accordingly such that $U = [U_1', U_2']$ and U_1 (U_2) are $T \times (N-r)$ $(T \times r)$ submatrices. A system of equations results from solving $Y_2 = F\Lambda_2' + U_2$ for F and inserting the result into the first set of equations:

$$\begin{aligned} Y_1 &= (Y_2 - U_2)(\Lambda_2')^{-1}\Lambda_1' + U_1 \\ &= Y_2\Theta' + V, \end{aligned} \tag{3.3}$$

where $\Theta = \Lambda_1\Lambda_2^{-1}$ and $V = U_1 - U_2\Theta'$. Accordingly Θ yields an estimator for the renormalized loading matrix $B^* = [\Theta', I_r]'$ and $B_\perp^* = [I_r, \Theta']'$.

The i'th equation of system (3.3) can be consistently estimated based on the following $N - r - 1$ moment conditions

$$E(y_{kt} v_{it}) = 0, \quad k = 1, \ldots, i-1, \; i+1, \ldots, N-r, \tag{3.4}$$

that is, we do not employ y_{it} and $y_{n+1,t}, \ldots, y_{Nt}$ as instruments as they are correlated with v_{it}. Accordingly, a Generalized Method of Moments (GMM) estimator based on $(N-r)(N-r-1)$ moment conditions can be constructed to estimate the $n \cdot r$ parameters in the matrix Θ. An important problem with this estimator is

[3] To see that the PC estimator yields a consistent estimator of the factor space for $\Sigma = \sigma^2 I_N$ let B denote the matrix of r eigenvectors of Ψ. It follows that $B'\Psi B_\perp = B'\Lambda\Omega\Lambda' B_\perp$. The latter expression becomes zero if $B = \Lambda Q$, where Q is some regular $r \times r$ matrix.

that the number of instruments increases rapidly as N increases. It is well known that, if the number of instruments is large relative to the number of observations, the GMM estimator may have poor properties in small samples. Furthermore, if $n^2 - n > T$, the weight matrix for the GMM estimator is singular. Therefore it is desirable to construct a GMM estimator based on a smaller number of instruments. Breitung (2005) proposes a just-identified IV etimator based on equation specific instruments that do not involve y_{it} and $y_{n+1,t}, \ldots, y_{Nt}$.

In the case of homogeneous variances (i.e. $\Sigma = \sigma^2 I_N$) the PC estimator is the maximum likelihood (ML) estimator assuming that y_t is normally distributed. In the general case with $\Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_N^2)$ the ML estimator minimizes the function $\ell^* = tr(S\Sigma^{-1}) + \log|\Sigma|$ (cf. Jöreskog, 1969). Various iterative procedures have been suggested to compute the ML estimator from the set of highly nonlinear first order conditions. For large factor models (with $N > 20$, say) it has been observed that the convergence of the usual maximization algorithms is quite slow and in many cases the algorithms have difficulty in converging to the global maximum.

3.3 Approximate Factor Models

The fairly restrictive assumption of the strict factor model can be relaxed if it is assumed that the number of variables (N) tends to infinity (cf. Chamberlain and Rothshild, 1983; Stock and Watson, 2002a and Bai, 2003). First, it is possible to allow for (weak) serial correlation of the idiosyncratic errors. Thus, the PC estimator remains consistent if the idiosyncratic errors are generated by (possibly different) stationary ARMA processes. However, persistent and non-ergodic processes such as the random walk are ruled out. Second, the idiosyncratic errors may be weakly cross-correlated and heteroskedastic. This allows for finite 'clusters of correlation' among the errors. Another way to express this assumption is to assume that all eigenvalues of $E(u_t u_t') = \Sigma$ are bounded. Third, the model allows for weak correlation among the factors and the idiosyncratic components. Finally, $N^{-1}\Lambda'\Lambda$ must converge to a positive definite limiting matrix. Accordingly, on average the factors contribute to all variables with a similar order of magnitude. This assumption rules out the possibility that the factors contribute only to a limited number of variables, whereas for an increasing number of remaining variables the loadings are zero.

Beside these assumptions a number of further technical assumptions restrict the moments of the elements of the random vectors f_t and u_t. With these assumptions Bai (2003) establishes the consistency and asymptotic normality of the PC estimator for Λ and f_t. However, as demonstrated by Bovin and Ng (2005a) the small sample properties may be severely affected when (a part of) the data is cross-correlated.

3.4 Specifying the Number of Factors

In practice, the number of factors necessary to represent the correlation among the variables is usually unknown. To determine the number of factors empirically a number of criteria were suggested. First, the eigenvalues of the sample correlation matrix R may roughly indicate the number of common factors. Since $tr(R) =$

$N = \sum_{i=1}^{N} \mu_i$, where μ_i denotes the i'th eigenvalue of R (in descending order), the fraction of the total variance explained by k common factors is $\tau(k) = (\sum_{i=1}^{k} \mu_i)/N$. Unfortunately, there is no generally accepted limit for the explained variance that indicates a sufficient fit. Sometimes it is recommended to include those factors with an eigenvalue larger than unity, since these factors explain more than an 'average factor'.

In some applications (typically in psychological or sociological studies) two or three factors explain more than 90 percent of the variables, whereas in macroeconomic panels a variance ratio of 40 percent is sometimes considered as a reasonable fit.

A related method is the 'Scree-test'. Cattell (1966) observed that the graph of the eigenvalues (in descending order) of an uncorrelated data set forms a straight line with an almost horizontal slope. Therefore, the point in the eigenvalue graph where the eigenvalues begin to level off with a flat and steady decrease is an estimator of the sufficient number of factors. Obviously such a criterion is often fairly subjective because it is not uncommon to find more than one major break in the eigenvalue graph and there is no unambiguous rule to use.

Several more objective criteria based on statistical tests are available that can be used to determine the number of common factors. If it is assumed that r is the true number of common factors, then the idiosyncratic components u_t should be uncorrelated. Therefore it is natural to apply tests that are able to indicate a contemporaneous correlation among the elements of u_t. The score test is based on the sum of all relevant $N(N-1)/2$ squared correlations. This test is asymptotically equivalent to the LR test based on (two times) the difference of the log-likelihood of the model assuming r_0 factors against a model with an unrestricted covariance matrix. An important problem of these tests is that they require $T >> N >> r$. Otherwise the performance of these tests is quite poor. Therefore, in typical macroeconomic panels which include more than 50 variables these tests are not applicable.

For the approximate factor model Bai and Ng (2002) suggested information criteria that can be used to estimate the number of factors consistently as N and T tend to infinity. Let $V(k) = (NT)^{-1} \sum_{t=1}^{T} \hat{u}_t' \hat{u}_t$ denote the (overall) sum of squared residuals from a k-factor model, where $\hat{u}_t = y_t - \widehat{B} \hat{f}_t$ is the $N \times 1$ vector of estimated idiosyncratic errors. Bai and Ng (2002) suggest several variants of the information criterion, where the most popular statistic is

$$IC_{p2}(k) = \log[V(k)] + k \left(\frac{N+T}{NT} \right) \log[\min\{N,T\}] .$$

The estimated number of factors ($\hat{k}$) is obtained from minimizing the information criterion in the range $k = 0, 1, \ldots, kmax$ where $kmax$ is some pre-specified upper bound for the number of factors. As N and T tend to infinity, $\hat{k} \xrightarrow{p} r$, i.e., the criterion is (weakly) consistent.

3.5 Dynamic Factor Models

The dynamic factor model is given by

$$y_t = \Lambda_0 g_t + \Lambda_1 g_{t-1} + \cdots + \Lambda_m g_{t-m} + u_t \,, \tag{3.5}$$

where $\Lambda_0, \ldots, \Lambda_m$ are $N \times q$ matrices and g_t is a vector of q stationary factors. As before, the idiosyncratic components of u_t are assumed to be independent (or weakly dependent) stationary processes.

Forni *et al.* (2004) suggest an estimation procedure of the innovations of the factors $\eta_t = g_t - E(g_t|g_{t-1}, g_{t-2}, \ldots)$. Let $f_t = [g_t', g_{t-1}', \ldots, g_{t-m}']'$ denote the $r = (m+1)q$ vector of 'static' factors such that

$$y_t = \Lambda^* f_t + u_t \,, \tag{3.6}$$

where $\Lambda^* = [\Lambda_0, \ldots, \Lambda_m]$. In a first step the static factors f_t are estimated by PC. Let $\hat{f}_t$ denote the vector of estimated factors. It is important to note that a (PC) estimator does not estimate the original vector f_t but some 'rotated' vector Qf_t such that the components of (Qf_t) are orthogonal. In a second step a VAR model is estimated:

$$\hat{f}_t = A_1 \hat{f}_{t-1} + \cdots + A_p \hat{f}_{t-p} + e_t \,. \tag{3.7}$$

Since $\hat{f}_t$ includes estimates of the lagged factors, some of the VAR equations are identities (at least asymptotically) and, therefore, the rank of the residual covariance matrix $\widehat{\Sigma}_e = T^{-1} \sum_{t=p+1}^{T} \hat{e}_t \hat{e}_t'$ is q, as $N \to \infty$. Let $\widehat{W}_r$ denote the matrix of q eigenvectors associated with the q largest eigenvalues of $\widehat{\Sigma}_e$. The estimate of the innovations of the dynamic factors results as $\hat{\eta}_t = \widehat{W}_r' \hat{e}_t$. These estimates can be used to identify structural shocks that drive the common factors (cf. Forni *et al.*, 2004, Giannone *et al.*, 2002).

An important problem is to determine the number of dynamic factors q from the vector of r static factors. Forni *et al.* (2004) suggest an informal criterion based on the portion of explained variances, whereas Bai and Ng (2005) and Stock and Watson (2005) suggest consistent selection procedures based on principal components. Breitung and Kretschmer (2005) propose a test procedure based on the canonical correlation between $\hat{f}_t$ and $\hat{f}_{t-1}$. The i'th eigenvalue from a canonical correlation analysis can be seen as an R^2 from a regression of $\hat{v}_i' \hat{f}_t$ on $\hat{f}_{t-1}$, where $\hat{v}_i$ denotes the associated eigenvector. If there is a linear combination of $\hat{f}_t$ that corresponds to a lagged factor, then this linear combination is perfectly predictable and, therefore, the corresponding R^2 (i.e. the eigenvalue) will tend to unity. On the other hand, if the linear combination reproduces the innovations of the original factor, then this linear combination is not predictable and, therefore, the eigenvalue will tend to zero. Based on this reasoning, information criteria and tests of the number of factors are suggested by Breitung and Kretschmer (2005).

Forni *et al.* (2000, 2002) suggest an estimator of the dynamic factors in the frequency domain. This estimator is based on the frequency domain representation of the factor model given by

$$f_y(\omega) = f_\chi(\omega) + f_u(\omega) \,,$$

where $\chi_t = \Lambda_0 f_t + \cdots + \Lambda_m f_{t-m}$ denotes the vector of common components of y_t, f_χ is the associated spectral density matrix, f_y is the spectral density matrix of y_t and f_u is the (diagonal) spectral density matrix of u_t. Dynamic principal components analysis applied to the frequencies $\omega \in [0, \pi]$ (Brillinger, 1981) yields a consistent estimate of the spectral density matrix $f_\chi(\omega)$. An estimate of the common components χ_{it} is obtained by computing the time domain representation of the process from an inversion of the spectral densities. The frequency domain estimator yields a two-sided filter such that $\tilde{f}_t = \sum_{j=-\infty}^{\infty} \widehat{\Psi}_j' y_{t-j}$, where, in practice, the infinite limits are truncated. Forni *et al.* (2005) also suggest a one-sided filter which is based on a conventional principal component analysis of the transformed vector $\tilde{y}_t = \widehat{\Sigma}^{-1/2} y_t$, where $\widehat{\Sigma}$ is the (frequency domain) estimate of the covariance matrix of u_t. This one-sided estimator can be used for forecasting based on the common factors.

3.6 Overview of Existing Applications

Dynamic factor models were traditionally used to construct economic indicators and for forecasting. More recently, they have been applied to macroeconomic analysis, mainly with respect to monetary policy and international business cycles. We briefly give an overview of existing applications of dynamic factor models in these four fields, before providing a macro analytic illustration.

3.6.1 Construction of Economic Indicators

The two most prominent examples of monthly coincident business cycle indicators, to which policy makers and other economic agents often refer, are the Chicago Fed National Activity Index[4] (CFNAI) for the US and EuroCOIN for the Euro area. The CFNAI estimate, which dates back to 1967, is simply the first static principal component of a large macro data set. It is the most direct successor to indicators which were first developed by Stock and Watson but retired by the end of 2003. EuroCOIN is estimated as the common component of Euro-area GDP based on dynamic principal component analysis. It was developed by Altissimo *et al.* (2001) and is made available from 1987 onwards by the CEPR.[5] Measures of core inflation have been constructed analogously (e. g. Cristadoro *et al.*, 2001, and Kapetanios, 2004, for the Euro area and Kapetanios, 2004, for the UK).

3.6.2 Forecasting

Factor models are widely used in central banks and research institutions as a forecasting tool. The forecasting equation typically has the form

$$y_{t+h}^h = \mu + a(L) y_t + b(L) \hat{f}_t + e_{t+h}^h \ , \tag{3.8}$$

where y_t is the variable to be forecasted at period $t+h$ and e_{t+h} denotes the h-step ahead prediction error. Accordingly, information used to forecast y_t are the past of

[4] See http://www.chicagofed.org/economic_research_and_data/cfnai.cfm.
[5] See http://www.cepr.org/data/eurocoin/.

the variable and the common factor estimates $\hat{f}_t$ extracted from an additional data set.

Factor models have been used to predict real and nominal variables in the US (e. g. Stock and Watson, 1999, 2002a,b; Giacomini and White, 2003; Banerjee and Marcellino, 2003), in the Euro area (e. g. Forni *et al.*, 2000, 2003; Camba-Mendez and Kapetanios, 2004; Marcellino *et al.*, 2003; Banerjee *et al.*, 2003), for Germany (Schumacher and Dreger, 2004; Schuhmacher, 2005), for the UK (Artis *et al.*, 2004) and for the Netherlands (den Reijer, 2005). The factor model forecasts are generally compared to simple linear benchmark time series models, such as AR models, AR models with single measurable leading indicators and VAR models. More recently, they have also been compared with pooled single indicator forecasts or forecasts based on 'best' single indicator models or groups of indicators derived using automated selection procedures (PCGets) (e. g. Banerjee and Marcellino, 2003; Watson, 2003). Pooling variables versus combining forecasts is a particularly interesting comparison, since both approaches claim to exploit a lot of information.[6]

Overall, results are quite encouraging, and factor models are often shown to be more successful in terms of forecasting performance than smaller benchmark models. Three remarks are, however, in order. First, the forecasting performance of factor models apparently depends on the types of variable one wishes to forecast, the countries/regions of interest, the underlying data sets, the benchmark models and horizons. Unfortunately, a systematic assessment of the determinants of the relative forecast performance of factor models is still not available. Second, it may not be sufficient to include just the first or the first few factors. Instead, a factor which explains not much of the entire panel, say, the fifth or sixth principal component, may be important for the variable one wishes to forecast (Banerjee and Marcellino, 2003). Finally, the selection of the variables to be included in the data set is ad hoc in most applications. The same data set is often used to predict different variables. This may, however, not be adequate. Instead, one should only include variables which exhibit high explanatory power with respect to the variable that one aims to forecast (see also Bovin and Ng, 2005b).

3.6.3 Monetary Policy Analysis

Forni *et al.* (2004) and Giannone *et al.* (2002, 2004) identify the main macroeconomic shocks in the US economy and estimate policy rules conditional on the shocks. Sala (2003) investigates the transmission of common Euro-area monetary policy shocks to individual EMU countries. Cimadomo (2003) assesses the proliferation of economy-wide shocks to sectors in the US and examines if systematic monetary policy has distributional and asymmetric effects across sectors. All these studies rely on the structural dynamic factor model developed by Forni *et al.* (2004). Bernanke *et al.* (2005), Stock and Watson (2005) and Favero *et al.* (2005) use a different but related approach. The two former papers address the problem of omitted

[6]The models are further used to investigate the explanatory power of certain groups of variables, for example financial variables (Forni *et al.*, 2003) or variables summarizing international influences for domestic activity (see, for example, Banerjee *et al.*, 2003) who investigate the ability of US variables or factors to predict Euro-area inflation and output growth.

variables bias inherent in many simple small-scale VAR models. They show for the US that the inclusion of factors in monetary VARs, denoted by factor-augmented VAR (FAVAR) models, can eliminate the well-known price puzzle in the US. Favero *et al.* (2005) confirm these findings for the US and for some individual Euro-area economies. They further demonstrate that the inclusion of factors estimated from dynamic factor models in the instrument set used for estimation of Taylor rules increases the precision of the parameters estimates.

3.6.4 International Business Cycles

Malek Mansour (2003) and Helbling and Bayoumi (2003) estimate a world and, respectively, a G7 business cycle and investigate to what extent the common cycle contributes to economic variation in individual countries. Eickmeier (2004) investigates the transmission of structural shocks from the US to Germany and assesses the relevance of the various transmission channels and global shocks, thereby relying on the Forni *et al.* (2004) framework. Marcellino *et al.* (2000) and Eickmeier (2005) investigate economic comovements in the Euro-area. They try to give the common Euro-area factors an economic interpretation by relating them to individual countries and variables using correlation measures.

3.7 Empirical Application

Our application sheds some light on economic comovements in Europe by fitting the large-scale dynamic factor model to a large set of macroeconomic variables from European monetary union (EMU) member countries and central and eastern European countries (CEECs). We determine the dimension of the Euro-area economy, i.e. the number of macroeconomic driving forces which are common to all EMU countries and which explain a significant share of the overall variance in the set and we make some tentative interpretation. Most importantly, our application addresses the recent discussion on whether the CEECs should join the EMU.

One of the criteria that should be satisfied is the synchronization of business cycles. In what follows, we investigate how important Euro-area factors are for the CEECs compared to the current EMU members. In addition, the heterogeneity of the influences of the common factors across the CEECs is examined.[7]

[7] A more comprehensive study based on a slightly different data set is provided by Eickmeier and Breitung (2005).

Table 3.1: **Criteria for selecting the number of factors.**

	Bai and Ng criteria			Variance shares of PCs	
r	IC_{p1}	IC_{p2}	IC_{p3}	Static PCs	Dynamic PCs
1	−0.096	−0.091	−0.109	0.159	0.211
2	−0.105*	−0.095*	−0.131	0.248	0.326
3	−0.100	−0.084	−0.138*	0.317	0.418
4	−0.082	−0.061	−0.133	0.371	0.494
5	−0.065	−0.039	−0.129	0.423	0.555
6	−0.037	−0.006	−0.114	0.464	0.608
7	−0.014	0.023	−0.103	0.504	0.656
8	0.012	0.054	−0.090	0.541	0.698
9	0.036	0.084	−0.078	0.575	0.734
10	0.066	0.118	−0.062	0.604	0.768

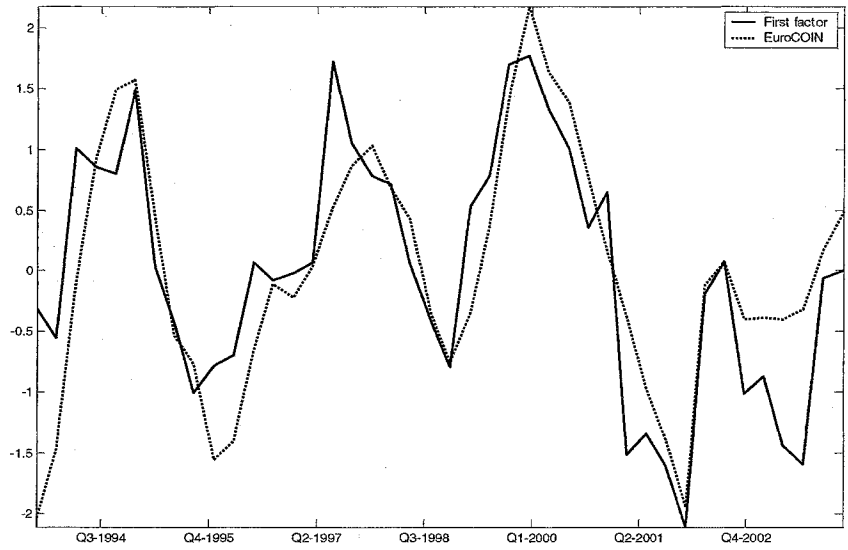
Note: The maximal number of factors for the Bai and Ng (2002) criteria is $r_{max} = 10$. The cumulative variance shares present the variance share explained by the first r principal components (PC). An asterisk indicates the minimum.

Our data set contains 41 aggregate Euro-area time series, 20 key variables of each of the core Euro-area countries (Austria, Belgium, France, Germany, Italy, Netherlands, Spain), real GDP and consumer prices for the remaining Euro-area economies (Finland, Greece, Ireland, Luxembourg, Portugal) and for eight CEECs (Czech Republic, Estonia, Hungary, Lithuania, Latvia, Poland, Slovenia, Slovak Republic) as well as some global[8] variables.[9]

Overall, we include $N = 208$ quarterly series. The sample ranges from 1993 Q1 to 2003 Q4. The factor analysis requires some pre-treatment of the data. Series exhibiting a seasonal pattern were seasonally adjusted. Integrated series were made stationary through differencing. Logarithms were taken of the series which were not in rates or negative, and we removed outliers. We standardized the series to have a mean of zero and a variance of one.

[8] Among the global variables are US GDP and world energy prices. Studies have shown that fluctuations in these variables may influence the Euro area (see, for example, Jiménez-Rodríguez and Sánchez, 2005; Peersman, 2005).

[9] The aggregate Euro-area series are taken from the data set underlying the ECB's area wide model (for a detailed description see Fagan *et al.*, 2001). The remaining series mainly stem from OECD and IMF statistics.



Note: The monthly EuroCOIN series was converted into a quarterly series. It was normalized to have a mean of zero and a variance of one.

Figure 3.1: *Euro-area business cycle estimates.*

The series are collected in the vector $N \times 1$ vector y_t $(t = 1, 2, \ldots, T)$. It is assumed that y_t follows an approximate dynamic factor model as described in Section 3. The r common Euro-area factors collected in f_t are estimated by applying static principal component analysis to the correlation matrix of y_t. On the basis of the IC_{p3} criterion of Bai and Ng (2002), we choose $r = 3$, although the other two criteria suggest $r = 2$ (Table 1). One reason is that factors are still estimated consistently if the number of common factors is overestimated, but not if it is underestimated (Stock and Watson, 2002b; Kapetanios and Marcellino, 2003; Artis *et al.*, 2004). Another reason is that two factors explain a relatively low share of the total variance (25 percent), whereas three factors account for 32 percent which is more consistent with previous findings for macroeconomic Euro-area data sets (Table 1).[10]

The common factors f_t do not bear a direct structural interpretation. One reason is that f_t may be a linear combination of the q 'true' dynamic factors and their lags. Using the consistent Schwarz criterion of Breitung and Kretschmer (2005), we obtain $q = 2$, conditional on $r = 3$. That is, one of the two static factors enter the factor model with a lag. Informal criteria are also used in practice. Two dynamic principal components explain 33 percent (Table 1). This is comparable to the variance explained by the r static factors. The other criterion consists in requiring each dynamic principal component to explain at least a certain share, for example 10 percent, of the total variance. This would also suggest $q = 2$.

[10]Those range between 32 and 55 percent (Marcellino *et al.*, 2000; Eickmeier, 2005; Altissimo *et al.*, 2001).

Even if the dynamic factors were separated from their lags, they could not be given a direct economic meaning, since they are only identified up to a linear transformation. Some tentative interpretation of the factors is given nevertheless. In business cycle applications, the first factor is often interpreted as a common cycle. Indeed, as is obvious from Figure 1, our first factor is highly correlated with EuroCOIN and can therefore be interpreted as the Euro-area business cycle. To facilitate the interpretation of the other factors, the factors may be rotated to obtain a new set of factors which satisfies certain identifying criteria, as done in Eickmeier (2005). Another possibility consists in estimating the common structural shocks behind f_t using structural vector autoregression (SVAR) and PC techniques as suggested by Forni *et al.* (2004). This would also allow us to investigate how common Euro-area shocks spread to the CEECs.

Table 3.2: *Variance shares explained by the common factors.*

	Δ GDP		Δ GDP
AUT	0.42	CZ	0.03
BEL	0.60	ES	0.08
FIN	0.19	HU	0.18
FRA	0.66	LT	0.03
GER	0.60	LV	0.03
GRC	0.07	PL	0.07
IRE	0.27	SI	0.11
ITA	0.44	SK	0.05
LUX	0.44		
NLD	0.54		
PRT	0.09		
ESP	0.14		
Mean all countries	0.25	Std. all countries	0.22
Mean EMU	0.37	Std. EMU	0.21
Mean EMU - GPI	0.45	Std. EMU - GPI	0.18
Mean CEECs	0.07	Std. CEECs	0.05

Note: EMU - GPI denotes the Euro area less Greece, Portugal and Ireland.

Table 2 shows how much of the variance of output growth in CEECs and EMU countries is explained by the Euro-area factors. On average, the common factors explain a larger part of output growth in EMU economies (37 percent) compared to the CEECs (7 percent). Interestingly, the shares of the peripheral countries (Greece, Portugal and Ireland) are smaller than the corresponding shares in a number of CEECs. Of the latter, Hungary and Slovenia exhibit the largest variance shares explained by the Euro-area factors. The dispersion across EMU countries is about four times as large as the dispersion across the CEECs. The difference is somewhat lower when Greece, Portugal and Ireland are excluded from the EMU group.[11]

[11] These small peripheral countries were found to exhibit a relatively low synchronization with the rest of the Euro area and are sometimes treated separately (e.g. Korhonen, 2003).

3.8 Conclusion

In this paper we have reviewed and complemented recent work on dynamic factor models. By means of an empirical application we have demonstrated that these models turn out to be useful in investigating macroeconomic problems such as the economic consequences for central and eastern European countries of joining the European Monetary Union. Nevertheless, several important issues remain unsettled. First it turns out that the determination of the number of factors representing the relevant information in the data set is still a delicate issue. Since Bai and Ng (2002) have made available a number of consistent information criteria it has been observed that alternative criteria may suggest quite different number of factors. Furthermore, the results are often not robust and the inclusion of a few additional variables may have a substantial effect on the number of factors.

Even if dynamic factors may explain more than a half of the total variance it is not clear whether the idiosyncratic components can be treated as irrelevant 'noise'. It may well be that the idiosyncratic components are important for the analysis of macroeconomic variables. On the other hand, the loss of information may even be more severe if one focusses on a few variables (as in typical VAR studies) instead of a small number of factors. Another important problem is to attach an economic meaning to the estimated factors. As in traditional econometric work, structural identifying assumptions may be employed to admit an economic interpretation of the factors (cf. Breitung, 2005). Clearly, more empirical work is necessary to assess the potentials and pitfalls of dynamic factor models in empirical macroeconomic.

References

Altissimo, F., Bassanetti, A., Cristadoro, R., Forni, M., Hallin, M., Lippi, M., Reichlin, L. (2001). EuroCOIN: A real time coincident indicator of the euro area business cycle. CEPR Working Paper 3108.

Artis, M., Banerjee, A., Marcellino, M. (2004). Factor forecasts for the UK. EUI Florence, Mimeo.

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica* **71** 135–171.

Bai, J., Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica* **70** 191–221.

Bai, J., Ng, S. (2005). Determining the number of primitive shocks in factor models. New York University, Mimeo.

Banerjee, A., Marcellino, M. (2003). Are there any reliable leading indicators for US inflation and GDP growth? IGIR Working Paper 236.

Banerjee, A., Marcellino, M., Masten, I. (2003). Leading indicators for Euro area inflation and GDP growth. CEPR Working Paper 3893.

Bernanke, B. S., Boivin, J. (2003). Monetary policy in a data-rich environment. *Journal of Monetary Economics* **50** 525–546.

BERNANKE, B. S., BOIVIN, J., ELIASZ, P. (2005). Measuring the effects of monetary policy: A factor-augmented vector autoregressive (FAVAR) approach. *Quarterly Journal of Economics* **120** 387–422.

BOVIN, J., NG, S. ((). 2. *0* **0** 5.a). Are more data always better for factor analysis?Journal of Econometrics (forthcoming).

BOVIN, J., NG, S. (2005b). Understanding and comparing factor-based forecasts. Columbia Business School, Mimeo.

BREITUNG, J., KRETSCHMER, U. (2005). Identification and estimation of dynamic factors from large macroeconomic panels. University of Bonn, Mimeo.

BREITUNG, J. (2005). Estimation and inference in dynamic factor models. University of Bonn, Mimeo.

BRILLINGER, D. R. (1981). *Time series data analysis and theory.* Holt, Rinehart and Winston, New York.

CAMBA-MENDEZ, G., KAPETANIOS, G. (2004). Forecasting Euro area inflation using dynamic factor measures of underlying inflation. ECB Working Paper 402.

CATELL, R. B. (1966). The Scree test for the number of factors. *Multivariate Behavioral Research* **1** 245–276.

CHAMBERLAIN, G., ROTHSCHILD, M. (1983). Arbitrage, factor structure and mean-variance analysis in large asset markets. *Econometrica* **51** 1305–1324.

CIMADOMO, J. (2003). The effects of systematic monetary policy on sectors: A factor model analysis. ECARES - Université Libre de Bruxelles, Mimeo.

CRISTADORO, R., FORNI, M., REICHLIN, L., VERONESE, G. (2001). A core inflation index for the Euro area. CEPR Discussion Paper 3097.

DEN REIJER, A. H. J. (2005). Forecasting Dutch GDP using large scale factor models. DNB Working Paper 28.

EICKMEIER, S. (2004). Business cycle transmission from the US to Germany - a structural factor approach. Bundesbank Discussion Paper 12/2004, revised version.

EICKMEIER, S. (2005). Common stationary and non-stationary factors in the Euro area analyzed in a large-scale factor model. Bundesbank Discussion Paper 2/2005.

EICKMEIER, S., BREITUNG, J. (2005). How synchronized are central and east European economies with the Euro area? Evidence from a structural factor model. Bundesbank Discussion Paper 20/2005.

FAGAN, G., HENRY, J., MESTRE, R. (2001). An area wide model (AWM) for the Euro area. ECB Working Paper 42.

FAVERO, C., MARCELLINO, M., NEGLIA, F. (2005). Principal components at work: The empirical analysis of monetary policy with large datasets. *Journal of Applied Econometrics* **20** 603–620.

FORNI, M., GIANNONE, D. LIPPI, F., REICHLIN, L. (2004). Opening the Black Box: Structural factor models versus structural VARS. Université Libre de Bruxelles, Mimeo.

FORNI, M., HALLIN, M., LIPPI, F., REICHLIN L. (2000). The generalized dynamic factor model: Identification and estimation. *Review of Economics and Statistics* **82** 540–554..

FORNI, M., HALLIN, M., LIPPI, F., REICHLIN L. (2002). The generalized dynamic factor model: consistency and convergence rates. *Journal of Econometrics* **82** 540–554.

FORNI M., HALLIN, M., LIPPI, F., REICHLIN, L. (2003). Do financial variables help forecasting inflation and real activity in the Euro area?. *Journal of Monetary Economics* **50** 1243–1255.

FORNI M., HALLIN, M., LIPPI, F., REICHLIN, L. (2005). The generalized dynamic factor model: one-sided estimation and forecasting. *Journal of the American Statistical Association* **100** 830–840.

GIACOMINI R., WHITE H. (2003). Tests of conditional predictive ability. Boston College Working Papers in Economics 572, Boston College Department of Economics.

GIANNONE, D., SALA, L., REICHLIN, L. (2002). Tracking Greenspan: Systematic and unsystematic monetary policy revisited. ECARES-ULB, Mimeo.

GIANNONE, D., SALA, L., REICHLIN, L. (2004). Monetary policy in real time. forthcoming in: NBER Macroeconomics Annual.

HELBLING, T., BAYOUMI, T. (2003). Are they all in the same boat? The 2000-2001 growth slowdown and the G7-business cycle linkages. IMF Working Paper, WP/03/46.

JIMÉNEZ-RODRÍGUEZ, M., M. SÁNCHEZ (2005). Oil price shocks and real GDP growth: empirical evidence for some OECD countries. *Applied Economics* **37** 201–228.

JÖRESKOG, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrica* **34** 183–202.

KAPETANIOS, G. (2004). A note on modelling core inflation for the UK using a new dynamic factor estimation method and a large disaggregated price index dataset. *Economics Letters* **85** 63–69.

KAPETANIOS, G., MARCELLINO, M. (2003). A comparison of estimation methods for dynamic factor models of large dimensions. Queen Mary University of London, Working Paper 489.

KORHONEN, I. (2003). Some empirical tests on the integration of economic activity between the Euro area and the accession countries: A note. *Economics of Transition* **11** 1–20.

MALEK MANSOUR, J. (2003). Do national business cycles have an international origin?. *Empirical Economics* **28** 223–247.

MARCELLINO, M., STOCK, J. H., WATSON, M. W. (2000). A dynamic factor analysis of the EMU. IGER Bocconi, Mimeo.

MARCELLINO, M., STOCK, J. H., WATSON, M. W. (2003). Macroeconomic forecasting in the Euro area: vountry-specific versus Euro wide information. *European Economic Review* **47** 1–18.

PEERSMAN, G. (2005). What caused the early millenium slowdown? Evidence based on vector autoregressions. *Journal of Applied Econometrics* **20** 185–207.

SALA, L. (2003). Monetary policy transmission in the Euro area: A factor model approach. IGER Bocconi, Mimeo.

SCHUMACHER, C. (2005). Forecasting German GDP using alternative factor models based on large datasets. Bundesbank Discussion Paper 24/2005.

SCHUMACHER, C., DREGER, C. (2004). Estimating large-scale factor models for economic activity in Germany: Do they outperform simpler models?. *Jahrbücher für Nationalökonomie und Statistik* **224** 731–750.

STOCK, J. H., WATSON, M. W. (1999). Forecasting inflation. *Journal of Monetary Economics* **44** 293–335.

STOCK, J. H., WATSON, M. W. (2002a). Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics* **20** 147–162.

STOCK, J. H., WATSON, M. W. (2002b). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association* **97** 1167–1179.

STOCK, J. H., WATSON, M. W. (2005). Implications of dynamic factor models for VAR analysis. Princeton University, Mimeo.

WATSON, M. W. (2003). Macroeconomic forecasting using many predictors. In *Advances in Economics and Econometrics: Theory and Applications.* (Dewatripont, M., Hansen, L. P., Turnovsky, S. J. eds.), Vol. III, Eighth World Congress, Cambridge University Press, Cambridge.

4 Unit Root Testing*

Jürgen Wolters[1] and Uwe Hassler[2]

[1] Fachbereich Wirtschaftswissenschaft, Freie Universität Berlin
wolters@wiwiss.fu-berlin.de
[2] Fachbereich Wirtschaftswissenschaften, J.W. Goethe Universität Frankfurt
hassler@wiwi.uni-frankfurt.de

Summary. The occurrence of unit roots in economic time series has far reaching consequences for univariate as well as multivariate econometric modelling. Therefore, unit root tests are nowadays the starting point of most empirical time series studies. The oldest and most widely used test is due to Dickey and Fuller (1979). Reviewing this test and variants thereof we focus on the importance of modelling the deterministic component. In particular, we survey the growing literature on tests accounting for structural shifts. Finally, further applied aspects are addressed, for instance, how to get the size correct and obtain good power at the same time.

4.1 Introduction

A wide variety of economic time series is characterized by trending behaviour. This raises the important question how to statistically model the long-run component. In the literature, two different approaches have been used. The so-called trend stationary model assumes that the long-run component follows a time polynomial, which is often assumed to be linear, and added to an otherwise stationary autoregressive moving average (ARMA) process. The difference stationary model assumes that differencing is required to obtain stationarity, i.e. that the first difference of a time series follows a stationary and invertible ARMA process. This implies that the level of the time series has a unit root in its autoregressive (AR) part. Unit root processes are also called integrated of order 1, I(1).

Since the seminal paper by Nelson and Plosser (1982) economists know that modelling the long-run behaviour by trend or difference stationary models has far-reaching consequences for the economic interpretation. In a trend stationary model

*We thank Mu-Chun Wang for producing the figures, and an anonymous referee for comments improving the presentation.

the effects of shocks are only temporary implying that the level of the variable is not influenced in the long run. In contrast a shock has permanent effects in a difference stationary model, meaning that the level of the variable will be shifted permanently after the shock has occurred.

Traditional econometrics assumes stationary variables (constant means and time-independent autocorrelations). This is one of the reasons why applied economists very often transform non-stationary variables into stationary time series. According to the two above-mentioned models this can be done by eliminating deterministic trends in the case of a trend stationary model or by taking first differences in the case of a difference stationary model. But what happens if the wrong transformation is applied? The papers by Chan *et al.* (1977), Nelson and Kang (1981) and Durlauf and Phillips (1988) investigate this problem. Eliminating the non-stationarity in a trend stationary model by taking first differences has two effects: one gets rid of the linear trend, but the stationary stochastic part is overdifferenced, implying spurious short-run cycles. If, on the other hand, it is tried to eliminate the non-stationarity in a difference stationary model by taking the residuals of a regression on a constant and on time as explanatory variables, spurious long-run cycles are introduced. These depend on the number of observations used in the regression. In this case artificial business cycles are produced that lead to wrong economic interpretations.

Moreover, regressing independent difference stationary processes on each other leads to the problem of spurious regressions as Granger and Newbold (1974) have demonstrated in a simulation study. Later on Phillips (1986) gave the theoretical reasoning for this phenomenon: The usual t-statistics diverge to infinity in absolute value, while the R^2 does not converge to zero, hence indicating spurious correlation between independent difference stationary processes. Granger (1981) and Engle and Granger (1987) offered a solution to the spurious regression problem by introducing the concept of cointegration.

The above discussion clearly indicates that the analysis of non-stationary time series requires a serious investigation of the trending behaviour. Therefore, formal tests are needed which allow to distinguish between trend stationary and difference stationary behaviour of time series. Such tests have first been developed by Fuller (1976) and Dickey and Fuller (1979, 1981) (DF test, or augmented DF test, ADF). In the meantime a lot of extensions and generalizations have been published which also are presented in different surveys such as Dickey *et al.* (1986), Diebold and Nerlove (1990), Campbell and Perron (1991), Hassler (1994), Stock (1994) and Phillips and Xiao (1998).

Due to page limitations we will present here only the (augmented) Dickey-Fuller approach for testing the null hypothesis of difference stationarity. The related semi-parametric approach developed by Phillips (1987) and Phillips and Perron (1988) is not presented, and the extension to panel unit root tests is not considered, see Breitung and Pesaran (2005) for a recent overview. Furthermore, we do not deal with tests for seasonal unit roots as proposed e.g. by Hylleberg *et al.* (1990), or tests having stationarity in the maintained hypothesis as Kwiatkowski *et al.* (1992). We rather focus on modelling the deterministic part of the time series under investigation. This is very important in case of structural breaks, since neglecting

deterministic shifts may result in misleading conclusions.

The paper is structured as follows. In Section 2, Dickey-Fuller unit root tests are described and discussed. The third section deals with important applied aspects regarding size and power. Section 4 turns to the handling of structural breaks.

4.2 Dickey-Fuller Unit Root Tests

4.2.1 Model

For the rest of the paper we assume the following data generating process (DGP):

$$y_t = d_t + x_t\,, \quad t = 1, \ldots, T\,. \tag{4.1}$$

The observed variable (y_t) is composed of a deterministic component d_t and a purely stochastic component x_t. The deterministic part may consist of a constant, seasonal dummy variables, a linear trend or a step dummy. The stochastic component is assumed to be a zero mean AR(p) process,

$$x_t = \alpha_1 x_{t-1} + \alpha_2 x_{t-2} + \ldots + \alpha_p x_{t-p} + u_t\,, \tag{4.2}$$

with $\alpha_p \neq 0$ and u_t being white noise. The AR(p) model in (4.2) can be reparameterized as

$$x_t = \rho x_{t-1} + \sum_{i=1}^{p-1} a_i \Delta x_{t-i} + u_t \tag{4.3}$$

with

$$\rho = \sum_{j=1}^{p} \alpha_j \quad \text{and } a_i = -\sum_{j=i+1}^{p} \alpha_j\,, i = 1, \ldots, p-1\,.$$

If the lag polynomial of x_t

$$1 - \alpha_1 z - \alpha_2 z^2 - \ldots - \alpha_p z^p = 0$$

has a unit root, then it holds

$$\rho = \sum_{j=1}^{p} \alpha_j = 1\,. \tag{4.4}$$

Substituting (4.3) in (4.1) we get the following expression for the observable variable y_t:

$$y_t = d_t - \rho d_{t-1} - \sum_{i=1}^{p-1} a_i \Delta d_{t-i} + \rho\, y_{t-1} + \sum_{i=1}^{p-1} a_i \Delta y_{t-i} + u_t\,.$$

Subtracting y_{t-1} on both sides of this equation, we obtain the Augmented Dickey-Fuller (ADF) regression:

$$\Delta y_t = d_t - \rho d_{t-1} - \sum_{i=1}^{p-1} a_i \Delta d_{t-i} + (\rho - 1)\, y_{t-1} + \sum_{i=1}^{p-1} a_i \Delta y_{t-i} + u_t\,. \tag{4.5}$$

Note that in addition to the (lagged) level of the deterministic part its lagged changes are included, too.

In the case that $d_t = c$, a constant, the ADF regression is given as

$$\Delta y_t = a + (\rho - 1)y_{t-1} + \sum_{i=1}^{k} a_i \Delta y_{t-i} + u_t\,, \quad t = k+2, \ldots, T\,, \quad (4.6)$$

with $a = (1-\rho)c$, meaning that under the null hypothesis the process is I(1) without drift. If the stochastic component x_t follows an AR(p) process then $k = p - 1$ in (4.6). More generally, however, x_t may be an ARMA process with invertible moving average component. In that case, Said and Dickey (1984) propose to approximate the ARMA stucture by autoregressions of order k where the lag length k has to grow with the sample size.

In the case of a linear trend, $d_t = c + mt$, we get from (4.5) as ADF regression

$$\Delta y_t = a + bt + (\rho - 1)y_{t-1} + \sum_{i=1}^{k} a_i \Delta y_{t-i} + u_t\,, \quad t = k+2, \ldots, T\,, \quad (4.7)$$

with

$$a = c(1-\rho) + \rho m - m\sum_{i=1}^{k} a_i \text{ and } b = m(1-\rho)\,.$$

Under the null of a unit root ($\rho = 1$) the trend term in (4.7) vanishes, while the constant term contains not only the slope parameter m but also the coefficients a_i of the short run dynamic. Under the null it holds that $E(\Delta y_t) \neq 0$, and hence y_t displays a stochastic trend, I(1), as well as a deterministic one because $E(y_t)$ grows linearly with t. Such series are called integrated with drift.

4.2.2 Distribution

The null hypothesis to be tested is that x_t and hence y_t is integrated of order one. Under the null hypothesis there is a unit root in the AR polynomial and we have because of (4.4):

$$H_0 : \rho - 1 = 0\,.$$

This hypothesis can be tested directly by estimating (4.5) with least squares and using the t-statistic of $\hat{\rho}-1$. It is a one-sided test that rejects in favour of stationarity if $\hat{\rho} - 1$ is significantly negative. The limiting distribution was discovered by Fuller (1976) and Dickey and Fuller (1979). It turned out that it is not centered around zero (but rather shifted to the left) and not symmetrical. In particular, limiting standard normal theory is invalid for $\rho = 1$, while it does apply in case of stationarity ($|\rho| < 1$). Similarly, the t-distribution is not a valid guideline for unit root testing in finite samples. The limiting distribution is very sensitive to the specification of d_t. Fuller (1976) and Dickey and Fuller (1979) consider three cases: No deterministics ($d_t = 0$), just a constant, and a constant and a linear trend. Critical values for these cases have first been provided by simulation in Fuller (1976, Table 8.5.2,

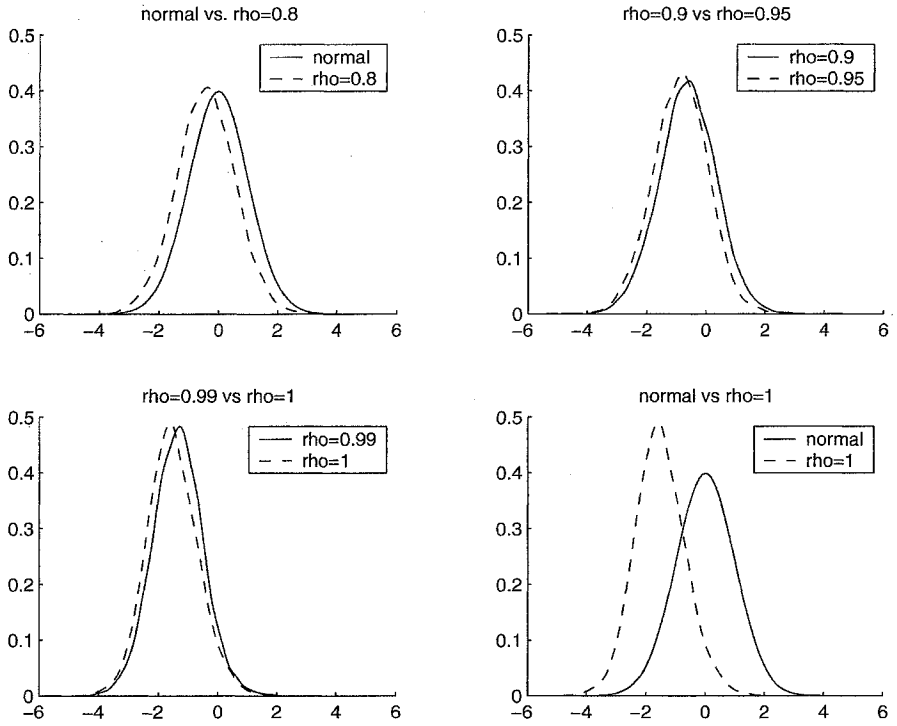


Figure 4.1: *Estimated density functions ($T = 100$) and standard normal.*

p. 373)[1]. Nowadays, somewhat more precise critical values are widely employed, which have been derived using more intensive simulations by MacKinnon (1991, p. 275). For a test with $T = 100$ at the 5 % level the critical values are -2.89 and -3.46, respectively, if d_t contains only a constant and if d_t includes a constant and a linear trend. These values are larger in absolute terms than the corresponding critical value of the t-distribution which is -1.65. Using the incorrect t-distribution the null hypothesis would be rejected much too often. The decision would wrongly be in favour of stationarity or trendstationarity despite the fact that the time series contains an I(1) component.

Theoretically, $\rho = 1$ is a singularity in the parameter space in that for any $|\rho| < 1$ limiting normality holds true, while $\rho = 1$ results in the non-normal Dickey-Fuller distributions. Some economists blame this as being artificial and claim that the true ρ equals 1 with probability zero. According to this, the normal approximation should be applied throughout. In practice however, and that means in finite samples, the distinction between stationarity and I(1) is not so clear-cut[2]. Evans and Savin (1981, p. 763) observed that for ρ '*near but below unity this distribution function is*

[1]Critical values for polynomials in t up to order 5 are derived in Ouliaris *et al.* (1989).

[2]In fact, Phillips (1987a) presented a unifying local-to-unity theory bridging the gap from stationarity to I(1) by introducing a time dependent ρ, $\rho_T = \exp\left(\frac{c}{T}\right)$.

very poorly approximated by the limiting normal even for large values of T". Similar evidence has been collected by Evans and Savin (1984) and Nankervis and Savin (1985). Evans and Savin (1981, 1984) considered a normalization different from the t-statistic, while Nankervis and Savin (1985) considered the usual studentization but did not provide graphs for the t-statistics. Therefore, we want to add the results of a small Monte Carlo study here[3]. The true DGP is (with $y_0 = 0$)

$$y_t = \rho y_{t-1} + u_t, \quad u_t \sim \mathcal{N}(0,1)\,,$$

for $t = 1, ..., T = 100$. Relying on

$$\Delta y_t = \hat{a} + (\hat{\rho} - 1) y_{t-1} + \hat{u}_t\,,$$

the usual t-statistic is computed. Figure 1 displays density estimates that are constructed from 10,000 replications by smoothing with a normal kernel. From these results we learn that even for ρ considerably smaller than 1 the standard normal approximation may be a very bad guideline in finite samples.

4.3 Size and Power Considerations

Since the work by Schwert (1989) it has been documented in several papers that the DF test may be over-sized in situations of practical importance[4]. Hence, proposals how to control for the probability of a type I error have attracted a lot of attention in the last decade. At the same time DF tests are blamed for poor power[5], and many papers tackled the problem to increase power. Several aspects related to those topics are addressed next.

4.3.1 Lag Length Selection

The lag length k in (4.6) and (4.7) has to be chosen to ensure that the residuals empirically follow a white noise process. Said and Dickey (1984) prove that the ADF test in (4.5) is valid if the true DGP is an ARMA process of unknown order, provided that the lag length k in the autoregression increases with the sample size but at a lower rate. A proof under more general conditions was recently provided by Chang and Park (2002). In practice, the choice of k is a crucial and difficult exercise. On the one hand, a growing number of lags reduces the effective sample while the number of estimated parameters is increased, and this reduction in degrees of freedom will result in a loss of power. On the other hand, k has to be large enough for the residuals to be approximately uncorrelated in order for the limiting theory to be

[3] All programming was done by Mu-Chun Wang in MATLAB.

[4] Kim and Schmidt (1993) established experimentally that conditionally heteroskedastic errors have little effect on DF tests as long as there is only moderate heteroskedasticity. But Valkanov (2005) showed that with strongly heteroskedastic data the use of asymptotic critical DF values leads to grossly oversized tests.

[5] Gonzalo and Lee (1996), in contrast, illustrated by means of Monte Carlo experiments that testing for $\rho = 1$ results in rejection frequencies very similar to those available if $|\rho| < 1$. In case of fractionally integrated alternatives, however, Hassler and Wolters (1994) showed that the ADF test has little power.

valid. Empirical researchers often start with a maximum lag length k_{max} and follow a sequential general-to-specific strategy, i.e. reduce lags until reaching significance at a prespecified level. Here, significance testing builds on a limiting standard normal distribution. Alternatively, k may be determined relying on information criteria. The performance of the two strategies in finite samples has been investigated by Ng and Perron (1995). In particular, information criteria tend to choose k too small to get the size correct. Therefore, Ng and Perron (2001) proposed adequately modified criteria.

The AR approximation is particularly poor in case of moving average roots close to one. Consider as DGP

$$\Delta y_t = a + u_t - \theta u_{t-1}, \quad |\theta| < 1 .$$

With θ close to one the polynomial $1 - \theta L$ almost cancels with $\Delta = 1 - L$, and the true null of integration will be rejected frequently, resulting in a test where the empirical size is above the nominal one, cf. Schwert (1989). In such a situation one may use the procedure proposed by Said and Dickey (1985) that explicitly takes into account the MA component of a series.

4.3.2 Deterministic Components

When performing unit root tests an appropriate specification of the deterministics in (4.5) is of crucial importance. First, consider the case where the true DGP is trend stationary. If the ADF regression without detrending is applied, then the test has asymptotically no power, which was shown by West (1987) and Perron (1988). In finite samples the null hypothesis of a unit root is rarely rejected, and is never rejected in the limit. Hence, we propose to include a linear trend as in (4.7) whenever a series is suspicious of a linear trend upon visual inspection[6]. Notice that the decision about time as regressor may not build on the standard t-statistic of the estimate $\hat{b}$. Second, assume the other way round that a detrended test is performed from (4.7) while the data does not contain a linear trend. In this situation a test from (4.6) without detrending would be more powerful. The effect of ignoring eventual mean shifts in the DGP when specifying d_t will be discussed in the following section, while the effect of neglected seasonal deterministics will be touched upon in the next subsection.

The treatment of the deterministic component plays a major role when it comes to power of unit root tests. Elliott *et al.* (1996) and Hwang and Schmidt (1996) proposed point optimal unit root tests with maximum power against a given local alternative $\rho_T = 1 - \frac{c}{T}$ for some specified constant $c > 0$. Power gains are obtained by efficiently removing the deterministic component under the alternative (using Generalized Least Squares, GLS). Use of GLS, however, amounts to the following procedure, see also Xiao and Phillips (1998). First, compute quasi-differences of the observed variable and the deterministic regressors,

$$\Delta_c y_t = y_t - \left(1 - \frac{c}{T}\right) y_{t-1}, \quad \Delta_c z_t = z_t - \left(1 - \frac{c}{T}\right) z_{t-1},$$

[6] Ayat and Burridge (2000) investigated a more rigorous sequential procedure to determine the appropriate deterministics when testing for unit roots.

where z_t is a deterministic vector such that d_t from (4.1) is parameterized as $d_t = \gamma' z_t$. Second, estimate the vector γ by simply regressing $\Delta_c y_t$ on $\Delta_c z_t$. Third, apply the ADF test without deterministics to the residuals from the second step. The distribution and hence critical values depend on the choice of c. Yet another approach to obtain more powerful unit root tests has been advocated by Shin and So (2001). They proposed to estimate γ by $\hat{\gamma}_t$ with information only up to t, and remove the deterministic component from y_t as follows: $\tilde{y}_t = y_t - \hat{\gamma}'_{t-1} z_{t-1}$. Applying an ADF type regression to $\tilde{y}_t$ results in a limiting distribution again different from Dickey-Fuller. Critical values have been provided by Shin and So (2001, Table II) for the simplest case of a constant where $z_t = 1$.

Leybourne *et al.* (2005) explored both asymptotic and finite-sample properties of five more powerful modifications of the DF test. They favoured two tests studied by Taylor (2002) and Leybourne (1995). The latter relies on the usual DF statistic and the test statistic computed from the reversed time series. The test proposed by Taylor (2002) builds on recursive detrending by ordinary least squares.

4.3.3 Span vs. Frequency

In practice, it often happens that data are available only for a fixed time span due to some structural breaks or institutional changes. In such a situation it is often recommended to use data with higher frequency to increase the number of observations and hence the power of tests. To that end people often work with monthly data instead of quarterly or annual observations. Perron (1989) has analyzed the power of some unit root tests when the sampling interval is varied but the time span is hold fixed. A general outcome of his computer experiments is that tests over short time span have low power, which is not significantly enhanced by choosing a shorter sampling interval. For related results see also Shiller and Perron (1985).

Moreover, in many cases higher frequency of observations comes at the price of additional seasonal dynamics that have to be modelled. In case of deterministic seasonal patterns it is important to remove the seasonality by including seasonal dummies in the regression. Dickey *et al.* (1986) prove that the inclusion of seasonal dummies instead of a constant does not affect the limiting distribution of DF tests, while Demetrescu and Hassler (2005) demonstrate that neglecting seasonal deterministic results in tests with low power and bad size properties at the same time. Another way of removing seasonal deterministics is simply to work with seasonal differences, which, however, can not be recommended in general. Hassler and Demetrescu (2005) argue that seasonal differencing may introduce artificial persistence into a time series and may hence create spurious unit roots.

Given a fixed time span of data the purpose of unit root testing is not to investigate the true nature of some abstract economic process but to describe the degree of persistence in a given sample. Even if difference stationarity is not a plausible theoretical model as $T \to \infty$ for economic series such as inflation rates, interest rates or unemployment rates, the unit root hypothesis may still provide an empirically valid description. In that sense the significance against $H_0 : \rho = 1$ may be understood as strength of mean-reversion in a given sample. Similarly, Juselius (1999, pp. 264) argues that '*the order of integration of a variable is not in general a property of an economic variable but a convenient statistical approximation to distinguish between*

the short-run, medium-run and long-run variation in the data'.

4.4 Structural Breaks

4.4.1 Ignoring Breaks

Consider for the moment a regression with a constant only,

$$\Delta y_t = \hat{a} + (\hat{\rho} - 1) y_{t-1} + \hat{u}_t \,. \tag{4.8}$$

It is now assumed that the regression (4.8) is misspecified: The true process is I(0), but it displays a break in the mean at time λT,

$$y_t = \begin{cases} x_t \,, & t < \lambda T \\ x_t + \mu \,, & t \geq \lambda T \end{cases} \,, \quad x_t \sim I(0), \tag{4.9}$$

where $\lambda \in (0,1)$, and $\mu \neq 0$. Given (4.9), neither the null nor the alternative hypothesis of the DF test holds true. Perron (1990) proves that $\hat{\rho}$ converges to a value that approaches 1 as the break $|\mu| > 0$ is growing. A corresponding result is found in Perron (1989a) in case of the detrended version of the DF test,

$$\Delta y_t = \hat{a} + \hat{b}\, t + (\hat{\rho} - 1) y_{t-1} + \hat{u}_t \,. \tag{4.10}$$

This means that for a considerable break $\mu \neq 0$ the wrong null of a unit root is hardly rejected. A high probability of a type II error arises because the DF regression is misspecified in that it does not account for the structural break in the data. See also the intuitive discussion in Rappoport and Reichlin (1989). Moreover, Perron (1989a) investigates the situation of trend stationary series with a break:

$$y_t = \begin{cases} x_t + \delta\, t \,, & t < \lambda T \\ x_t \,(+\mu) + (\delta + \tau)\, t \,, & t \geq \lambda T \end{cases} \,, \quad x_t \sim I(0) \,.$$

For $\tau \neq 0$ we have a break in the slope of the trend, while there may be an additional shift in the level ($\mu \neq 0$) or not. With this assumption Perron (1989a) investigates $\hat{\rho}$ from the detrended DF test (4.10). His asymptotic formulae were corrected by Montañés and Reyes (1998), but without changing the empirically relevant fact: Given a trend stationary series with a break in the linear trend there is little chance to reject the false null hypothesis of integration. Those results opened a new research avenue aiming at the discrimination between unit roots and structural shifts as potential causes of economic persistence.

Quite surprisingly, the opposite feature to that discovered by Perron (1989a, 1990) has been established by Leybourne *et al.* (1998a): If a unit root process is subject

to a mean shift, then the probability of rejecting the null of a unit root is not equal to the level of the test[7]. The authors assume

$$y_t = \begin{cases} x_t\,, & t < \lambda T \\ x_t + \mu\,, & t \geq \lambda T \end{cases}\,, \quad x_t \sim I(1)\,,$$

with $\mu \neq 0$. Leybourne *et al.* (1998a) prove that the limiting distribution of the DF statistic depends on λ if μ is growing with T . Experimentally, they establish that the empirical level is well above the nominal one in case of an early break, $\lambda \leq 0.1$; if, however, the break occurs in the second half of the sample, then the DF test is conservative in the sense that the rejection frequency is below the nominal level. More generally, Kim *et al.* (2004) have shown that the results of the ADF test (4.6) including only a constant term is highly unpredictable, if the true deterministic is a broken trend.

4.4.2 Correcting for Breaks

To avoid spurious unit roots due to structural breaks, Perron (1989a, 1990) suggested to test for integration after removing structural breaks[8]. Unfortunately, this changes the limiting distributions depending on the break fraction $\lambda \in (0, 1)$. To correct for a break in the level we need the step dummy

$$s_t(\lambda) = \begin{cases} 0\,, & t < \lambda T \\ 1\,, & t \geq \lambda T \end{cases}\,.$$

Now, all the deterministic components are removed from the observed series in a first step by an OLS regression[9],

$$y_t = \tilde{a} + \tilde{\mu}\, s_t(\lambda)\ (+\tilde{b}\, t\,) + \tilde{x}_t\,. \tag{4.11}$$

Next, the zero mean residuals $\tilde{x}_t$ are tested for a unit root. However, the validity of the asymptotic percentiles requires the inclusion of the impulse dummy

$$\Delta s_t(\lambda) = s_t(\lambda) - s_{t-1}(\lambda)\,.$$

The necessity to include (lagged values of) $\Delta s_t(\lambda)$ has been recognized only by Perron and Vogelsang (1992, 1993) although it is well motivated by (4.5):

$$\Delta\tilde{x}_t = (\hat{\rho} - 1)\tilde{x}_{t-1} + \sum_{i=1}^{k} \hat{a}_i \Delta\tilde{x}_{t-i} + \sum_{i=0}^{k} \hat{\alpha}_i \Delta s_{t-i}(\lambda) + \hat{u}_t\,. \tag{4.12}$$

[7] This, however, is only true in finite samples or if the break is growing with the number of observations. If in contrast the break is finite, then the level of the DF test is not affected asymptotically, see Amsler and Lee (1995). Still, power considerations suggest a correction for potential breaks.

[8] Perron (1994) provides a very accessible survey.

[9] Lütkepohl *et al.* (2001) and Saikkonen and Lütkepohl (2001) suggested to remove the deterministic components efficiently in the sense of Elliott *et al.* (1996). This approach of working with quasi-differences has the advantage of yielding limiting distributions independent of λ. The same property has the test by Park and Sung (1994).

Critical values when testing for $\rho = 1$ in (4.12) are found in Perron (1989a, 1990). They depend on the break fraction λ that is assumed to be known. Similarly, Perron (1989a) considered modifications of the detrended DF test allowing for a shift in the slope of the linear time trend.

If λ is not known the break point can be estimated from the data. Zivot and Andrews (1992) e.g. suggested to vary λ and to compute the test statistic $ADF(\lambda)$ for each regression. Then the potential break point can be determined as $\hat{\lambda} = \arg\min_\lambda ADF(\lambda)$. Confer also Banerjee *et al.* (1992) and Christiano (1992).

4.4.3 Smooth Transitions and Several Breaks

When defining the step dummy $s_t(\lambda)$ we assumed a sudden change at $t = \lambda T$. But even if the cause of a change occurs instantaneously, its effect most likely evolves gradually over a period of transition. To account for that effect Perron (1989a) proposed the so-called innovational outlier model, which assumes that *'the economy responds to a shock to the trend function the same way as it reacts to any other shock'*, Perron (1989a, p. 1380). This amounts to adding a step dummy variable to the augmented Dickey-Fuller regression instead of applying the ADF test after removing all deterministics. Leybourne *et al.* (1998b) considered unit root testing in the presence of more general deterministic smooth transition functions, see also Lin and Teräsvirta (1994), however without providing asymptotic theory. Limiting results under smooth transitions have been established in Saikkonen and Lütkepohl (2001) for known breakpoint, and in Saikkonen and Lütkepohl (2002) in case the date of the break is not known a priori. For a comparison of related tests see also Lanne *et al.* (2002).

Lumsdaine and Papell (1997) allowed for two break points where both break dates are assumed to be unknown. Park and Sung (1994) dealt with the case of several breaks, however at known time. Kapetanios (2005) combined both features, i.e. more than two breaks occuring at unknown break points. Our personal opinion, however, is that the data-driven estimation of m break dates for a given series should not be a recommended strategy unless it is possible to reveal an economic event or institutional change behind each eventual break.

References

Ajat, L., Burridge, P. (2000). Unit root tests in the presence of uncertainty about the non-stochastic trend. *Journal of Econometrics* **95** 71–96.

Amsler, Ch., Lee, J. (1995). An LM test for a unit root in the presence of a structural change. *Econometric Theory* **11** 359–368.

Banerjee, A., Lumsdaine, R. L., Stock, J. H. (1992). Recursive and sequential tests of the unit-root and trend-break hypothesis: Theory and international evidence. *Journal of Business & Economic Statistics* **10** 271–287.

Breitung, J., Pesaran, H. (2005). Unit roots and cointegration in panel data. University of Bonn, Mimeo.

CAMPBELL, J. Y., PERRON, P. (1991). Pitfalls and opportunities: What macroeconomists should know about unit roots. In *NBER Marcoeconomics Annual 1991* (O.J. Blanchard, S. Fisher, eds.), 141–201. MIT Press, Cambridge.

CHAN, K. H., HAYYA, J. C., ORD, J. K. (1977). A note on trend removal methods: The case of polynomial regressions versus variate differencing. *Econometrica* **45** 737–744.

CHANG, Y., PARK, J. Y. (2002). On the asymptotics of ADF tests for unit roots. *Econometric Reviews* **21** 431–447.

CHRISTIANO, L. J. (1992). Searching for a break in GNP. *Journal of Business & Economic Statistics* **10** 237–250.

DEMETRESCU, M., HASSLER, U. (2005). Effect of neglected deterministic seasonality on unit root tests. *Statistical Papers* (forthcoming).

DICKEY, D. A., FULLER, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association* **74** 427–431.

DICKEY, D. A., FULLER, W. A. (1981). Likelihood ratio statistics for autoregressive time series with a unit root. *Econometrica* **49** 1057–1072.

DICKEY, D. A., BELL, W. R., MILLER, R. B. (1986). Unit roots in time series models: Tests and implications. *American Statistician* **40** 12–26.

DIEBOLD, F. X., NERLOVE, M. (1990). Unit roots in economic time series: A selective survey. In *Advances in Econometrics: Cointegration, Spurious Regressions, and Unit Roots* (T. B. Fomby, G. F. Rhodes, eds.), 3–70. JAI Press, Greenwich.

DURLAUF, S., PHILLIPS. P. C. B. (1988). Trends versus random walks in time series analysis. *Econometrica* **56** 1333–1354.

ELLIOTT, G., ROTHENBERG, TH. J., STOCK, J. H. (1996). Efficient tests for an autoregressive unit root. *Econometrica* **64** 813–836.

ENGLE, R. F., GRANGER, C. W. J. (1987). Co-integration and error correction: Representation, estimation, and testing. *Econometrica* **55** 251–276.

EVANS, G. B. A., SAVIN, N. E. (1981). Testing for unit roots: 1. *Econometrica* **49** 753–779.

EVANS, G. B. A., SAVIN, N. E. (1984). Testing for unit roots: 2. *Econometrica* **52** 1241–1270.

FULLER, W. A. (1976). *Introduction to Statistical Time Series.* Wiley, New York.

GONZALO, J., LEE, T.-H. (1996). Relative power of t type tests for stationary and unit root processes. *Journal of Time Series Analysis* **17** 37–47.

GRANGER, C. W. J. (1981). Some properties of time series data and their use in econometric model specification. *Journal of Econometrics* **16** 121–130.

GRANGER, C. W. J., NEWBOLD, P. (1974). Spurious regressions in econometrics. *Journal of Econometrics* **2** 111–120.

HASSLER, U. (1994). Einheitswurzeltests - Ein Überblick. *Allgemeines Statistisches Archiv* **78** 207–228.

HASSLER, U., DEMETRESCU, M. (2005). Spurious persistence and unit roots due to seasonal differencing: The case of inflation rates. *Journal of Economics and Statistics* **225** 413–426.

HASSLER, U., WOLTERS, J. (1994). On the power of unit root tests against fractional alternatives. *Economics Letters* **45** 1–5.

HWANG, J., SCHMIDT, P. (1996). Alternative methods of detrending and the power of unit root tests. *Journal of Econometrics* **71** 227–248.

HYLLEBERG, S., ENGLE, R. F., GRANGER, C. W. J., YOO, B. S. (1990). Seasonal integration and cointegration. *Journal of Econometrics* **44** 215–238.

JUSELIUS, K. (1999). Models and relations in economics and econometrics. *Journal of Economic Methodology* **6** 259–290.

KAPETANIOS, G. (2005). Unit-root testing against the alternative hypothesis of up to m structural breaks. *Journal of Time Series Analysis* **26** 123–133.

KIM, K., SCHMIDT, P. (1993). Unit roots tests with conditional heteroskedasticity. *Journal of Econometrics* **59** 287–300.

KIM, T. W., LEYBOURNE, S., NEWBOLD, P. (2004). Behaviour of Dickey-Fuller unit-root tests under trend misspecification. *Journal of Time Series Analysis* **25** 755–764.

KWIATKOWSKI, D., PHILLIPS, P. C. B., SCHMIDT, P., SHIN, Y. (1992). Testing the null hypothesis of stationarity against the alternative of unit root. *Journal of Econometrics* **54** 159–178.

LANNE, M., LÜTKEPOHL, H., SAIKKONEN, P. (2002). Comparison of unit root tests for time series with level shifts. *Journal of Time Series Analysis* **23** 667–685.

LEYBOURNE, S. J (1995). Testing for unit roots using forward and reverse Dickey-Fuller regressions. *Oxford Bulletin of Economics and Statistics* **57** 559–571.

LEYBOURNE, S. J., KIM, T. W., NEWBOLD, P. (2005). Examination of some more powerful modifications of the Dickey-Fuller test. *Journal of Time Series Analysis* **26** 355–369.

LEYBOURNE, S. J., MILLS, T. C., NEWBOLD, P. (1998a). Spurious rejections by Dickey-Fuller tests in the presence of a break under the null. *Journal of Econometrics* **87** 191–203.

LEYBOURNE, S., NEWBOLD, P., VOUGAS, D. (1998b). Unit roots and smooth transitions. *Journal of Time Series Analysis* **19** 83–97.

LIN, C. J., TERÄSVIRTA, T. (1994). Testing the constancy of regression parameters against continuous structural change. *Journal of Econometrics* **62** 211–228.

LUMSDAINE, R. L., PAPELL, D. H. (1997). Multiple trend breaks and the unit-root hypothesis. *The Review of Economics and Statistics* **79** 212–218.

LÜTKEPOHL, H., MÜLLER, C., SAIKKONEN, P. (2001). Unit root tests for time series with a structural break when the break point is known. In *Nonlinear Statistical Modeling: Essays in Honor of Takeshi Amemiya* (C. Hsiao, K. Morimune, J. L. Powell, eds.) 327–348. Cambridge University Press, Cambridge.

MACKINNON, J. G. (1991). Critical Values for Co-Integration Tests. In *Long-Run Economic Relationships* (R. F. Engle and C. W. J. Granger, eds.), 267–276. Oxford University Press, Oxford.

MONTAÑÉS, A., REYES, M. (1998). Effect of a shift in the trend function on the Dickey-Fuller unit root test. *Econometric Theory* **14** 355–363.

NANKERVIS, J. C., SAVIN, N. E. (1985). Testing the autoregressive parameter with the t statistic. *Journal of Econometrics* **27** 143–161.

NELSON, C. R., KANG, H. (1981). Spurious periodicity in inappropriately detrended time series. *Econometrica* **49** 741–751.

NELSON, C. R., PLOSSER, C. I. (1982). Trends and random walks in macro-economic time series: Some evidence and implications. *Journal of Monetary Economics* **10** 139–162.

NG, S., PERRON, P. (1995). Unit root tests in ARMA models with data-dependent methods for the selection of the truncation lag. *Journal of the American Statistical Association* **90** 268–281.

NG, S., PERRON, P. (2001). Lag length selection and the construction of unit root tests with good size and power. *Econometrica* **69** 1519–1554.

OULIARIS, S., PARK, J. Y., PHILLIPS, P. C. B. (1989). Testing for a unit root in the presence of a maintained trend. In *Advances in Econometrics and Modelling* (B. Raj, ed.), 7–28. Kluwer Academic Publishers, Dordrecht.

PARK, J. Y., SUNG, J. (1994). Testing for unit roots in models with structural change. *Econometric Theory* **10** 917–936.

PERRON, P. (1988). Trends and random walks in macroeconomic time series: Further evidence from a new approach. *Journal of Economic Dynamics and Control* **12** 297–332.

PERRON, P. (1989). Testing for a random walk: A simulation experiment of power when the sampling interval is varied. In *Advances in Econometrics and Modelling* (B. Raj, ed.), 47–68. Kluwer Academic Publishers, Dordrecht.

PERRON, P. (1989a). The great crash, the oil price shock, and the unit root hypothesis. *Econometrica* **57** 1362–1401.

PERRON, P. (1990). Testing for a unit root in a time series with a changing mean. *Journal of Business & Economic Statistics* **8** 153–162.

PERRON, P. (1994). Trend, unit root and structural change in macroeconomic time series. In *Cointegration for the Applied Economist* (B. B. Rao, ed.), 113–146. St. Martin's Press, New York.

PERRON, P., VOGELSANG, T. J. (1992). Testing for a unit root in a time series with a changing mean: Corrections and Extensions. *Journal of Business & Economic Statistics* **10** 467–470.

PERRON, P., VOGELSANG, T. J. (1993). Erratum. *Econometrica* **61** 248–249.

PHILLIPS, P. C. B. (1986). Understanding spurious regressions in econometrics. *Journal of Econometrics* **33** 311–340.

PHILLIPS, P. C. B. (1987). Time series regression with a unit root. *Econometrica* **55** 277–301.

PHILLIPS, P. C. B. (1987a). Towards a unified asymptotic theory for autoregression. *Biometrika* **74** 535–47.

PHILLIPS, P. C. B., PERRON, P. (1988). Testing for a unit root in time series regression. *Biometrika* **75** 335–346.

PHILLIPS, P. C. B., XIAO, Z. (1998). A primer on unit root testing. *Journal of Economic Surveys* **12** 423–469.

RAPPOPORT, P., REICHLIN, L. (1989). Segmented trends and non-stationary time series. *The Economic Journal* **99** 168–177.

SAID, S. E., DICKEY, D. A. (1984). Testing for unit roots in ARMA(p,q)-models with unknown p and q. *Biometrika* **71** 599–607.

SAID, S. E., DICKEY, D. A. (1985). Hypothesis testing in ARIMA(p, 1, q) models. *Journal of the American Statistical Association* **80** 369–374.

SAIKKONEN, P., LÜTKEPOHL, H. (2001). Testing for unit roots in time series with level shifts. *Allgemeines Statistisches Archiv* **85** 1–25.

SAIKKONEN, P., LÜTKEPOHL, H. (2002). Testing for a unit root in a time series with a level shift at unknown time. *Econometric Theory* **18** 313–348.

SCHWERT, G. W. (1989). Tests for unit roots: A Monte Carlo investigation. *Journal of Business & Economic Statistics* **7** 147–158.

SHILLER, R. J., PERRON, P. (1985). Testing the random walk hypothesis. *Economics Letters* **18** 381–386.

SHIN, D. W., SO, B. S. (2001). Recursive mean adjustment for unit root tests. *Journal of Time Series Analysis* **22** 595–612.

STOCK, J. H. (1994). Unit roots, structural breaks and trends. In *Handbook of Econometrics, Volume IV* (R. F. Engle, D. L. McFadden, eds.), 2739–2841. Elsevier, Amsterdam.

TAYLOR, A. M. R. (2002). Regression-based unit root tests with recursive mean adjustment for seasonal and non-seasonal time series. *Journal of Business & Economic Statistics* **20** 269–281.

WEST, K. D. (1987). A note on the power of least squares tests for a unit root. *Economics Letters* **24** 249–252.

VALKANOV, R. (2005). Functional central limit theorem approximations and the distribution of the Dickey-Fuller test with strongly heteroskedastic data. *Economics Letters* **86** 427–433.

XIAO, Z., PHILLIPS, P. C. B. (1998). An ADF coefficient test for a unit root in ARMA models of unknown order with empirical applications to US economy. *Econometrics Journal* **1** 27–44.

ZIVOT, E., ANDREWS, D. W. K. (1992). Further evidence on the great crash, the oil-price shock, and the unit-root hypothesis. *Journal of Business & Economic Statistics* **10** 251–270.

5 Autoregressive Distributed Lag Models and Cointegration*

Uwe Hassler[1] and Jürgen Wolters[2]

[1] Fachbereich Wirtschaftswissenschaften, J.W. Goethe Universität Frankfurt
hassler@wiwi.uni-frankfurt.de
[2] Fachbereich Wirtschaftswissenschaft, Freie Universität Berlin
wolters@wiwiss.fu-berlin.de

Summary. This paper considers cointegration analysis within an autoregressive distributed lag (ADL) framework. First, different reparameterizations and interpretations are reviewed. Then we show that the estimation of a cointegrating vector from an ADL specification is equivalent to that from an error-correction (EC) model. Therefore, asymptotic normality available in the ADL model under exogeneity carries over to the EC estimator. Next, we review cointegration tests based on EC regressions. Special attention is paid to the effect of linear time trends in case of regressions without detrending. Finally, the relevance of our asymptotic results in finite samples is investigated by means of computer experiments. In particular, it turns out that the conditional EC model is superior to the unconditional one.

5.1 Introduction

The autoregressive distributed lag model (ADL) is the major workhorse in dynamic single-equation regressions. One particularly attractive reparameterization is the error-correction model (EC). Its popularity in applied time series econometrics has even increased, since it turned out for nonstationary variables that cointegration is equivalent to an error-correction mechanism, see Granger's representation theorem in Engle and Granger (1987). By differencing and forming a linear combination of the nonstationary data, all variables are transformed equivalently into an EC model

*We thank Vladimir Kuzin for excellent research assistance and Surayyo Kabilova for skillful word processing. Moreover, we are grateful to an anonymous referee for clarifying comments.

with stationary series only.

Working on feedback control mechanisms for stabilization policy, Phillips (1954, 1957) introduced EC models to economics. Sargan (1964) used them to estimate structural equations with autocorrelated residuals, and Hendry popularized their use in econometrics in a series of papers[1]. According to Hylleberg and Mizon (1989, p. 124) *'the error correction formulation provides an excellent framework within which it is possible to apply both the data information and the information available from economic theory'*. A survey on specification, estimation and testing of EC models is given by Alogoskoufis and Smith (1995). The present paper contributes to this literature in that it treats some aspects of testing cointegration and asymptotic normal inference of the cointegrating vector estimated from an EC format.

The rest of the paper is organized as follows. The next section reviews different reparameterizations and interpretations of ADL models. Then we use that the cointegrating vector computed from the ADL model is equivalent to the one estimated from EC in order to use results by Pesaran and Shin (1998) on asymptotic normality. Section 4 turns to cointegration testing from EC regressions. We review t-type and F-type test statistics, and pay particular attention to the role of linear time trends. The relevance of our asymptotic results in finite samples is investigated through Monte Carlo experiments in Section 5. A detailled summary is contained in the final section.

5.2 Assumptions and Representations

The autoregressive distributed lag model of order p and n, ADL(p,n), is defined for a scalar variable y_t as

$$y_t = \sum_{i=1}^{p} a_i \, y_{t-i} + \sum_{i=0}^{n} c_i' \, x_{t-i} + \varepsilon_t \,, \tag{5.1}$$

where ε_t is a scalar zero mean error term and x_t is a K-dimensional column vector process. Typically, a constant is included in (5.1), which we neglect here for brevity. The coefficients a_i are scalars while c_i' are row vectors. Using the lag operator L applied to each component of a vector, $L^k \, x_t = x_{t-k}$, it is convenient to define the lag polynomial $a(L)$ and the vector polynomial $c(L)$,

$$\begin{aligned} a(L) &= 1 - a_1 L - \ldots - a_p \, L^p \,, \\ c(L) &= c_0 + c_1 L + \ldots + c_n \, L^n \,. \end{aligned}$$

Now, it is straightforward to write (5.1) more compactly:

$$a(L) y_t = c'(L) x_t + \varepsilon_t \,.$$

[1] Davidson *et al.* (1978), Hendry (1979), and Hendry *et al.* (1984). It is noteworthy that A.W. Phillips, Sargan as well as Hendry were professors at the London School of Economics. A personal view on the history of EC models is given in the interview of Hendry by Ericsson (2004).

In order to obtain dynamic stability, it is maintained that

$$a(z) = 0 \quad \Rightarrow \quad |\, z \,| > 1 \quad \text{for } z \in \mathbb{C}\,. \tag{5.2}$$

Under this condition there exists an absolutely summable infinite expansion of the inverted polynomial $a^{-1}(L)$:

$$a^{-1}(L) = \frac{1}{a(L)} = \sum_{j=0}^{\infty} a_j^* L^j, \quad \sum_{j=0}^{\infty} |a_j^*| < \infty\,.$$

Invertibility of $a(L)$ hence yields the following representation:

$$y_t = \frac{c'(L)}{a(L)}\, x_t + e_t\,, \quad a(L)\, e_t = \varepsilon_t\,,$$

where e_t has a stable autoregressive structure of order p. Expanding $a^{-1}(L)$ provides an infinite distributed lag representation,

$$y_t = \left(\sum_{j=0}^{\infty} a_j^* L^j\right)\left(\sum_{j=0}^{n} c_j\, L^j\right)' x_t + e_t = \sum_{j=0}^{\infty} b_j'\, x_{t-j} + e_t\,, \tag{5.3}$$

where b_j are the vectors of dynamic multipliers derived by the method of indetermined coefficients. The vector of long-run multipliers of the ADL(p, n) model may therefore be easily computed from:

$$\beta := \frac{c(1)}{a(1)} = \sum_{j=0}^{\infty} b_j\,. \tag{5.4}$$

It is worth mentioning that (5.1) is suitable for estimation but in order to obtain an economic interpretation of the parameters one has to consider a transformation like (5.3).

Different reparameterizations have been discussed in the literature, see e.g. Wickens and Breusch (1988). By re-arranging the x's one obtains with $\Delta = 1 - L$:

$$y_t = \sum_{i=1}^{p} a_i y_{t-i} + a(1)\beta' x_t - \sum_{i=0}^{n-1}\left(\sum_{j=i+1}^{n} c_j\right)' \Delta x_{t-i} + \varepsilon_t\,, \tag{5.5}$$

where y_t is related to its own past, to contemporaneous x_t and differences Δx_{t-i}. The use of this specification has been suggested for cointegration analysis by Pesaran and Shin (1998). A further variant relates y_t to x_t and differences of both variables. By subtracting $\left(\sum_{i=1}^{p} a_i\right) y_t$ and re-normalizing, (5.5) yields:

$$y_t = \frac{-1}{a(1)} \sum_{i=0}^{p-1}\left(\sum_{j=i+1}^{p} a_j\right) \Delta y_{t-i} + \beta' x_t - \frac{1}{a(1)} \sum_{i=0}^{n-1}\left(\sum_{j=i+1}^{n} c_j\right)' \Delta x_{t-i} + \varepsilon_t\,.$$

This representation due to Bewley (1979) has the advantage that the long-run multipliers β are the coefficients of x_t. However, the contemporaneous Δy_t on the

right-hand side is correlated with ε_t, which renders OLS invalid. Nevertheless, the use of $y_{t-1}, \ldots, y_{t-p-1}$ and $x_t, \ldots, x_{t-n+1}$ as instruments allows for consistent instrumental variable estimation.

One further transformation will turn out to be fruitful for cointegration testing and estimation. Notice that

$$\sum_{i=1}^{p} a_i y_{t-i} - y_{t-1} = -a(1) y_{t-1} - \sum_{i=1}^{p-1} \left(\sum_{j=i+1}^{p} a_j \right) \Delta y_{t-i} \,.$$

Using this result and $x_t = x_{t-1} + \Delta x_t$, (5.5) yields the error-correction format:

$$\begin{aligned} \Delta y_t \;=\; & -a(1) \left(y_{t-1} - \beta' x_{t-1} \right) - \sum_{i=1}^{p-1} \left(\sum_{j=i+1}^{p} a_j \right) \Delta y_{t-i} \\ & + \left(a(1)\beta - \sum_{j=1}^{n} c_j \right)' \Delta x_t - \sum_{i=1}^{n-1} \left(\sum_{j=i+1}^{n} c_j \right)' \Delta x_{t-i} + \varepsilon_t \,. \end{aligned}$$

The interpretation relies on a long-run equilibrium relation, $y = \beta' x$. The error-correction mechanism is the adjustment of y_t via $a(1)$ to equilibrium deviations in the previous period, $y_{t-1} - \beta' x_{t-1}$. In the following, this equation will often be rewritten as

$$\Delta y_t = \gamma\, y_{t-1} + \theta' x_{t-1} + \sum_{i=1}^{p-1} \alpha_i \Delta y_{t-i} + \sum_{i=0}^{n-1} \phi_i' \Delta x_{t-i} + \varepsilon_t \,, \tag{5.6}$$

where

$$\gamma = -a(1), \quad \theta = a(1)\,\beta = -\gamma\,\beta \,, \tag{5.7}$$

and α_i as well as ϕ_i are defined in an obvious manner.

Since the work by Engle and Granger (1987), cointegration of nonstationary processes is known to be equivalent to a data generating error-correction process. For the rest of the paper we assume that y_t and x_t are integrated of order one, I(1), i.e. differencing is required to obtain stationarity. If there exists a linear combination of the nonstationary processes, $y_t - \beta' x_t$, $\beta \neq 0$, which is stationary, then y_t and x_t are called cointegrated. The cointegration rank is at most one, and x_t does not adjust towards equilibrium.

Assumption: (i) The vector $(y_t, x_t')'$ of length $K+1$ is $I(1)$. (ii) The vector x_t alone is not cointegrated. (iii) In case of cointegration, x_t does not adjust to past equilibrium deviations $(y_{t-1} - \beta'\, x_{t-1})$.

Further, we assume a correctly specified error-correction equation in the following sense.

Assumption: (i) The errors ε_t are serially independent with variance σ^2, $\varepsilon_t \sim iid(0, \sigma^2)$. (ii) The errors are uncorrelated with $\Delta\, x_{t+h}$, for all $h \in \mathbb{Z}$.

These assumptions summarize (A1) through (A5) in Pesaran and Shin (1998, p. 375). The case of several linearly independent cointegrating vectors or the situation where Δx_t adjusts to lagged deviations, too, is beyond the scope of a single-equation framework, see e.g. Lütkepohl (2006) in this volume.

Assumption 2 (ii) was made to ensure exogeneity of Δx_t. It may seem very restrictive for applied work. Working with normally distributed data, however, we do not need it because Johansen (1992) proved assuming a Gaussian vector EC model for $(y_t, x_t')'$ that Assumption 1 (iii) alone is sufficient for weak exogeneity of Δx_t, cf. also Urbain (1992, Prop. 1). In fact, he thus showed that under Assumption 1 (iii) alone the single-equation analysis is equivalent to maximum likelihood estimation of the full system (Johansen, 1992, Corollary 1).

5.3 Inference on the Cointegrating Vector

In this section we assume that y_t and x_t are cointegrated, and the interest focusses on estimating and testing β given T observations. It is well known since Phillips and Durlauf (1986) or Stock (1987) that the static OLS estimator,

$$\hat{y}_t = \hat{\alpha} + \hat{\beta}' x_t, \quad t = 1, \ldots, T,$$

is super-consistent. Under exogeneity, it further holds (cf. Phillips and Park, 1988) that $T(\hat{\beta} - \beta)$ converges to a normal distribution, where the variance depends on the long-run variance (or spectral density at frequency zero) of $y_t - \beta' x_t$. This parameter may be difficult to estimate in finite samples. Moreover, already Banerjee *et al.* (1986) observed that static OLS may be biased in finite samples due to ignoring short-run dynamics. An alternative approach dating back to Stock (1987) relies on estimating (5.6):

$$\Delta y_t = \hat{c} + \hat{\gamma}\, y_{t-1} + \hat{\theta}' x_{t-1} + \sum_{i=1}^{p-1} \hat{\alpha}_i \Delta y_{t-i} + \sum_{i=0}^{n-1} \hat{\phi}_i' \Delta x_{t-i} + \hat{\varepsilon}_t\,. \quad (5.8)$$

A natural candidate for estimating β is now from (5.8) because of (14.2.2)

$$\hat{\beta}_{EC} = -\frac{\hat{\theta}}{\hat{\gamma}}\,. \quad (5.9)$$

Below we will obtain limiting normality of $T(\hat{\beta}_{EC} - \beta)$ under exogeneity by drawing upon results by Pesaran and Shin (1998), who consider the OLS estimation of (5.5):

$$y_t = \tilde{c} + \sum_{i=1}^{p} \tilde{a}_i\, y_{t-i} + \tilde{\theta}' x_t + \sum_{i=0}^{n-1} \tilde{\phi}_i' \Delta x_{t-i} + \tilde{\varepsilon}_t\,. \quad (5.10)$$

As estimator for β they propose because of (14.2.2):

$$\hat{\beta}_{PS} = \frac{\tilde{\theta}}{1 - \sum_{i=1}^{p} \tilde{a}_i}\,. \quad (5.11)$$

Pesaran and Shin (1998, Theorem 2.4 or 3.2) establish limiting normality under the stated assumptions.

Proposition: Under Assumptions 1 and 2 and under cointegration it holds as $T \to \infty$:

$$\left[\sum_{t=1}^{T}(x_t - \bar{x})(x_t - \bar{x})'\right]^{0.5}\left(\hat{\beta}_{PS} - \beta\right) \sim \mathcal{N}_K\left(0, \frac{\sigma^2}{(a(1))^2} I_K\right),$$

where I_K denotes the identity matrix.

Remark: A It is noteworthy that $\sum(x_t - \bar{x})(x_t - \bar{x})'$ diverges with T^2, so that $\hat{\beta}_{PS}$ converges with the expected super-consistent rate T. Although normality arises just like in the stationary case, the rate of convergence differs from the situation where x_t is I(0). Moreover, σ^2 and $a(1)$ may be estimated consistently:

$$\tilde{\sigma}^2 = \frac{1}{T-m}\sum_{t=1}^{T}\tilde{\varepsilon}_t^2, \quad \tilde{a}(1) = 1 - \sum_{i=1}^{p}\tilde{a}_i,$$

where $m = K(n+1)+p+1$ denotes the number of estimated parameters including a constant. Finally, by demeaning x_t in Proposition 1, we assume that the regression equation contains an intercept. The result continues to hold, if a linear time trend as additional regressor is allowed for.

Remark: B In practice, Assumption 2 (ii) may be too restrictive, and (lagged values of) Δx_t may be correlated with ε_t. To account for that, Pesaran and Shin (1998) propose to simply include the corresponding difference Δx_{t-k} as additional regressor in (5.10) in case that $k \geq n$.

Since (5.6) is a linear transformation of (5.5), it turns out that the regression (5.8) is a linear transformation of (5.10). Using the techniques by Wickens and Breusch (1988) we can establish the following result. The proof is tedious but not difficult, details are available upon request.

Proposition: For the OLS regressions (5.10) and (5.8) it holds:

$$\hat{\gamma} = \sum_{i=1}^{p}\tilde{a}_i - 1, \quad \hat{\theta} = \tilde{\theta}, \quad \hat{\varepsilon}_t = \tilde{\varepsilon}_t,$$

and consequently: $\hat{\beta}_{EC} = \hat{\beta}_{PS}$.

As a corollary to Propositions 1 and 2, $\hat{\beta}_{EC}$ follows a limiting normal distribution. Consider a t type statistic testing for the kth component $\beta^{(k)}$, $k = 1, 2, \ldots, K$:

$$\tau_k = \frac{|\hat{\gamma}|\,(\hat{\beta}_{EC}^{(k)} - \beta^{(k)})}{\hat{\sigma}\sqrt{[(\sum(x_t - \bar{x})(x_t - \bar{x})')^{-1}]_{kk}}},$$

where $[\cdot]_{kk}$ denotes the entry on the principal diagonal of a matrix, and, obviously: $\hat{\sigma}^2 = \frac{1}{T-m}\sum_{t=1}^{T}\hat{\varepsilon}_t^2$ where again $m = K(n+1)+p+1$.

Corollary: Under the assumptions of Proposition 1 it holds for $k = 1, \ldots, K$:

$$\tau_k \sim \mathcal{N}(0,1),$$

as $T \to \infty$.

Concluding this section it should be noticed that estimation of and inference about β from linear or nonlinear dynamic regressions similar to (5.8) and (5.10) has been discussed by Stock (1987), Phillips (1988), Phillips and Loretan (1991), and Boswijk (1995), too.

5.4 Cointegration Testing

We consider tests for the null hypothesis of no cointegration building on the error-correction equation (5.6) augmented by a constant intercept and estimated by OLS, $t = 1, 2, \ldots, T$,

$$\Delta y_t = \widehat{c} + \widehat{\gamma}_\mu \, y_{t-1} + \widehat{\theta}'_\mu x_{t-1} + \sum_{i=1}^{p-1} \widehat{\alpha}_i \Delta y_{t-i} + \sum_{i=0}^{n-1} \widehat{\phi}'_i \Delta x_{t-i} + \widehat{\varepsilon}_t \,. \quad (5.12)$$

Sometimes empirical researchers wish to work with detrended series, which amounts to adding a linear time trend to the set of regressors[2]:

$$\Delta y_t = \widehat{c} + \widehat{\delta}\, t + \widehat{\gamma}_\tau \, y_{t-1} + \widehat{\theta}'_\tau x_{t-1} + \sum_{i=1}^{p-1} \widehat{\alpha}_i \Delta y_{t-i} + \sum_{i=0}^{n-1} \widehat{\phi}'_i \Delta x_{t-i} + \widehat{\varepsilon}_t \,. \quad (5.13)$$

Clearly, the linear trend will change all parameter estimates. For that reason γ and θ are now indexed with τ, while all other estimates are denoted by the same symbols as in (5.12) for convenience. Sometimes, (5.12) and (5.13) are called conditional (or structural) error-correction models, while unconditional (reduced form) models are obtained by restricting $\phi_0 = 0$ and excluding contemporaneous differences, Δx_t.

Given Assumption 1, the null hypothesis of no cointegration[3] may be parameterized as follows:

$$H_0 : \gamma_\mu = 0 \quad \text{or} \quad \gamma_\tau = 0 \,.$$

Under the alternative of cointegration equilibrium adjustment implies

$$H_1 : \gamma_\mu < 0 \quad \text{or} \quad \gamma_\tau < 0 \,.$$

Therefore, Banerjee *et al.* (1998) proposed the use of the conventional studentized t statistic relying on an OLS estimation of (5.12) or (5.13):

$$ECt_\mu = t_{\gamma_\mu = 0} \quad \text{or} \quad ECt_\tau = t_{\gamma_\tau = 0} \,.$$

[2]This is not so much a question of choosing the 'right model' but rather the economically more meaningful one, cf. Hassler (1999) for a discussion.

[3]Pesaran *et al.* (2001) consider a more general procedure where the order of integration of x_t is not assumed to be known. The null hypothesis then reads as 'there is no stable level relationship between y_t and x_t'. The limiting distribution depends on the I(0) or I(1) assumption, but Pesaran *et al.* (2001) propose to use critical values from both distributions as bounds: If the test statistic falls outside the bounds, a conclusive inference may be drawn without having to know the integration status of the underlying variables.

The null hypothesis is rejected for too small (negative) values.

Similarly, Boswijk (1994) suggested an F type test for

$$H_0 : \gamma_\mu = 0, \theta_\mu = 0 \quad \text{or} \quad \gamma_\tau = 0, \theta_\tau = 0 .$$

Let $F_{\gamma,\theta}$ denote the conventional F statistics from (5.12) or (5.13) testing for lack of significance. Then Boswijk (1994) considered

$$ECF_\mu = (K+1)F_{\gamma_\mu,\theta_\mu} \quad \text{or} \quad ECF_\tau = (K+1)F_{\gamma_\tau,\theta_\tau} .$$

Here, the null hypothesis is rejected for too large values. Boswijk (1994) suggested a further variant for (5.13), where the linear trend is restricted under H_0:

$$H_0 : \gamma_\tau = 0, \quad \theta_\tau = 0, \quad \delta = 0 .$$

The corresponding F type statistic tests for $K+2$ restrictions:

$$ECF_\tau^* = (K+2)F_{\gamma_\tau,\theta_\tau,\delta} .$$

In many economic applications it may occur that x_t is $I(1)$ with drift,

$$E(\Delta x_t) = d \neq 0 .$$

Still, empirical workers often wish to regress without detrending. However, the linear trend is growing with T while the stochastic trend is only of order $T^{0.5}$. Hence, the linear trend in the data,

$$\begin{aligned} x_t &= x_0 + d\,t + I(1) \\ &= O_p(1) + O_p(T) + O_p(T^{0.5}) , \end{aligned}$$

dominates the stochastic trend and hence affects the limiting distribution of ECt_μ from (5.12). Fortunately, critical values are nevertheless readily available.

Proposition Under Assumptions 1 and 2 and the null hypothesis of no cointegration, it holds as $T \to \infty$:

a) $ECt_\tau \xrightarrow{d} \mathcal{BDM}_\tau(K)$ for any $E(\Delta x_t)$;

b) $ECt_\mu \xrightarrow{d} \mathcal{BDM}_\mu(K)$ for $E(\Delta x_t) = 0$;

c) $ECt_\mu \xrightarrow{d} \mathcal{BDM}_\tau(K-1)$ for $E(\Delta x_t) \neq 0$, where $\mathcal{BDM}_\tau(0)$ stands for the detrended Dickey-Fuller distribution.

Convergence in distribution is denoted by $\xrightarrow{d}$. The following variables $\mathcal{BDM}_\mu(K)$ and $\mathcal{BDM}_\tau(K)$ represent functionals of vector standard Brownian motions of length K, which are demeaned and detrended, respectively. K denotes the number of variables contained in the vector x_t. Detailed expressions of those limiting distributions and simulated critical values can be found in Banerjee *et al.* (1998), who prove a) and b). The third result was established by Hassler (2000), and by detrended

Dickey-Fuller distribution we mean the limit of $\hat{\tau}_\tau$ in the notation by Dickey and Fuller (1979).

Remark: C In applied work it may be not so clear whether x_t is dominated by a linear trend or not. It hence may be dubious whether to use Proposition 3 b) or c) for inference. Looking at corresponding percentiles in Table 1, we learn that critical values from $\mathcal{BDM}_\tau(K-1)$ are smaller than those from $\mathcal{BDM}_\mu(K)$ for K being fixed. Usage of $\mathcal{BDM}_\tau(K-1)$ in case of regressions without detrending therefore results in a conservative test avoiding over-rejection, which may be the advisable strategy in practice.

Similar results to Proposition 3 are available for the F type statistics.

Proposition: Under Assumptions 1 and 2 and the null hypothesis of no cointegration, it holds as $T \to \infty$:

a) $ECF_\tau \xrightarrow{d} \mathcal{B}_\tau(K)$ for any $E(\Delta x_t)$;

b) $ECF_\tau^* \xrightarrow{d} \mathcal{B}_\tau^*(K)$ for any $E(\Delta x_t)$;

c) $ECF_\mu \xrightarrow{d} \mathcal{B}_\mu(K)$ for $E(\Delta x_t) = 0$.

Boswijk (1994) characterized the stochastic limits of the type $\mathcal{B}$ depending again on the number of I(1)-variables x_t and on the deterministics (with or without linear trend). However, there remains one question. How do linear trends in the data affect the limiting distribution of the F type test ECF_μ without detrending? Without proof we state motivated by Proposition 3 c) the following conjecture (a proof could follow the lines in Hassler, 2000).

Conjecture: When $E(\Delta x_t) \neq 0$, we conjecture under the assumptions of Proposition 4 for the regression without detrending:

$$ECF_\mu \xrightarrow{d} \mathcal{B}_\tau^*(K-1), \tag{5.14}$$

where in case of $K = 1$, $\mathcal{B}_\tau^*(0)$ is understood to be twice the limiting distribution of the Φ_3 statistic from Dickey and Fuller (1981); see Table VI in Dickey and Fuller (1981) for percentiles: $\Phi_3 \xrightarrow{d} \frac{1}{2}B_\tau^*(0)$.

The applicability of (5.14) in finite samples will be established by computer experiments in Section 5. The intuition behind this claim is again that the process x_t follows one common linear time trend and K stochastic I(1) trends. The linear trend dominates one stochastic trend. Therefore, in case of linear trends the following holds asymptotically: testing for $\theta_\mu = 0$ in (5.12) with θ_μ being of length K amounts to the same as if we tested for $\delta = 0$ and $\theta_\tau = 0$ where θ_τ was only $(K-1)$-dimensional in (5.13).

Examples of critical values of the distributions encountered in this section are given in Table 1. We observe that critical values of $\mathcal{B}_\tau^*(K-1)$ are shifted to the right relative to $\mathcal{B}_\mu(K)$. Analogously to Remark C the use of $\mathcal{B}_\tau^*(K-1)$ instead of $\mathcal{B}_\mu(K)$ in case of regressions without detrending hence results in a conservative procedure.

Table 5.1: **Critical Values**

	$\mathcal{BDM}_\mu(K)$	$\mathcal{BDM}_\tau(K-1)$	$\mathcal{B}_\mu(K)$	$\mathcal{B}^*_\tau(K-1)$
	$K=1$			
1 %	-3.78	-3.96	15.22	16.54
5 %	-3.19	-3.41	11.41	12.50
10 %	-2.89	-3.13	9.54	10.68
	$K=2$			
1 %	-4.06	-4.27	18.68	19.30
5 %	-3.48	-3.69	14.38	15.24
10 %	-3.19	-3.39	12.22	13.22
	$K=3$			
1 %	-4.46	-4.51	21.43	22.50
5 %	-3.74	-3.91	17.18	18.03
10 %	-3.42	-3.62	14.93	15.85
	$K=4$			
1 %	-4.57	-4.72	24.63	25.46
5 %	-3.97	-4.12	19.69	20.66
10 %	-3.66	-3.82	17.38	18.45
	$K=5$			
1 %	-4.70	-4.89	27.11	28.51
5 %	-4.27	-4.30	22.48	23.33
10 %	-3.82	-4.00	19.87	20.76

Note: The asymptotic critical values of $\mathcal{BDM}_\mu(K)$ and $\mathcal{BDM}_\tau(K-1)$ are taken from Banerjee *et al.* (1998, Table I), except for $\mathcal{BDM}_\tau(0)$ from Fuller (1996, Table 10.A.2). The percentiles of $\mathcal{B}_\mu(K)$ and $\mathcal{B}^*_\tau(K-1)$ are from Tables B.2 and B.5 in Boswijk (1994), except for $\mathcal{B}^*_\tau(0)$. The latter quantiles are twice the values found in Dickey and Fuller (1981, Table VI).

5.5 Monte Carlo Evidence

For simulation purposes we generated a bivariate process ($K = 1$) as

$$\begin{pmatrix} \Delta y_t \\ \Delta x_t \end{pmatrix} = \begin{pmatrix} -\gamma_1 \\ \gamma_2 \end{pmatrix} \left(y_{t-1} - x_{t-1} \right) + \begin{pmatrix} \frac{1}{2} & \frac{1}{4} \\ 0 & \frac{1}{2} \end{pmatrix} \begin{pmatrix} \Delta y_{t-1} \\ \Delta x_{t-1} \end{pmatrix} + \varepsilon_t \,, \tag{5.15}$$

$$\varepsilon_t \sim ii\mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix} \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right), \quad t = 1, 2, \ldots, T \,. \tag{5.16}$$

We consider the conditional error-correction regression,

$$\Delta y_t = \hat{c} + \hat{\gamma}\, y_{t-1} + \hat{\theta}\, x_{t-1} + \hat{\alpha}_1\, \Delta y_{t-1} + \hat{\phi}_0 \Delta x_t + \hat{\phi}_1\, \Delta x_{t-1} + \hat{\varepsilon}_t \,, \tag{5.17}$$

as well as the unconditional one without contemporaneous Δx_t:

$$\Delta y_t = \tilde{c} + \tilde{\gamma}\, y_{t-1} + \tilde{\theta}\, x_{t-1} + \tilde{\alpha}_1\, \Delta y_{t-1} + \tilde{\phi}_1\, \Delta x_{t-1} + \tilde{\varepsilon}_t \,. \tag{5.18}$$

Clearly, the unconditional regression (5.18) is only appropriate if $\rho = 0$; if $\rho \neq 0$, however, the inclusion of Δx_t is required to account for simultaneous correlation.

Throughout we present rejections at the nominal 5 % level that are obtained from 50 000 replications. All programming[4] was done in Ox Professional 3.30.

Table 2 contains results for the asymptotically normal cointegration estimator $\hat{\beta}_{EC}$, see Corollary 1. For the upper and the middle panel we assume $\gamma_2 = 0$ and $\gamma_1 \in \{0.2, 0.4, 0.6\}$. With growing γ_1 (i.e. error-correction adjustment) the experimental size improves. For $T = 100$ the test is oversized. With $T = 250$, the experimental level of the conditional regression is fairly close to the nominal one, and the correspondence is very good for $T = 1000$. Moreover, for $\rho > 0$, Assumption 2 (ii) is violated because Δx_t and the regression error are correlated. This turns the unconditional regression (5.18) invalid, while the conditional regression is not affected by ρ. This supports the proposal by Pesaran and Shin (1998) to add (lags of) Δx_t in case that Assumption 2 (ii) does not hold in order to maintain limiting normality, cf. Remark B. In the lower panel Assumption 1 (iii) is violated because $\gamma_2 = \gamma_1 \neq 0$. In this situation Δx_t is not exogeneous as proven by Johansen (1992). Therefore, even the conditional regression does not result in a limiting $\mathcal{N}(0, 1)$ distribution as is well demonstrated for $T = 1000$.

Table 3 displays findings for the cointegration tests ECt and ECF with $T = 100$ only. In the column '$\gamma = 0$' it holds $\gamma_1 = \gamma_2 = 0$, and the null hypothesis is true. The next three columns assume $\gamma_2 = 0$ and $\gamma_1 \in \{0.05, 0.1, 0.2\}$. The power increases with γ_1, and the t and F tests behave very similarly. The conditional regression including Δx_t produces tests that are robust with respect to ρ, while (5.18) results in dramatic power losses as ρ grows. In the next three columns ($\gamma_1 = 0$, $\gamma_2 \in \{0.05, 0.1, 0.2\}$) we do have cointegration but y_t does not adjust. Hence, the unconditional regression provides no power, while (5.17) still allows to reject, as long as $\rho \neq 0$. Only here it turns out that the F type test is slightly more powerful. In

[4] We thank Vladimir Kuzin for computational help.

Table 5.2: **Asymptotically Normal Cointegration Vector**

	$T = 100$			$T = 250$			$T = 1000$		
$\gamma_1 =$	0.2	0.4	0.6	0.2	0.4	0.6	0.2	0.4	0.6
$\gamma_2 = 0$	Conditional regression (5.17)								
$\rho = 0$	11.4	9.0	7.8	7.2	6.4	6.1	5.5	5.3	5.2
$\rho = 0.3$	11.1	9.1	8.0	7.3	6.5	6.2	5.6	5.3	5.3
$\rho = 0.6$	11.1	8.8	8.3	7.3	6.4	6.1	5.7	5.3	5.2
$\gamma_2 = 0$	Unconditional regression (5.18)								
$\rho = 0$	10.9	8.5	7.4	7.0	6.2	5.9	5.5	5.3	5.1
$\rho = 0.3$	12.2	10.2	9.4	9.2	8.3	8.1	7.5	7.3	7.4
$\rho = 0.6$	17.8	16.0	15.4	15.1	14.2	14.1	13.8	13.9	13.7
$\gamma_2 = \gamma_1$	Conditional regression (5.17)								
$\rho = 0$	22.6	20.9	21.0	19.0	18.3	18.4	17.9	17.6	17.5
$\rho = 0.3$	18.1	16.9	16.3	15.1	14.6	14.6	13.8	13.8	13.9
$\rho = 0.6$	13.7	12.7	12.5	11.4	10.7	10.5	10.2	10.0	10.1

Note: The true DGP is (7.1) with (5.16). We report the frequency of rejection of a two-sided test as in Corollary 1 at the 5 % significance level.

the last three columns Assumption 1 (iii) does not hold ($\gamma_2 = \gamma_1 \in \{0.05, 0.1, 0.2\}$). If $\rho \neq 0$, the power increases with growing ρ for the tests based on conditional regressions compared with $\gamma_2 = 0$, while in case of unconditional regressions the power is reduced.

Finally, Table 4 supports our conjecture. Here, we simulated $(K+1)$-dimensional random walks $(y_t, x_t')'$ independent of each other. Moreover, x_t contains a drift, which is identical in all components:

$$x_t = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} t + \sum_{i=1}^{t} \varepsilon_i.$$

Application of ECF_μ from (5.17) with critical values from $\mathcal{B}_\tau^*(K-1)$ provides a valid approximation as T increases.

5.6 Summary

We reviewed different parameterizations of the autoregressive distributed lag (ADL) model and stressed the equivalence with error-correction (EC) mechanisms. This motivates the following finding: the cointegrating vector and the residuals computed from the EC model are numerically identical to the ones constructed from the ADL

Table 5.3: **Cointegration Tests**

	$\gamma = 0$	γ_1 $(\gamma_2 = 0)$			γ_2 $(\gamma_1 = 0)$			$\gamma_1 = \gamma_2$		
	H_0	0.05	0.1	0.2	0.05	0.1	0.2	0.05	0.1	0.2
Conditional regression (5.17)										
					$\rho = 0$					
ECt_μ	6.1	30.0	75.4	99.6	2.7	1.5	0.7	22.1	48.1	79.8
ECF_μ	6.4	29.2	71.6	99.2	5.5	4.3	4.1	20.4	45.4	80.0
					$\rho = 0.3$					
ECt_μ	5.7	26.2	68.0	98.9	6.6	7.7	8.5	32.0	67.7	94.0
ECF_μ	6.4	23.9	64.0	98.4	8.6	10.0	13.1	30.6	66.6	94.1
					$\rho = 0.6$					
ECt_μ	5.8	25.6	67.3	98.8	14.6	26.7	45.1	48.2	87.3	99.2
ECF_μ	6.7	23.3	63.0	98.2	17.7	31.3	51.1	47.4	86.9	99.2
Unconditional regression (5.18)										
					$\rho = 0$					
ECt_μ	6.0	30.4	76.1	99.6	2.7	1.7	0.7	24.1	55.1	89.0
ECF_μ	6.4	29.1	72.3	99.3	5.7	4.3	4.2	22.0	51.8	88.4
					$\rho = 0.3$					
ECt_μ	5.5	22.9	60.8	96.3	2.3	1.4	0.7	16.1	35.7	70.8
ECF_μ	6.8	21.8	58.1	95.9	5.6	4.5	3.9	17.0	37.3	72.6
					$\rho = 0.6$					
ECt_μ	4.4	13.6	34.1	76.0	1.8	1.1	0.5	8.0	17.0	39.7
ECF_μ	6.4	15.3	38.7	80.9	5.5	4.5	4.1	12.4	24.4	48.4

Note: The true DGP is (7.1) with (5.16) and $T = 100$. We report the frequency of rejection at the 5 % significance level.

regression. Therefore, under the exogeneity conditions of Pesaran and Shin (1998) the limiting normality of the estimated cointegrating vector carries over to the EC model. Next, we review t-type and F-type tests for the null hypothesis of no cointegration proposed in an EC framework by Banerjee *et al.* (1998) and Boswijk (1994), respectively. Hassler (2000) treated the t-type test in the presence of linear trends in the data when regressions are run without detrending. Here, we treat the F-type test in the same situation. We refrain from proving the limiting distribution but support a conjecture by means of simulation evidence instead.

The main results of our Monte Carlo study are the following. First, in most cases the the t-type cointegration test is just as powerful as the F-type one. Second, we investigate the case that is of particular interest in applied work where Δx_t is correlated with the regression error. In this situation, the conditional regression (including contemporaneous Δx_t as regressor) still provides valid inference about

Table 5.4: **Conjecture**

	$T = 100$			$T = 250$			$T = 1000$		
	$K = 1$	$K = 2$	$K = 3$	$K = 1$	$K = 2$	$K = 3$	$K = 1$	$K = 2$	$K = 3$
1 %	1.35	1.84	1.89	1.06	1.49	1.46	0.98	1.34	1.26
5 %	5.40	6.18	6.60	5.17	5.80	5.77	4.89	5.50	5.47
10 %	10.37	11.45	11.78	9.96	11.04	11.18	9.72	10.70	10.53

Note: The true DGP is a random walk with drift. We report rejection frequencies of the F test applied to (5.17) with critical values from $\mathcal{B}_T^*(K-1)$.

the cointegration vector relying on the normal approximation. For this result to hold true it is crucial that Δx_t is exogenous in the sense that it does not adjust to past equilibrium deviations. Moreover, cointegration tests from the conditional regression are more powerful than those from unconditional ones. A general finding hence is that the conditional error-correction regression outperforms the unconditional one.

References

ALOGOSKOUFIS, G., SMITH, R. (1995). On error correction models: Specification, interpretation, estimation. In *Surveys in Econometrics* (L. Oxley, D. A. R. George, C. J. Roberts, S. Sayer, eds.), 139–170. Blackwell Publishers, Oxford.

BANERJEE, A., DOLADO, J. J., MESTRE, R. (1998). Error-correction mechanism tests for cointegration in a single-equation framework. *Journal of Time Series Analysis* **19** 267–283.

BANERJEE, A., DOLADO, J. J., HENDRY, D. F., SMITH, G. W. (1986). Exploring equilibrium relationships in econometrics through static models: Some monte carlo evidence. *Oxford Bulletin of Economics and Statistics* **48** 253–277.

BEWLEY, R. A. (1979). The direct estimation of the equilibrium response in a linear model. *Economics Letters* **3** 357–361.

BOSWIJK, H. P. (1994). Testing for an unstable root in conditional and structural error correction models. *Journal of Econometrics* **63** 37–60.

BOSWIJK, H. P. (1995). Efficient inference on cointegration parameters in structural error correction models. *Journal of Econometrics* **69** 133–158.

DAVIDSON, J. E.,H., HENDRY, D. F., SRBA, F., YEO, S. (1978). Econometric modelling of the aggregate time-series relationship between consumers' expenditure and income in the United Kingdom. *Economic Journal* **88** 661–692.

DICKEY, D. A., FULLER, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association* **74** 427–431.

DICKEY, D. A., FULLER, W. A. (1981). Likelihood ratio statistics for autoregressive time series with a unit root. *Econometrica* **49** 1057–1072.

ENGLE, R. F., GRANGER, C. W. J. (1987). Co-integration and error correction: Representation, estimation, and testing. *Econometrica* **55** 251–276.

ERICSSON, N. R. (2004). The ET interview: Professor David F. Hendry. *Econometric Theory* **20** 743–804.

FULLER, W. A. (1996). *Introduction to Statistical Time Series*. 2nd ed., Wiley, New York.

HASSLER, U. (1999). (When) Should cointegrating regressions be detrended? The case of a German money demand function. *Empirical Economics* **24** 155-172.

HASSLER, U. (2000). Cointegration testing in single error-correction equations in the presence of linear time trends. *Oxford Bulletin of Economics and Statistics* **62** 621–632.

HENDRY, D. F. (1979). Predictive failure and econometric modelling in macroeconomics: The transactions demand for money. In *Economic Modelling* (P. Ormerod, ed.), 217–242. Heinemann, London.

HENDRY, D. F., PAGAN, A., SARGAN, J. D. (1984). Dynamic specification. In *Handbook of Econometrics, II* (Z. Griliches, M.D. Intriligator, eds.), 1023–1100. North-Holland, Amsterdam.

HYLLEBERG, S., MIZON, G. E. (1989). Cointegration and error correction mechanisms. *The Economic Journal* **99** 113–125.

JOHANSEN, S. (1992). Cointegration in partial systems and the efficiency of single-equation analysis. *Journal of Econometrics* **52** 389–402.

LÜTKEPOHL, H. (2006). Structural vector autoregressive analysis for cointegrated variables. *Allgemeines Statistisches Archiv* **90** 75–88.

PESARAN, M. H., SHIN, Y. (1998). An autoregressive distributed-lag modelling approach to cointegration analysis. In *Econometrics and Economic Theory in the 20th Century. The Ragnar Frisch Centennial Symposium* (S. Strøm, ed.), 371–413. Cambridge University Press, Cambridge.

PESARAN, M. H., SHIN, Y., SMITH, R. J. (2001). Bounds testing approaches to the analysis of level relationships. *Journal of Applied Econometrics* **16** 289–326.

PHILLIPS, A. W. (1954). Stabilization policy in a closed economy. *Economic Journal* **64** 290–323.

PHILLIPS, A. W. (1957). Stabilization policy and the time form of lagged responses. *Economic Journal* **67** 265–277.

PHILLIPS, P. C. B. (1988). Reflections on econometric methodology. *The Economic Record* **64** 344–359.

PHILLIPS, P. C. B., DURLAUF, S. N. (1986). Multiple time series regressions with integrated processes. *Review of Economic Studies* **LIII** 473–495.

PHILLIPS, P. C. B., LORETAN, M. (1991). Estimating long-run economic equilibria. *Review of Economic Studies* **58** 407–436.

PHILLIPS, P. C. B., PARK, J. Y. (1988). Asymptotic equivalence of ordinary least squares and generalized least squares in regressions with integrated regressors. *Journal of the American Statistical Association* **83** 111-115.

SARGAN, J. D. (1964). Wages and prices in the United Kingdom: A study in econometric methodology. In *Econometric Analysis for National Economic Planning* (R. E. Hart, G. Mills, J. K. Whittaker, eds.), 25–54. Butterworths, London.

STOCK, J. H. (1987). Asymptotic properties of least-squares estimators of co-integrating vectors. *Econometrica* **55** 1035–1056.

WICKENS, M. R., BREUSCH, T. S. (1988). Dynamic specification, the long-run and the estimation of transformed regression models. *The Economic Journal* **98** 189–205.

URBAIN, J.-P. (1992). On weak exogeneity in error correction models. *Oxford Bulletin of Economics and Statistics* **54** 187–207.

6 Structural Vector Autoregressive Analysis for Cointegrated Variables*

Helmut Lütkepohl[1]

[1] Department of Economics, European University Institute
helmut.luetkepohl@iue.it

Summary: Vector autoregressive (VAR) models are capable of capturing the dynamic structure of many time series variables. Impulse response functions are typically used to investigate the relationships between the variables included in such models. In this context the relevant impulses or innovations or shocks to be traced out in an impulse response analysis have to be specified by imposing appropriate identifying restrictions. Taking into account the cointegration structure of the variables offers interesting possibilities for imposing identifying restrictions. Therefore VAR models which explicitly take into account the cointegration structure of the variables, so-called vector error correction models, are considered. Specification, estimation and validation of reduced form vector error correction models is briefly outlined and imposing structural short- and long-run restrictions within these models is discussed.

6.1 Introduction

In an influential article, Sims (1980) advocated the use of vector autoregressive (VAR) models for macroeconometric analysis as an alternative to the large simultaneous equations models that were in common use at the time. The latter models often did not account for the rich dynamic structure in time series data of quarterly or monthly frequency. Given that such data became more common in macro economic studies in the 1960s and 1970s, it was plausible to emphasize modelling of the dynamic interactions of the variables of interest. Sims also criticized the way the classical simultaneous equations models were identified and questioned the exogeneity assumptions for some of the variables which often reflect the prefer-

*I thank an anonymous reader for comments on an earlier draft of this paper that helped me to improve the exposition.

ences and prejudices of the model builders and are not necessarily fully backed by theoretical considerations. In contrast, in VAR models all observed variables are typically treated as a priori endogenous. Restrictions are imposed to a large extent by statistical tools rather than by prior beliefs based on controversial theories.

In a VAR analysis, the dynamic interactions between the variables are usually investigated by impulse responses or forecast error variance decompositions. These quantities are not unique, however. To identify those shocks or innovations and the associated impulse responses that reflect the actual ongoings in a given system of variables, usually also requires a priori assumptions which cannot be checked by statistical tools. Therefore *structural* VAR (SVAR) models were developed as a framework for incorporating identifying restrictions for the innovations to be traced out in an impulse response analysis.

In a parallel development it was discovered that the trending properties of the variables under consideration are of major importance for both econometric modelling and the associated statistical analysis. The spurious regression problem pointed out by Granger and Newbold (1974) showed that ignoring stochastic trends can lead to seriously misleading conclusions when modelling relations between time series variables. Consequently, the stochastic trends, unit roots or order of integration of the variables of interest became of major concern to time series econometricians and the concept of cointegration was developed by Granger (1981), Engle and Granger (1987), Johansen (1995) and many others. In this framework, the long-run relations are now often separated from the short-run dynamics. The cointegration or long-run relations are of particular interest because they can sometimes be associated with relations derived from economic theory. It is therefore useful to construct models which explicitly separate the long-run and short-run parts of a stochastic process. Vector error correction or equilibrium correction models (VECMs) offer a convenient framework for this purpose. They also open up the possibility to separate shocks or innovations with permanent and transitory effects. This distinction may be helpful in identifying impulse responses of interest. Therefore these models will be used as the framework in the following exposition.

A variable will be called *integrated of order d* ($I(d)$) if stochastic trends or unit roots can be removed by differencing the variable d times and a stochastic trend still remains after differencing only $d-1$ times. In line with this terminology, a variable without a stochastic trend or unit root is sometimes called $I(0)$. In the following, all variables are assumed to be either $I(0)$ or $I(1)$ to simplify matters. Hence, for a time series variable y_{kt}, it is assumed that the first differences, $\Delta y_{kt} \equiv y_{kt} - y_{k,t-1}$, have no stochastic trend. A set of $I(1)$ variables is called *cointegrated* if a linear combination exists which is $I(0)$. If a system consists of both $I(0)$ and $I(1)$ variables, any linear combination which is $I(0)$ is called a cointegration relation. Admittedly, this terminology is not in the spirit of the original idea of cointegration because it can happen that a linear combination of $I(0)$ variables is called a cointegration relation. In the present context, this terminology is a convenient simplification, however. Therefore it is used here.

Although in practice the variables will usually have nonzero means, polynomial trends or other deterministic components, it will be assumed in the following that deterministic terms are absent. The reason is that deterministic terms do not play

a role in impulse response analysis which is the focus of this study. Moreover, augmenting the models with deterministic terms is usually straightforward.

In the next section the model setup for structural modelling with cointegrated VAR processes will be presented. Estimation of the models is discussed in Section 3 and issues related to model specification are considered in Section 4. Conclusions follow in Section 5. The structural VECM framework of the present article was proposed by King *et al.* (1991) and a recent more general survey of structural VAR and VECM analysis with some examples was given by Breitung *et al.* (2004). Further references will be given in the following. The present article draws heavily on Lütkepohl (2005, Chapter 9), where further details can be found.

The following general notation will be used. The natural logarithm is abbreviated as log. For a suitable matrix A, $\mathrm{rk}(A)$ and $\det(A)$ denote the rank and determinant of A, respectively. Furthermore, for $n > m$, the $(n \times (n-m))$ matrix $A_\perp$ denotes an orthogonal complement of the $(n \times m)$ matrix A of rank m, that is, $A_\perp$ is such that $\mathrm{rk}(A_\perp) = n - m$ and $A_\perp' A = 0$. The orthogonal complement of a nonsingular square matrix is zero and the orthogonal complement of a zero matrix is an identity matrix of suitable dimension. Moreover, vec is the column stacking operator which stacks the columns of a matrix in a column vector and vech is the column stacking operator for symmetric square matrices which stacks the columns from the main diagonal downwards only. The $(n \times n)$ identity matrix is signified as I_n and $\mathbf{D}_n$ denotes the $(n^2 \times \frac{1}{2}n(n+1))$ duplication matrix defined such that for a symmetric $(n \times n)$ matrix A, $\mathrm{vec}(A) = \mathbf{D}_n \mathrm{vech}(A)$. The Kronecker product is denoted by $\otimes$.

6.2 The Model Setup

As mentioned earlier, it is assumed that all variables are at most $I(1)$ and that the data generation process can be represented as a VECM of the form

$$\Delta y_t = \alpha\beta' y_{t-1} + \Gamma_1 \Delta y_{t-1} + \cdots + \Gamma_{p-1}\Delta y_{t-p+1} + u_t\,, \quad t = 1, 2, \ldots, \tag{2.1}$$

where y_t is a K-dimensional vector of observable variables and α and β are $(K \times r)$ matrices of rank r. More precisely, β is the cointegration matrix and r is the cointegrating rank of the process. The term $\alpha\beta' y_{t-1}$ is sometimes referred to as error correction term. The Γ_j's, $j = 1, \ldots, p-1$, are $(K \times K)$ short-run coefficient matrices and u_t is a white noise error vector with mean zero and nonsingular covariance matrix Σ_u, $u_t \sim (0, \Sigma_u)$. Moreover, $y_{-p+1}, \ldots, y_0$ are assumed to be fixed initial conditions. Rewriting (2.1) in the levels of the variables gives a VAR(p) model of the form

$$y_t = A_1 y_{t-1} + \cdots + A_p y_{t-p} + u_t\,,$$

where $A_1 = \alpha\beta' + I_K + \Gamma_1$, $A_i = \Gamma_i - \Gamma_{i-1}$, $i = 2, \ldots, p-1$, and $A_p = -\Gamma_{p-1}$. Thus, the levels VAR form includes p lags when $p-1$ lagged differences are used in the VECM (2.1).

6.2.1 The Identification Problem

Impulse responses are often used to study the relationships between the variables of a dynamic model such as (2.1). In other words, the marginal effect of an impulse

to the system is traced out over time. The residuals u_t are the 1-step ahead forecast errors associated with the VECM (2.1). Tracing the marginal effects of a change in one component of u_t through the system may not reflect the actual responses of the variables because in practice an isolated change in a single component of u_t is not likely to occur if the component is correlated with the other components. Hence, in order to identify structural innovations which induce informative responses of the variables, uncorrelated or orthogonal impulses or shocks or innovations are usually considered.

In a VAR analysis the so-called AB-model of Amisano and Giannini (1997) provides a general framework for imposing structural restrictions. If cointegrated variables and VECMs are considered, the special case of a B-model setup is typically used. We will therefore focus on the B-model in the following. In that setup it is assumed that the structural innovations, say ε_t, have zero mean and identity covariance matrix, $\varepsilon_t \sim (0, I_K)$, and they are linearly related to the u_t such that

$$u_t = B\varepsilon_t\,.$$

Hence, $\Sigma_u = BB'$. Without further restrictions, the $(K \times K)$ matrix B is not uniquely specified by these relations. In fact, due to the symmetry of the covariance matrix, $\Sigma_u = BB'$ represents only $\frac{1}{2}K(K+1)$ independent equations. For a unique specification of the K^2 elements of B we need at least $\frac{1}{2}K(K-1)$ further restrictions. Some of them may be obtained via a more detailed examination of the cointegration structure of the model, as will be seen in the following.

According to Granger's representation theorem (see Johansen, 1995), the process y_t has the representation

$$y_t = \Xi \sum_{i=1}^{t} u_i + \sum_{j=0}^{\infty} \Xi_j^* u_{t-j} + y_0^*, \quad t = 1, 2, \ldots, \tag{2.2}$$

where the term y_0^* contains the initial values and the Ξ_j^*'s are absolutely summable so that the infinite sum is well-defined. Absolute summability of the Ξ_j^*'s implies that these matrices converge to zero for $j \to \infty$. Notice that the term $x_t \equiv \sum_{i=1}^{t} u_i = x_{t-1} + u_t$, $t = 1, 2, \ldots$, is a K-dimensional random walk. The long-run effects of shocks are represented by the term $\Xi \sum_{i=1}^{t} u_i$ which captures the common stochastic trends. The matrix Ξ can be shown to be of the form

$$\Xi = \beta_\perp \left[\alpha_\perp' \left(I_K - \sum_{i=1}^{p-1} \Gamma_i \right) \beta_\perp \right]^{-1} \alpha_\perp'.$$

It has rank $K - r$. Thus, there are $K - r$ independent common trends. Substituting $B\varepsilon_i$ for u_i in the common trends term in (2.2) gives $\Xi \sum_{i=1}^{t} u_i = \Xi B \sum_{i=1}^{t} \varepsilon_i$. Clearly, the long-run effects of the structural innovations are given by ΞB because the effects of an ε_t impulse vanish in $\sum_{j=0}^{\infty} \Xi_j^* B\varepsilon_{t-j}$ in the long-run.

The structural innovations ε_t represent a regular random vector with nonsingular covariance matrix. Hence, the matrix B has to be nonsingular. Thus, $\mathrm{rk}(\Xi B) = K - r$ and there can at most be r zero columns in the matrix ΞB. In other words, at most r of the structural innovations can have transitory effects and at least $K - r$ of them must have permanent effects. If a cointegrating rank r is diagnosed and r transitory

shocks can be justified, r columns of ΞB can be restricted to zero. Because the matrix has reduced rank $K-r$, each column of zeros stands for $K-r$ independent restrictions only. Thus, the r transitory shocks represent $r(K-r)$ independent restrictions only. Still, it is useful to note that restrictions can be imposed on the basis of knowledge of the cointegrating rank of the system which can be determined by statistical means, provided as many transitory shocks can be justified as there are linearly independent cointegration relations. For a unique specification of B, further theoretical considerations are required for imposing additional restrictions, however.

For just-identification of the structural innovations in the B-model we need a total of $K(K-1)/2$ independent restrictions. Given that $r(K-r)$ restrictions can be derived from the cointegration structure of the model, this leaves us with $\frac{1}{2}K(K-1)-r(K-r)$ further restrictions for just-identifying the structural innovations. More precisely, $r(r-1)/2$ additional restrictions are required for the transitory shocks and $(K-r)((K-r)-1)/2$ restrictions are needed to identify the permanent shocks (see, e.g., King *et al.*, 1991; Gonzalo and Ng, 2001). Together this gives a total of $\frac{1}{2}r(r-1)+\frac{1}{2}(K-r)((K-r)-1)=\frac{1}{2}K(K-1)-r(K-r)$ restrictions. The transitory shocks may be identified, for example, by placing zero restrictions on B directly and thereby specifying that certain shocks have no instantaneous impact on some of the variables. Clearly, it is not sufficient to impose arbitrary restrictions on B or ΞB. They have to be such that they identify the transitory and permanent shocks. For instance, the transitory shocks cannot be identified through restrictions on ΞB because they correspond to zero columns in that matrix. In other words, $r(r-1)/2$ of the restrictions have to be imposed on B directly. Generally, identifying restrictions are often of the form

$$C_{\Xi B}\mathrm{vec}(\Xi B)=c_l \quad \text{and} \quad C_s\mathrm{vec}(B)=c_s, \tag{2.3}$$

where $C_{\Xi B}$ and C_s are appropriate selection matrices to specify the long-run and contemporaneous restrictions, respectively, and c_l and c_s are vectors of suitable dimensions. In applied work, they are typically zero vectors. In other words, zero restrictions are specified in (2.3) for ΞB and B. The first set of restrictions can be written alternatively as

$$C_l\mathrm{vec}(B)=c_l, \tag{2.4}$$

where $C_l \equiv C_{\Xi B}(I_K \otimes \Xi)$ is a matrix of long-run restrictions on B.

So far we have just discussed a 'counting rule' and, hence, a necessary condition for identification. Even though the restrictions in (2.4) are linear restrictions, the full set of equations we have for B is a nonlinear one because the relation $\Sigma_u = BB'$ is nonlinear. Hence, the matrix B will only be identified locally in general. In particular, we may reverse the signs of the columns of B to find another valid matrix. If restrictions of the form

$$C_l\mathrm{vec}(B)=c_l \quad \text{and} \quad C_s\mathrm{vec}(B)=c_s \tag{2.5}$$

are available for B, a necessary and sufficient condition for local identification is that

$$\mathrm{rk}\begin{bmatrix} 2\mathbf{D}_K^{+}(B\otimes I_K) \\ C_l \\ C_s \end{bmatrix} = K^2,$$

where $\mathbf{D}_K^+$ is the Moore-Penrose inverse of the $(K^2 \times \frac{1}{2}K(K+1))$ duplication matrix $\mathbf{D}_K$ (see Lütkepohl, 2005, Proposition 9.4). Although the unknown parameter matrix B appears in this condition, it is useful in practice because it will fail everywhere in the parameter space or be satisfied everywhere except on a set of Lebesgue measure zero. Thus, if a single admissible B matrix can be found which satisfies the restrictions in (2.5) and for which also the rank condition holds, then local identification is ensured almost everywhere in the parameter space. Thus, trying an arbitrary admissible B matrix is a possibility for checking identification.

As an example, consider a model for $K = 3$ variables. Assuming that all variables are $I(1)$ and the cointegrating rank $r = 2$, then there can be two transitory shocks and one permanent shock. If two transitory shocks are assumed, the permanent shock is identified in this situation without further assumptions because $K - r = 1$ and, hence, the number of additional restrictions is $(K-r)((K-r)-1)/2 = 0$. Moreover, only $1 \; (= r(r-1)/2)$ further restriction is necessary to identify the two transitory shocks. Assuming a recursive structure for the two transitory shocks and placing the permanent shock first in the ε_t vector, the following restrictions are obtained:

$$\Xi B = \begin{bmatrix} * & 0 & 0 \\ * & 0 & 0 \\ * & 0 & 0 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} * & * & * \\ * & * & 0 \\ * & * & * \end{bmatrix}.$$

In these matrices the asterisks denote unrestricted elements. The two zero columns in ΞB represent two independent restrictions only because ΞB has rank $K-r = 1$. A third restriction is placed on B. The way it is specified, the third shock does not have an instantaneous effect on the second variable. Hence, there are $K(K-1)/2 = 3$ independent restrictions in total and the structural innovations are locally just-identified. In this case, uniqueness can be obtained, for instance, by fixing the signs of the diagonal elements of B.

In this example, with two zero columns in ΞB, it is also easy to see that it does not suffice to impose a further restriction on this matrix to identify B locally. To disentangle the two transitory shocks, we have to impose a restriction on B. In fact, it is necessary to restrict an element in the last two columns of B.

In the standard B-model with three variables which does not take into account the cointegration structure, at least $\frac{1}{2}K(K-1) = 3$ restrictions are needed for identification. In contrast, in the present VECM case, assuming that $r = 2$ and that there are two transitory shocks, only one restriction is required because two columns of ΞB are zero. Thus, the long-run restrictions from the cointegration structure of the variables may help in the identification of shocks of interest. As another example consider a bivariate system with one cointegrating relation. No further restriction is required to identify the permanent and transitory shocks in this case, if, say, the first shock is allowed to have permanent effects and the second one can have transitory effects only. Further examples may be found in Breitung *et al.* (2004) and more discussion of partitioning the shocks in permanent and transitory ones is given in Gonzalo and Ng (2001) and Fisher and Huh (1999) among others.

6.2.2 Computation of Impulse Responses and Forecast Error Variance Decompositions

If the matrix B is uniquely specified, impulse responses can be computed easily from the structural form parameters. Rewriting the VECM (2.1) in levels VAR form as

$$y_t = A_1 y_{t-1} + \cdots + A_p y_{t-p} + B\varepsilon_t,$$

and computing matrices

$$\Phi_i = \sum_{j=1}^{i} \Phi_{i-j} A_j, \quad i = 1, 2, \ldots,$$

with $\Phi_0 = I_K$ and $A_j = 0$ for $j > p$, the structural impulse response coefficients can be shown to be the elements of the matrices

$$\Theta_j = \Phi_j B, \quad j = 0, 1, 2, \ldots \tag{2.6}$$

(see Lütkepohl, 2005, for details).

Forecast error variance decompositions are alternative tools for analyzing the dynamic interactions between the variables of a VAR model. Denoting by $\omega_{kj}(h)$ the percentage contribution of variable j to the h-step forecast error variance of variable k, it can be shown that

$$\omega_{kj}(h) = (\theta_{kj,0}^2 + \cdots + \theta_{kj,h-1}^2) \Big/ \sum_{j=1}^{K} (\theta_{kj,0}^2 + \cdots + \theta_{kj,h-1}^2),$$

where $\theta_{kj,l}$ is the kj-th element of Θ_l. Because these forecast error variance components depend on the structural impulse responses, they also require identified innovations, that is, a uniquely specified matrix B, for a meaningful interpretation.

In practice, the parameters of the VECM are unknown and have to be estimated from the given time series data. This issue will be considered next. Computing the impulse responses and forecast error variance components from estimated rather than known parameters gives estimates of these quantities. Some implications of working with estimated impulse responses will also be considered in the next section.

6.3 Estimation

If the lag order and the cointegrating rank as well as structural identifying restrictions are given, estimation of a VECM can proceed by first estimating the reduced form parameters and then estimating B as described in the following.

6.3.1 Estimating the Reduced Form

The parameters of the reduced form VECM (2.1) can be estimated by the Johansen (1995) Gaussian maximum likelihood (ML) procedure. In presenting the estimators,

the following notation will be used, where a sample of size T and p presample values are assumed to be available: $\Delta Y = [\Delta y_1, \ldots, \Delta y_T]$, $Y_{-1} = [y_0, \ldots, y_{T-1}]$, $U = [u_1, \ldots, u_T]$, $\Gamma = [\Gamma_1 : \cdots : \Gamma_{p-1}]$ and $X = [X_0, \ldots, X_{T-1}]$ with

$$X_{t-1} = \begin{bmatrix} \Delta y_{t-1} \\ \vdots \\ \Delta y_{t-p+1} \end{bmatrix}.$$

Using this notation, the VECM (2.1) can be written compactly as

$$\Delta Y = \alpha\beta' Y_{-1} + \Gamma X + U. \tag{3.1}$$

Given a specific matrix $\alpha\beta'$, the equationwise least squares estimator of Γ is easily seen to be

$$\widehat{\Gamma} = (\Delta Y - \alpha\beta' Y_{-1})X'(XX')^{-1}. \tag{3.2}$$

Substituting this matrix for Γ in (3.1) and rearranging terms gives

$$\Delta Y M = \alpha\beta' Y_{-1} M + \widehat{U}, \tag{3.3}$$

where $M = I - X'(XX')^{-1}X$. Estimators for α and β can be obtained by canonical correlation analysis (see Anderson, 1984) or, equivalently, a reduced rank regression based on the model (3.3). Following Johansen (1995), the estimator may be determined by defining

$$S_{00} = T^{-1}\Delta Y M \Delta Y', \quad S_{01} = T^{-1}\Delta Y M Y'_{-1}, \quad S_{11} = T^{-1} Y_{-1} M Y'_{-1},$$

and solving the generalized eigenvalue problem

$$\det(\lambda S_{11} - S'_{01} S_{00}^{-1} S_{01}) = 0. \tag{3.4}$$

Denote the ordered eigenvalues by $\lambda_1 \geq \cdots \geq \lambda_K$ and the associated matrix of eigenvectors by $V = [b_1, \ldots, b_K]$. The generalized eigenvectors satisfy $\lambda_i S_{11} b_i = S'_{01} S_{00}^{-1} S_{01} b_i$ and they are assumed to be normalized such that $V' S_{11} V = I_K$. An estimator of β is then obtained by choosing

$$\widehat{\beta} = [b_1, \ldots, b_r]$$

and α is estimated as

$$\widehat{\alpha} = \Delta Y M Y'_{-1} \widehat{\beta} (\widehat{\beta}' Y_{-1} M Y'_{-1} \widehat{\beta})^{-1}, \tag{3.5}$$

that is, $\widehat{\alpha}$ may be regarded as the least squares estimator from the model

$$\Delta Y M = \alpha \widehat{\beta}' Y_{-1} M + \tilde{U}.$$

Using (3.2) a feasible estimator of Γ is $\widehat{\Gamma} = (\Delta Y - \widehat{\alpha}\widehat{\beta}' Y_{-1})X'(XX')^{-1}$. Under Gaussian assumptions, these estimators are ML estimators conditional on the presample values (Johansen, 1991, 1995). They are consistent and jointly asymptotically normal under more general assumptions than Gaussianity. The asymptotic distribution of $\widehat{\Gamma}$ is nonsingular so that standard inference may be used for the short-term parameters Γ_j.

Notice that for any nonsingular $(r \times r)$ matrix C, we may define $\alpha^* = \alpha C'$ and $\beta^* = \beta C^{-1}$ and get $\alpha\beta' = \alpha^*\beta^{*\prime}$. In order to estimate the matrices α and β consistently, it is necessary to impose identifying (uniqueness) restrictions. Without such restrictions, only the product $\alpha\beta'$ can be estimated consistently. An example of identifying restrictions which has received some attention in the literature, assumes that the first part of β is an identity matrix, $\beta' = [I_r : \beta'_{(K-r)}]$, where $\beta_{(K-r)}$ is a $((K-r)\times r)$ matrix. For instance, for $r = 1$, this restriction amounts to normalizing the coefficient of the first variable to be one. By a suitable rearrangement of the variables it can always be ensured that the normalization $\beta' = [I_r : \beta'_{(K-r)}]$ is possible. Test procedures exist for checking the normalization empirically if a proper ordering of the variables is not known a priori (Boswijk, 1996; Saikkonen, 1999).

Using this normalization, the parameters $\beta_{(K-r)}$ are identified so that inference becomes possible. Generally, if uniqueness restrictions are imposed, it can be shown that $T(\widehat{\beta} - \beta)$ and $\sqrt{T}(\widehat{\alpha} - \alpha)$ converge in distribution (Johansen, 1995). Hence, the estimator of β converges with the fast rate T and is therefore sometimes called *superconsistent*. In contrast, the estimator of α converges with the usual rate $\sqrt{T}$. It has an asymptotic normal distribution under general assumptions and, hence, it behaves like usual estimators in a model with stationary variables. In fact, its asymptotic distribution is the same that would be obtained if $\widehat{\beta}$ were replaced by the true cointegration matrix β and α were estimated by least squares from (3.3).

Although inference for α and β separately requires identifying restrictions, such constraints for α and β are not necessary for the impulse response analysis. In particular, the same matrices Ξ and Θ_j, $j = 0, 1, 2, \ldots$, are obtained for any pair of $(K \times r)$ matrices α and β that gives rise to the same product matrix $\alpha\beta'$.

6.3.2 Estimating the Structural Parameters

Replacing the reduced form parameters by their ML estimators gives the concentrated log-likelihood function

$$\log l_c(B) = \text{constant} - \frac{T}{2}\log|B|^2 - \frac{T}{2}\text{tr}(B'^{-1}B^{-1}\widetilde{\Sigma}_u), \tag{3.6}$$

where $\widetilde{\Sigma}_u = T^{-1}\sum_{t=1}^{T}\widehat{u}_t\widehat{u}'_t$ and the $\widehat{u}_t$'s are the estimated reduced form residuals. Maximization of this function with respect to B subject to the structural restrictions has to be done by numerical methods because a closed form solution is usually not available.

Suppose the structural restrictions for a VECM are given in the form of linear restrictions as in (2.5). For computing the parameter estimates, we may replace Ξ by its reduced form ML estimator,

$$\widehat{\Xi} = \widehat{\beta}_\perp\left[\widehat{\alpha}'_\perp\left(I_K - \sum_{i=1}^{p-1}\widehat{\Gamma}_i\right)\widehat{\beta}_\perp\right]^{-1}\widehat{\alpha}'_\perp,$$

and the restricted ML estimator of B can be obtained by optimizing the concentrated log-likelihood function (3.6) with respect to B, subject to the restrictions in (2.5), with C_l replaced by

$$\widehat{C}_l = C_{\Xi B}(I_K \otimes \widehat{\Xi})$$

(see Vlaar, 2004). Although this procedure results in a set of stochastic restrictions, from a numerical point of view we have a standard constrained optimization problem which can be solved by a Lagrange approach. Due to the fact that for a just-identified structural model the log-likelihood maximum is the same as for the reduced form, a comparison of the log-likelihood values can serve as a check for a proper convergence of the optimization algorithm used for structural estimation.

Under usual assumptions, the ML estimator of B, $\widehat{B}$ say, is consistent and asymptotically normal,

$$\sqrt{T}\mathrm{vec}(\widehat{B} - B) \xrightarrow{d} \mathcal{N}(0, \Sigma_{\widehat{B}}) .$$

Thus, the t-ratios of elements with regular asymptotic distributions can be used for assessing the significance of individual parameters if setting the associated element of B to zero is a valid restriction which leads to a nonsingular B matrix. The asymptotic distribution of $\widehat{B}$ is singular, however, because of the restrictions that have been imposed on B. Therefore F-tests will in general not be valid and have to be interpreted cautiously. Expressions for the covariance matrices of the asymptotic distributions in terms of the model parameters can be obtained by working out the corresponding information matrices (see Vlaar, 2004). For practical purposes, bootstrap methods are in common use for inference in this context.

Although in structural VAR and VECM analysis just-identified models are often used to minimize the risk of misspecification, the same approach can be used if there are over-identifying restrictions for B. In that case, $\widehat{B}\widehat{B}'$ will not be equal to the reduced form white noise covariance estimator $\widetilde{\Sigma}_u$, however. Still the estimator of B will be consistent and asymptotically normal under general conditions. The LR statistic,

$$\lambda_{LR} = T(\log|\widehat{B}\widehat{B}'| - \log|\widetilde{\Sigma}_u|) , \tag{3.7}$$

can be used to check the over-identifying restrictions. It has an asymptotic χ^2-distribution with degrees of freedom equal to the number of over-identifying restrictions if the null hypothesis holds.

6.3.3 Estimation of Impulse Responses

The impulse responses are estimated by replacing all unknown quantities in (2.6) by estimators. Suppose the structural form coefficients are collected in a vector $\boldsymbol{\alpha}$ and denote its estimator by $\widehat{\boldsymbol{\alpha}}$. For inference purposes it is important to note that any specific impulse response coefficient θ is a (nonlinear) function of $\boldsymbol{\alpha}$ and it is estimated as

$$\widehat{\theta} = \theta(\widehat{\boldsymbol{\alpha}}) . \tag{3.8}$$

If $\widehat{\boldsymbol{\alpha}}$ is asymptotically normal, that is,

$$\sqrt{T}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) \xrightarrow{d} \mathcal{N}(0, \Sigma_{\widehat{\boldsymbol{\alpha}}}) , \tag{3.9}$$

then $\widehat{\theta}$ is also asymptotically normally distributed,

$$\sqrt{T}(\widehat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma_\theta^2) , \tag{3.10}$$

and the variance of the asymptotic distribution is

$$\sigma_\theta^2 = \frac{\partial\theta}{\partial\boldsymbol{\alpha}'}\Sigma_{\widehat{\boldsymbol{\alpha}}}\frac{\partial\theta}{\partial\boldsymbol{\alpha}} . \tag{3.11}$$

Here $\partial\theta/\partial\boldsymbol{\alpha}$ denotes the vector of first order partial derivatives of θ with respect to the elements of $\boldsymbol{\alpha}$. The result (3.11) holds if σ_θ^2 is nonzero, which follows if $\Sigma_{\widehat{\boldsymbol{\alpha}}}$ is nonsingular and $\partial\theta/\partial\boldsymbol{\alpha} \neq 0$. In general the covariance matrix $\Sigma_{\widehat{\boldsymbol{\alpha}}}$ will not be nonsingular for cointegrated systems, however, due to the superconsistency of $\widehat{\boldsymbol{\beta}}$. Moreover, the impulse responses generally consist of sums of products of the VAR coefficients and, hence, the partial derivatives will also be sums of products of such coefficients. Therefore the partial derivatives will also usually be zero in parts of the parameter space. Thus, $\sigma_\theta^2 = 0$ may hold and, hence, $\widehat{\theta}$ may actually converge at a faster rate than $\sqrt{T}$ in parts of the parameter space (cf. Benkwitz *et al.*, 2000).

It was found, however, that even under ideal conditions where the asymptotic theory holds, it may not provide a good guide for small sample inference. Therefore bootstrap methods are often used to construct confidence intervals for impulse responses (e.g., Kilian, 1998; Benkwitz *et al.*, 2001). In the present context, these methods have the additional advantage that they avoid deriving explicit forms of the rather complicated analytical expressions for the asymptotic variances of the impulse response coefficients. Unfortunately, bootstrap methods generally do not overcome the problems due to zero variances in the asymptotic distributions of the impulse responses and they may provide confidence intervals which do not have the desired coverage level even asymptotically (see Benkwitz *et al.*, 2000, for further discussion).

Although we have discussed the estimation problems in terms of impulse responses, similar problems arise for forecast error variance components. In fact, these quantities are proportions and they are therefore always between zero and one. In other words, they are bounded from below and above. Moreover, the boundary values are possible values as well. This feature makes inference even more delicate.

So far it was assumed that a model and identifying structural restrictions are given. In practice this is usually not the case. While the structural restrictions normally come from theoretical considerations or institutional knowledge, there is a range of statistical tools for specifying the reduced form of a VECM. These tools will be summarized briefly in the next section.

6.4 Model Specification and Validation

The general approach to structural VECM analysis is to specify a reduced form first and then impose structural restrictions that can be used in an impulse response analysis. To specify the reduced form VECM, the lag order and the cointegrating rank have to be chosen. Most procedures for specifying the latter quantity require that the lag order is already known whereas order selection can be done without prior knowledge of the cointegrating rank. Therefore lag order selection is typically based on a VAR process in levels without imposing a cointegration rank restriction. Standard model selection criteria of the form

$$\mathrm{Cr}(m) = \log\det(\widetilde{\Sigma}_u(m)) + c_T\varphi(m) \tag{4.1}$$

can be used for that purpose. Here $\widetilde{\Sigma}_u(m) = T^{-1}\sum_{t=1}^{T}\widehat{u}_t\widehat{u}_t'$ is the residual covariance matrix estimator for a model with lag order m and $\varphi(m)$ is a function

which penalizes large VAR orders. For instance, $\varphi(m)$ may represent the number of parameters which have to be estimated in a VAR(m) model. The quantity c_T is a sequence that depends on the sample size T. For example, for Akaike's AIC, $c_T = 2/T$ and for the popular Hannan-Quinn criterion, $c_T = 2 \log\log T/T$. The term $\log\det(\widetilde{\Sigma}_u(m))$ measures the fit of a model with order m. It decreases (or at least does not increase) when m increases because there is no correction for degrees of freedom in the covariance matrix estimator. The criterion chosen by the analyst is evaluated for $m = 0, \ldots, p_{\max}$, where $p_{\max}$ is a prespecified upper bound and the order p is estimated so as to minimize the criterion. Rewriting the levels VAR(p) model in VECM form, there are $p - 1$ lagged differences that may be used in the next stage of the analysis, where the cointegrating rank is chosen.

Once the lag order is specified the cointegrating rank can be chosen by defining the matrix $\Pi = \alpha\beta'$ and testing a sequence of null hypotheses $H_0(0) : \mathrm{rk}(\Pi) = 0$, $H_0(1) : \mathrm{rk}(\Pi) = 1, \ldots, H_0(K-1) : \mathrm{rk}(\Pi) = K - 1$ against the rank being greater than the one specified in the null hypothesis. The rank for which the null hypothesis cannot be rejected for the first time is then used in the next stages of the analysis. A range of test statistics is available for use in this testing sequence (see, e.g., Hubrich *et al.*, 2001, for a recent survey). The most popular tests in applied work are Johansen's (1995) likelihood ratio tests. They are easy to compute because the Gaussian likelihood function is easy to maximize for any given cointegrating rank, as shown in Section 3.1.

When a reduced form model has been specified, a range of tools can be used for model checking. For example, tests for residual autocorrelation and structural stability may be used (see Lütkepohl, 2005, for details). Finally, once a satisfactory reduced form is available, the structural restrictions may be imposed and the model can then be used for impulse response analysis.

6.5 Conclusions

In this article a brief overview of some important issues related to structural modelling based on VARs with cointegrated variables was given. Generally, using a standard VAR analysis, the impulse responses are the relevant tools for interpreting the relationships between the variables. Unfortunately, they are not unique and subject matter knowledge is required to specify those impulses and their associated responses which reflect the actual ongoings in the system of interest. It was discussed how the cointegration properties of the variables can help in specifying identifying restrictions properly. In particular, the cointegrating rank specifies the maximum number of transitory shocks in a system with cointegrated variables. This rank in turn can be determined by statistical procedures. As a final note it may be worth mentioning that the software JMulTi (Lütkepohl and Krätzig, 2004) provides easy access to the necessary computations for a structural VECM analysis.

References

AMISANO, G., GIANNINI, C. (1997). *Topics in Structural VAR Econometrics.* 2nd ed., Springer, Berlin.

ANDERSON, T. W. (1984). *An Introduction to Multivariate Statistical Analysis.* John Wiley, New York.

BENKWITZ, A., LÜTKEPOHL, H., NEUMANN, M. (2000). Problems related to bootstrapping impulse responses of autoregressive processes. *Econometric Reviews* **19** 69–103.

BENKWITZ, A., LÜTKEPOHL, H., WOLTERS, J. (2001). Comparison of bootstrap confidence intervals for impulse responses of German monetary systems. *Macroeconomic Dynamics* **5** 81–100.

BOSWIJK, H. P. (1996). Testing identifiability of cointegrating vectors. *Journal of Business & Economic Statistics* **14** 153–160.

BREITUNG, J., BRÜGGEMANN, R., LÜTKEPOHL, H. (2004). Structural vector autoregressive modeling and impulse responses. In *Applied Time Series Econometrics* (H. Lütkepohl, M. Krätzig, eds.), pp. 159–196, Cambridge University Press, Cambridge.

ENGLE, R. F., GRANGER, C. W. J. (1987). Cointegration and error correction: Representation, estimation and testing. *Econometrica* **55** 251–276.

FISHER, L. A., HUH, H. (1999). Weak exogeneity and long-run and contemporaneous identifying restrictions in VEC models. *Economics Letters* **63** 159–165.

GONZALO, J., NG, S. (2001). A systematic framework for analyzing the dynamic effects of permanent and transitory shocks. *Journal of Economic Dynamics & Control* **25** 1527–1546.

GRANGER, C. W. J. (1981). Some properties of time series data and their use in econometric model specification. *Journal of Econometrics* **16** 121–130.

GRANGER, C. W. J., NEWBOLD, P. (1974). Spurious regressions in eonometrics. *Journal of Econometrics* **2** 111–120.

HUBRICH, K., LÜTKEPOHL, H., SAIKKONEN, P. (2001). A review of systems cointegration tests. *Econometric Reviews* **20** 247–318.

JOHANSEN, S. (1988). Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control* **12** 231–254.

JOHANSEN, S. (1991). Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models. *Econometrica* **59** 1551–1581.

JOHANSEN, S. (1995). *Likelihood-based Inference in Cointegrated Vector Autoregressive Models.* Oxford University Press, Oxford.

KILIAN, L. (1998). Small-sample confidence intervals for impulse response functions. *Review of Economics and Statistics* **80** 218–230.

KING, R. G., PLOSSER, C. I., STOCK, J. H., WATSON, M. W. (1991). Stochastic trends and economic fluctuations. *American Economic Review* **81** 819–840.

LÜTKEPOHL, H. (2005). *New Introduction to Multiple Time Series Analysis.* Springer, Berlin.

LÜTKEPOHL, H., KRÄTZIG, M. (EDS.) (2004). *Applied Time Series Econometrics.* Cambridge University Press, Cambridge.

SAIKKONEN, P. (1999). Testing normalization and overidentification of cointegrating vectors in vector autoregressive processes. *Econometric Reviews* **18** 235–257.

SIMS, C. A. (1980). Macroeconomics and reality. *Econometrica* **48** 1–48.

VLAAR, P. J. G. (2004). On the asymptotic distribution of impulse response functions with long-run restrictions. *Econometric Theory* **20** 891–903.

7 Econometric Analysis of High Frequency Data

Helmut Herwartz[1]

[1]Institut für Statistik und Ökonometrie, Christian Albrechts–Universität zu Kiel
herwartz@stat-econ.uni-kiel.de

Summary: Owing to enormous advances in data acquisition and processing technology the study of high (or ultra) frequency data has become an important area of econometrics. At least three avenues of econometric methods have been followed to analyze high frequency financial data: Models in tick time ignoring the time dimension of sampling, duration models specifying the time span between transactions and, finally, fixed time interval techniques. Starting from the strong assumption that quotes are irregularly generated from an underlying exogeneous arrival process, fixed interval models promise feasibility of familiar time series techniques. Moreover, fixed interval analysis is a natural means to investigate multivariate dynamics. In particular, models of price discovery are implemented in this venue of high frequency econometrics. Recently, a sound statistical theory of 'realized volatility' has been developed. In this framework high frequency log price changes are seen as a means to observe volatility at some lower frequency.

7.1 Introduction

With the enormous advances in computer technology, data acquisition, storage and processing has become feasible at higher and higher frequencies. In the extreme case of ultra high frequency financial data the analyst has access to numerous characteristics, called marks, of each transaction (price and quantity traded, corresponding bid and ask quotes etc.) and to the time of its occurrence, measured in seconds. As a consequence, numerous financial market microstructure hypotheses undergo empirical tests based on ultra frequency data. Typical issues in this vein of microstructure analysis are, for instance, the informational content of traded volumes for (future) prices (Karpoff, 1987), the relation between prices and clustering of transactions (Easley and O'Hara, 1992), or the significance of bid ask spreads as a means to identify the presence of informed traders in the market (Admati and Pfleiderer, 1988). From an econometric perspective such hypotheses naturally require an analysis of the marks in tick time, and, eventually motivate a duration model. The

methodology for the analysis of marked point processes as well as durations has experienced substantial progress since the introduction of Autoregressive Conditional Duration (ACD) models by Engle and Russell (1998). For a recent overview the reader may consult Engle and Russell (2005).

Another area of market microstructure modeling is information diffusion across markets trading the same asset or close substitutes. Then, it is of interest if independent price discovery (Schreiber and Schwartz, 1986) takes place in some major market or, alternatively, if the efficient price is determined over a cross section of interacting exchanges. Following Harris *et al.* (1995) or Hasbrouck (1995) price discovery is investigated by means of vector error correction models (VECM) mostly after converting transaction time to fixed time intervals of 1, 10 or 30 minutes, say. Although the latter conversion goes at the cost of loosing information on the trading intensity, it appears inevitable since the price quotations of interest are collected as a vector valued variable. Owing to irregular time spacing of quotes the statistical analysis of fixed interval data has to cope with methodological issues arising from the incidence of missing values. A condensed review over econometric approaches to model price discovery will be given in Section 7.2.

Apart from market microstructure modeling high frequency data have recently attracted large interest in econometrics as a means to estimate conditional volatility of asset prices at lower frequencies (Anderson *et al.*, henceforth, ABDL, 2001, 2003). Owing to its consistency for the process of conditional variances this estimator has particular appeal since it makes the latent volatility observable in the limit. A sound statistical theory on 'realized volatility' is now available making it a strong competitor to parametric approaches to modeling time varying second order moments. Section 7.3 will provide theoretical and empirical features of 'realized volatility'.

7.2 Price Discovery

A particular issue in empirical finance is the analysis of dynamic relationships between markets trading simultaneously a given security. Since cross sectional price differentials cannot persist, it is of interest, if the involved market places contribute jointly to the fundamental value of the asset or if particular markets lead the other. The process of incorporating new information into the efficient price has become popular as price discovery (Schreiber and Schwartz, 1986).

Starting from stylized features of high frequency data this section will first discuss the scope of fixed interval techniques. Then, the VECM is motivated as a means to address price discovery and a few empirical results are provided. A formal measure of a particular market's contribution to price discovery is derived along with a brief formalization of the VECM. Finally, parameter estimation in case of missing values is discussed.

7.2.1 Nonsynchronous Trading and Fixed Interval Analysis

Financial markets do not operate continuously but rather show strong concentration of activity. Intraday periodicity of marks has been documented e.g. by Engle and Russell (1998). On stock markets, for instance, volatility and trade frequency show distinct U-shaped patterns over the trading day. Intradaily seasonality is also typical for the decentralized foreign exchange (FX) markets which are open 24 hours a day, seven days a week (Dacorogna *et al.*, 1993). Making the strong assumption that marks are irregularly generated from an underlying discrete time process such that the order arrival and the quote process are independent (Ghysels *et al.*, 1997) quotations are conveniently obtained as the last observation recorded in a particular time interval of fixed length, one minute say. Alternatively some interpolation scheme may be applied.

Although the analysis of intradaily return processes conditional on observed patterns of quoting might suffer from interdependence between quotes and intensity such an approach is almost indispensable if relationships between informationally linked markets are investigated. The extent to which time dependence affects the stochastic properties of marks remains an empirical question. Goodhart and O'Hara (1997) conclude that, empirically, time dependence is still an open issue and is likely problem specific. Jordá and Marcellino (2003) argue that numerous features which are typically reported for high frequency FX rates, e.g. intraday seasonality and volatility clustering, may be attributed to irregular sampling from a homogeneous and continuous log price process, as, for instance, a random walk. Interestingly, they find by means of Monte Carlo simulation that sampling at fixed equidistant intervals best recovers the properties of the underlying process.

7.2.2 Motivation and Applications of VECM

To characterize the dynamics of price discovery VECM (Engle and Granger, 1987; Johansen, 1995) have been motivated by Harris *et al.* (1995). For the econometric analysis it is assumed that (log) price processes are integrated of order one, i.e. stationary after taking first differences. Arbitrage trading will guarantee that cross sectional prices do not persistently deviate from each other. Thus, the VECM representation for cointegrated time series variables is a suggestive framework for the econometric analysis of price discovery.

De Jong *et al.* (1998) and Herwartz (2001) analyze joint dynamics of the JPY/DEM quotes and an implied price constructed from the USD exchange rates of these currencies. It is found that the highly liquid USD markets and the JPY/DEM market both contribute to JPY/DEM price formation. Grammig *et al.* (2005) confirm for three major German stocks with cross listed counterparts at the New York Stock Exchange (NYSE) that the home market (Frankfurt Stock Exchange) is the major source of price discovery but fails to determine the fundamental asset price completely. Adopting a VECM specified in tick time Harris *et al.* (1995) investigate discovery of IBM stock prices by the NYSE and other US stock exchanges in the Midwest and Pacific area. According to their findings NYSE prices show a stronger response to interregional price differentials in comparison with the remaining mar-

ket places. Huang (2002) estimates the contribution of two particular electronic communication networks (ECN) and Nasdaq market makers to price discovery for the 30 most heavily traded shares at Nasdaq in June 1998 and November 1999. It is found that the purely anonymous trade operating via ECN contributes substantially to price formation, and, in addition, the relative contribution of ECN has vastly increased over time.

7.2.3 The Vector Error Correction Model

Let $p_t = (p_{1t}, p_{2t}, \ldots, p_{Nt})'$ denote a vector of prices for one security observed in time t over a cross section of N markets. The VECM of order $(q-1)$ is derived from an unrestricted vector autoregression of order q and reads as (Johansen, 1995):

$$\Delta p_t = \Pi p_{t-1} + \Gamma_1 \Delta p_{t-1} + \ldots + \Gamma_{p-1} \Delta p_{t-p+1} + u_t \,. \tag{7.1}$$

The error vector u_t is serially uncorrelated with covariance matrix $E[u_t u_t'] = \Omega = [\omega_{ij}]$, $i, j = 1, \ldots, N$. In the present context u_t may be interpreted as innovations to the fundamental value of the asset. Market efficiency will ensure that differentials between integrated log prices, as e. g. $p_{1t} - p_{2t}$, $p_{1t} - p_{3t}$, $\ldots$, $p_{1t} - p_{Nt}$, will be stationary, and thereby provide $N-1$ cointegrating relationships. Owing to cointegration the matrix Π in (7.1) has rank $N-1$ and allows the factorization $\Pi = \alpha\beta'$, where α is a $N \times (N-1)$ matrix and β' is e.g.

$$\beta' = (\mathbf{1}, I_{N-1}^{(-)}) \,, \tag{7.2}$$

with $\mathbf{1}$ and $I_{N-1}^{(-)}$ denoting a unit column vector and minus the identity matrix of dimension $N-1$, respectively. Weak exogeneity of a particular market j will result in rowwise zero restrictions in α, $\alpha_j = 0$, and implies that market j does not adjust towards a new equilibrium in direct response to price differentials. Apart from this assessment of market j's short run contribution to price discovery, the market specific impact on equilibrium prices is also of interest. Formalizing the latter Hasbrouck (1995) introduced a measure closely related to the forecast error variance decompositions. To derive the relative contribution of market j to prices in the long run it is convenient to start from the MA-representation of stationary price changes

$$\Delta p_t = \Lambda(L) u_t, \; \Lambda(L) = \Lambda_0 + \Lambda_1 L + \Lambda_2 L^2 + \ldots \,. \tag{7.3}$$

Integrating (7.3) one obtains the so-called common trend representation of p_t, i.e.

$$p_t = p_0 + \Lambda(1)\xi_t + \Lambda^*(L) u_t, \; \xi_t = \sum_{i=1}^{t} u_i \,, \tag{7.4}$$

where $\Lambda^*(z)$ contains scalar polynomials in z. According to the interpretation of u_t carrying market news the quantity $\Lambda(1) \sum_{i=1}^{t} u_i$ measures the permanent impact of news on prices. Owing to cointegration linking the variables in p_t $\beta'\Lambda(1) = 0$ (Engle and Granger, 1987). From the particular structure of β' given in (7.2) it follows that the N rows in $\Lambda(1)$ are identical. Let λ denote any row of $\Lambda(1)$ and λ_j its typical element. Then, the total variance of the common stochastic trend is

$\lambda\Omega\lambda'$. In case of a diagonal structure of Ω the latter is comprised by N distinct sources of news and, thus, the share of total variance going back to the j-th market is

$$s_j = \frac{\lambda_j^2 \omega_{jj}}{\lambda\Omega\lambda'} . \tag{7.5}$$

In empirical practice, however, Ω is likely not diagonal, i.e. error terms in u_t exhibit contemporaneous correlation. Therefore one may use a Cholesky factorization of Ω to isolate the underlying structural innovations e_t as

$$e_t = C^{-1} u_t \Leftrightarrow u_t = Ce_t, \Omega = CC', C \text{ lower triangular.}$$

Then, the contribution of market j to price discovery is

$$s_j^* = \frac{((\lambda C)_j)^2}{\lambda\Omega\lambda'} . \tag{7.6}$$

Similar to orthogonalized impulse responses (Lütkepohl, 1991), however, s_j^* will depend on the ordering of markets in p_t. Therefore alternative orderings of the variables in p_t should be studied to obtain upper and lower bounds for s_j^*. Hasbrouck (1995) points out that contemporaneous correlation in Ω is likely to be a consequence of temporal aggregation. Thus, sampling at very high frequencies will to some extent guard against spurious interpretations of the information shares defined above. Frijns and Schotman (2003) propose an information criterion assessing the prevalence of microstructure noise in dealer quotes. Owing to its dependence on the employed calendar time intervals it could be used to determine a particular sampling frequency in a data driven way.

7.2.4 Parameter Estimation with Incomplete Samples

Fixed interval samples may consist of both, price quotations and missing values. To cope with such sampling schemes de Jong and Nijman (1997) introduce a moment estimator for VECM. As a competing procedure one may regard quasi maximum likelihood (QML) estimation using the Kalman Filter for recursive log likelihood evaluation of a state space representation implied by (7.1) (Kohn and Ansley, 1986; Jones, 1980). Apart from capturing irregular sampling, the state space form will allow numerous generalizations of linear VECM which are natural when analyzing dynamics of empirical price processes.

Within the state space framework an observable output ($\mathbf{x}_t$) depends on the state of a dynamic system in time t (ξ_t), exogenous influences (w_t), and some unpredictable zero mean errors (η_t). The state itself depends on the previous state and an unpredictable error (ϵ_t). Formally these dependencies are denoted by means of the so-called observation and state equation, respectively:

$$\underset{(n\times 1)}{\mathbf{x}_t} = \underset{(n\times w)(w\times 1)}{B\ w_t} + \underset{(n\times r)(r\times 1)}{H\ \xi_t} + \underset{(n\times 1)}{\eta_t} , \tag{7.7}$$

$$\underset{(r\times 1)}{\xi_t} = \underset{(r\times r)(r\times 1)}{F\ \xi_{t-1}} + \underset{(r\times 1)}{\epsilon_t} . \tag{7.8}$$

For QML estimation both vector processes η_t and ϵ_t are assumed to follow a multivariate normal distribution and

$$E[\eta_t \eta_s'] = \begin{cases} Q \text{ if } t = s \\ 0 \quad \text{else} \end{cases}, \; E[\, \epsilon_t \epsilon_s'] = \begin{cases} R \text{ if } t = s \\ 0 \quad \text{else} \end{cases}, \; E[\eta_t \epsilon_s] = 0 \text{ for all } t, s\,.$$

Setting $B = Q = 0$ the model in (7.1) can be given as a special case of (7.7) and (7.8). In this case the dynamic model is specified for the log price series. The observation equation becomes an identity and R is just a zero matrix except for its upper left block which is equal to Ω. The specification in (7.7) and (7.8) is already sufficient to nest a variety of dynamic models including cointegrated moving average (MA) specifications employed by De Jong *et al.* (1998). Notwithstanding, it is worthwhile to mention a few veins of generalizations that may be particularly fruitful for an analysis of fixed interval financial data.

Firstly, it should be noted that all parameter matrices B, H, F, Q and R may exhibit deterministic or stochastic time dependence. In the former case one may think of periodic systems formalizing intraday seasonalities. Stochastic parameter variation is typically implemented by relating model parameters to predetermined variables. To provide a particular example, the assumption of time invariant covariances Q and R may be criticized when modeling financial returns, often showing marked volatility clustering. Popular models to account for time varying second order moments, as the family of (generalized) autoregressive conditional heteroskedastic (G)ARCH models (Engle, 1982; Bollerslev, 1986) or stochastic volatility (SV) models (Taylor, 1986), may be implemented via a state space model (Goodhart *et al.*, 1993). The normality assumption may be justified by means of QML theory but, secondly, it is likely at odds with stylized features of high frequency data. The Kalman Filter, however, may also be generalized to take non normality of innovations explicitly into account when implementing the iterative updating schemes. Thirdly, the linear relationships characterizing the conditional mean of the model in (7.1) can be generalized to capture nonlinear error correction dynamics formalized, for instance, by means of smooth transition (Teräsvirta, 1994) as in Herwartz (2001). Taking transaction costs implicitly into account one may also formalize a threshold cointegration model (Lo and Zivot, 2001).

7.3 Realized Volatility

Volatility clustering characterizes price processes observed at speculative markets at ultra (by transaction), high, intermediate (daily) to lower frequencies (weekly, monthly). In the sequel of its introduction the GARCH model and its numerous parametric, semi- and nonparametric variants as well as SV models have been successfully applied in empirical studies of higher order dynamics of speculative prices. Merely the number of competing approaches capturing some but hardly all stylized facts of empirical returns indicates that merits and weaknesses of particular models are to some degree sample specific. Therefore one may a-priori opt for a model free approach to volatility estimation which has recently been popularized under the notion of 'realized volatility' by ABDL (2001, 2003) or Barndorff-Nielsen and Shephard, henceforth BS, (2002a,b).

This section will first note some early applications and recent contributions to the theory of realized volatility. Then, two asymptotic features of the variance estimator will be discussed in turn, namely consistency and conditional normality. Empirical properties of realized variance are collected in an own subsection, as well as recent modifications of the estimator.

7.3.1 Measuring Volatility from High Frequency Data

The basic idea of estimating lower frequency variances by summing up the squares of uncorrelated higher frequency returns has some tradition in empirical finance. For instance, Schwert (1989) constructs monthly (stock) variance estimates as a sum over squared daily returns and, similarly, Schwert (1990) exploits intraday returns to estimate daily variances. A sound statistical theory on realized volatility is, however, the focus of numerous recent contributions to financial econometrics as e.g. ABDL (2001, 2003), BS (2002a,b). For a detailed review over the field the reader may consult Andersen *et al.* (2005). Owing to both, computational feasibility and theoretical underpinning, realized volatility methods suggest themselves also for an analysis of (realized) conditional covariances (BS 2004, Andersen *et al.*, 2004, henceform ABDW) - an area of empirical finance where multivariate parametric models, GARCH or SV, crucially suffer from the curse of dimensionality.

Since the statistical theory on realized variance has gained substantially from the central limit theory in BS (2002a,b) going beyond the consistency of realized variance (ABDL, 2001, 2003) the discussion of asymptotic properties is mostly taken from BS (2002a). The statistical concepts are provided for the univariate case. To familiarize with the theory of realized covariation or realized beta the reader is referred to BS (2004) or ABDW (2004).

7.3.2 Consistency of Realized Variances

It is assumed that in continuous time the log price process of a speculative asset $(p(\tau), \tau \geq 0)$ is a special semimartingale (Back, 1991) and may be given as the solution of a stochastic differential equation (SDE)

$$dp(\tau) = \mu(\tau)d\tau + s(\tau)dW(\tau)\,, \tag{7.9}$$

where $\mu(\tau)$ is a possibly time varying drift term. The spot volatility process$(s(\tau), \tau \geq 0)$ is strictly stationary by assumption, locally square integrable and independent of the standard Brownian motion $W(\tau)$. Possible candidates to formalize the spot variance process are provided, for instance, within the class of constant elasticity of variance processes nesting in particular the GARCH diffusion model (Nelson, 1990). Positive Ornstein-Uhlenbeck processes are considered by BS (2002a, b). It is worthwhile to point out that the model in (7.9) allows $\mu(\tau)$ to depend on $s^2(\tau)$, e. g. $\mu(\tau) = \alpha\tau + \beta s^2(\tau)$, thereby formalizing a risk premium. Note that in case with constant drift and variance the SDE in (7.9) collapses to the model introduced by Merton (1973).

Discrete compounded returns measured over a sequence of intervals of length δ (one

day, say) are obtained as

$$r_t = p(t\delta) - p((t-1)\delta),\ t = 1, 2, \ldots . \tag{7.10}$$

>From the specification in (7.9) it is apparent that the latter will exhibit a mixed normal distribution, i. e.

$$r_t|\sigma_t^2 \sim N(0, \sigma_t^2),\ \sigma_t^2 = \sigma^2(t\delta) - \sigma^2((t-1)\delta),\ \sigma^2(\tau) = \int_0^\tau s^2(u)du\,. \tag{7.11}$$

Following BS (2002a) σ_t^2 will be referred to as the actual variance. As specified, the log price process $p(\tau) = \int_0^\tau dp(u)du$ allows to recover the integrated variance $\sigma^2(\tau) = \int_0^\tau ds^2(u)du$. For this purpose consider a sequence of partitions of the interval $[0, \tau]$, $0 = \tau_0^q < \tau_1^q < \ldots, < \tau_{M_q}^q = \tau$, where $\sup_m(\tau_{m+1}^q - \tau_m^q) \to 0$ as $q \to \infty$. Then, the theory of quadratic variation (Protter, 1990) yields asymptotically

$$\text{plim}_{q\to\infty}\left[\sum_{m=1}^{M_q}(p(\tau_m^q) - p(\tau_{m-1}^q))^2\right] = \sigma^2(\tau)\,.$$

Note that asymptotic theory applies with regard to the number of intra-interval price observations (M_q). The latter result may be specialized to estimate actual daily variances from intraday returns ($r_{m,t}$). To formalize the latter, sampling at an equidistant grid of time instants $t - l_0\delta, t - l_1\delta, \ldots,$
$t - l_M\delta, l_i = (M-i)/M$, is assumed without loss of generality, i. e.

$$r_{m,t} = p\left((t-1)\delta + \frac{m\delta}{M}\right) - p\left((t-1)\delta + \frac{(m-1)\delta}{M}\right), m = 1, \ldots, M\,. \tag{7.12}$$

Then, the sum of intraday squared price changes

$$\hat{\sigma}_t^2 = \sum_{m=1}^{M} r_{m,t}^2 \tag{7.13}$$

is an estimate of σ_t^2 called realized variance. To illustrate the basic mechanism obtaining $\hat{\sigma}_t^2$ as a consistent estimator of σ_t^2 consider the issue of estimating the daily constant variance, δs^2, in the Merton (1973) model by means of rescaled intradaily returns observed over day t:

$$\hat{\sigma}_t^2 = \delta\hat{s}^2 = \frac{\delta}{M}\sum_{m=1}^{M} r_{m,t}^2\left(\frac{\delta}{M}\right)^{-1} = \sum_{m=1}^{M} r_{m,t}^2\,. \tag{7.14}$$

Since

$$E[r_{m,t}^2] = s^2\left(\frac{\delta}{M}\right) + \mu^2\left(\frac{\delta}{M}\right)^2$$
$$\text{and } E[r_{m,t}^4] = 3s^4\left(\frac{\delta}{M}\right)^2 + 6\mu^2 s^2\left(\frac{\delta}{M}\right)^3 + \mu^4\left(\frac{\delta}{M}\right)^4,$$

the variance of squared intraday returns is seen to be

$$\mathrm{Var}[r_{m,t}^2] = 2s^4\left(\frac{\delta}{M}\right)^2 + 4\mu^2 s^2\left(\frac{\delta}{M}\right)^3 . \tag{7.15}$$

Now, $\mathrm{Var}[\hat{s}^2] \xrightarrow{p} 0$ as $M \to \infty$ such that $\delta\hat{s}^2$ is consistent for the actual daily variance δs^2. Note that the realized variance is unbiased in case $\mu(\tau) = 0$.

The former observation that the use of high frequency returns will deliver a consistent estimator for the variance parameter (7.14) is well established since Merton (1973). A central contribution of the theory underlying realized volatility is that basically the same arguments carry over to the case of time varying but continuous spot variance. Moreover, it is worthwhile to point out that realized variance remains consistent even in case the log price differential in (7.9) is augmented with a jump component (ABDL, 2001, 2003). Similar to the speciality of the SDE in (7.9), it is essential that potentially deterministic parts of the jump process are offset by jump innovation risk to keep the risk return relationship balanced. Although consistency of realized variance is maintained in presence of jumps there is not yet a central limit theory of realized volatility available allowing for jumps. For this reason the less general class of log price processes in (7.9) is considered here.

In the construction of the realized variance, apparently, possible intra-interval seasonalities of the spot variance $s^2(\tau)$ is safely ignored. Intraday seasonality of volatility is, however, well documented and deserves particular attention when adopting parametric volatility models. In case of constant spot variance the result in (7.15) is also informative for the rate of convergence. Rescaling $\hat{s}^2$ by $\sqrt{M}$ will deliver an estimator with nondegenerate variance.

For the more general class of stochastic volatility models BS (2002a) start from defining a realized variance error as

$$\begin{aligned} u_t &= \hat{\sigma}_t^2 - \sigma_t^2 && (7.16)\\ &\xrightarrow{d} \sum_{m=1}^{M} \sigma_{m,t}^2(\varepsilon_{m,t}^2 - 1),\ \varepsilon_{m,t} \sim \mathrm{iid}\, N(0,1) . && (7.17) \end{aligned}$$

It is seen that unconditionally

$$\begin{aligned} \mathrm{Var}[u_t] &= 2ME[(\sigma_{1,t}^2)^2] \\ &= 2M(\mathrm{Var}[\sigma_{1,t}^2] + (E[\sigma_{1,t}^2])^2) , \end{aligned}$$

where intradaily variances $\sigma_{m,t}^2$ are similarly defined as returns in (7.12), i. e.

$$\sigma_{m,t}^2 = \sigma^2\left((t-1)\delta + \frac{m\delta}{M}\right) - \sigma^2\left((t-1)\delta + \frac{(m-1)\delta}{M}\right), \ m = 1, \ldots, M .$$

The unconditional moments of $\sigma_{1,t}^2$ will depend on the corresponding quantities of $s^2(\tau)$. Taking the expectation one has immediately

$$ME[\sigma_{1,t}^2] = \delta E[s^2(\tau)] .$$

Moreover, as shown in BS (2002a),

$$\lim_{M\to\infty}(M^2\mathrm{Var}[\sigma_{1,t}^2]) = \delta^2\mathrm{Var}[s^2(\tau)]\,. \tag{7.18}$$

Combining the latter two results it turns out that for appropriately rescaled realized variance errors a nondegenerate distribution is obtained, since

$$\lim_{M\to\infty}\mathrm{Var}[\sqrt{M}(\hat{\sigma}_t^2-\sigma_t^2)] = 2\delta^2(\mathrm{Var}[s^2(\tau)] + (E[s^2(\tau)])^2)\,. \tag{7.19}$$

7.3.3 Conditional Normality of Realized Variances

So far, consistency of realized variance has been derived mainly under the assumption of local square integrability of the spot variance process $s^2(\tau)$. To derive the asymptotic distribution of the realized variance error in (7.19) BS (2002a) make the stronger assumption that $s^2(\tau)$ is of locally bounded variation, implying the (local) existence of fourth order moments of $s(\tau)$. Note that the strategy of strengthening the underlying set of assumptions to obtain stronger asymptotic results parallels the case of dependent linear processes (White, 1984). In case $s^2(\tau)$ is of locally bounded variation as $M\to\infty$

$$\sqrt{\frac{M}{2\delta}}\frac{(\hat{\sigma}_t^2-\sigma_t^2)}{\sqrt{\sigma_t^4}} \xrightarrow{d} N(0,1),\ \sigma_t^4 = \int_{(t-1)\delta}^{t\delta} s^4(u)du\,. \tag{7.20}$$

Practical application of the result in (7.20) is, however, infeasible since the process of fourth order moments, $s^4(\tau)$, is not observed. Similar to the quadratic variation result, however, a consistent estimator for σ_t^4 is available from

$$\frac{M}{\delta}\sum_{m=1}^{M} r_{m,t}^4 \xrightarrow{p} 3\sigma_t^4\,. \tag{7.21}$$

After substitution of unknown quantities in (7.20) by consistent estimators given in (7.21) a feasible counterpart of (7.20) is obtained as

$$\sqrt{\frac{3}{2}}\frac{(\hat{\sigma}_t^2-\sigma_t^2)}{\sqrt{\sum_{m=1}^{M} r_{m,t}^4}} \xrightarrow{d} N(0,1)\,. \tag{7.22}$$

It is worthwhile to underscore that the result in (7.20) holds even in case of a nonzero drift term $\mu(\tau)$ in (7.9) such that the deterministic part of the SDE has only third order effects on realized variance. An implication of the central limit result in (7.20) is that unconditionally realized variance errors will have a mixed normal distribution as $M\to\infty$. Investigating the accuracy of the asymptotic approximation for finite M BS (2005) simulate volatility diffusions as positive Ornstein-Uhlenbeck processes and find that for moderate values of M the tails of the conditional distribution in (7.22) differ considerably from the normal limit. For log transformed realized variances, however, the implied asymptotic results

$$\sqrt{\frac{M}{2\delta}}\frac{(\ln(\hat{\sigma}_t^2)-\ln(\sigma_t^2))}{\sqrt{\sigma_t^4/(\sigma_t^2)^2}} \xrightarrow{d} N(0,1)$$

with feasible counterpart

$$\sqrt{\frac{3}{2}} \frac{(\ln(\hat{\sigma}_t^2) - \ln(\sigma_t^2))}{\sqrt{(\sum_{m=1}^{M} r_{m,t}^4)/(\hat{\sigma}_t^2)^2}} \xrightarrow{d} N(0,1)$$

are almost achieved for $M = 48$.

7.3.4 Stylized Features of Realized Volatility

In the following, stylized facts of realized variances estimated throughout for the daily frequency are provided. Moreover, some results on realized covariance and correlation are stated briefly.

Constructing realized variance measures from 30 minute log changes of the DEM / USD and JPY / USD exchange rates ABDL (2003) document firstly that over a ten year period the unconditional distribution of daily log FX rate changes standardized by realized volatility $\hat{\sigma}_t$ is well described by the Gaussian distribution. It is worthwhile to point out that parametric volatility models, GARCH or SV, mostly deliver innovation estimates which exhibit significant excess kurtosis. The latter experience has motivated the use of leptokurtic innovation distributions in parametric volatility models (Bollerslev, 1987) or semiparametric treatment (Engle and Gonzalez-Rivera, 1991). The unconditional distribution of realized volatility ($\hat{\sigma}_t$) is skewed to the right which is even more pronounced for the realized variances ($\hat{\sigma}_t^2$). In contrast, the distribution of the natural logarithm of realized volatility ($\ln(\hat{\sigma}_t)$) is (almost) symmetric and may be well approximated by a normal distribution. ABDL (2001) confirm the latter results for realized variances and its transformations based on five minute intraday returns.

Turning to conditional dynamics of realized variances, volatilities and log volatilities strong persistence is diagnosed. For instance, regarding the autocorrelation function of log volatility significant autocorrelations are found for both FX rates up to lag 70 or more. Interestingly, the autocorrelation pattern becomes almost uninformative after fractionally differentiating log volatility. To cope with the apparent long memory feature stationary, fractionally integrated time series models (Granger, 1980) have been successfully employed to model and forecast the conditional mean of log realized volatility (ABDL, 2003). With respect to the comovement of the two major USD rates ABDL (2001) illustrate that realized covariances are skewed to the right while the unconditional distribution of realized correlations is almost symmetric and approximately normal. The autocorrelation function of the latter reveals long memory.

ABDE (2001) use five minute returns to determine daily variances, covariances and correlations of the 30 US stocks listed in the DOW Jones Industrial Average over a period of five years. Remarkably, it turned out that stylized features characterizing the FX markets' realized second order moments carry over to the stock market as well. Particular summary statistics for standardized returns or log volatilities turned out to be quite homogeneous over the cross section of 30 stocks. The estimated degree of fractional integration observed over the stock markets' log realized volatilities varied closely around $d = 0.35$ whereas corresponding estimates reported

in ABDL (2003) for the DEM/USD and JPY/USD markets are 0.39 and 0.41, respectively. Moreover, stock market correlations appear to be positively related with the level of realized volatilities such that the gains of portfolio diversification are reduced in periods of high volatility. ABDE (2001) also analyze correlation between lagged returns and realized variances. The so-called leverage effect (Black, 1976) describing that asset prices show higher variation in the sequel of bad news in comparison with good news of comparable size is also found for realized variances, but of minor quantitative importance.

7.3.5 Bias Correction

In principal, the accuracy of realized variance estimates increases with the frequency of intradaily returns. Sampling at very high frequencies, however, is likely to be affected by microstructure noise going back e. g. to bid-ask bounces (negative autocorrelation), splitting of large orders into a sequence of smaller orders (positive autocorrelation) or inventory adjustments. In this case the assumption of an underlying semi martingale approximating the (log) price process is clearly not appropriate. High frequency return autocorrelation will cause realized variance to underestimate (positive autocorrelation) or overestimate (negative autocorrelation) the underlying actual variance. Aït-Sahalia *et al.* (2005) discuss in detail microstructure effects on realized variance estimates. To guard against microstructure noise ABDE (2001) whiten high frequency returns by means of an MA(1) filter. Starting from the decomposition

$$\left(\sum_{m=1}^{M} r_{m,t}\right)^2 = \sum_{m=1}^{M} r_{m,t}^2 + 2\sum_{m=1}^{M-1}\sum_{n=m+1}^{M} r_{m,t}r_{n,t} \tag{7.23}$$

Oomen (2003) illustrates that in presence of market microstructure noise the second term in (7.23), a sum of realized autocovariances, is likely to be substantial when M is large but will disappear when the sampling frequency decreases. A suitable intraday sampling frequency making realized variance an unbiased estimator of the actual variance is the highest possible frequency with approximately zero autocovariance component. When sampling at lower frequencies, however, unbiased estimation goes at the cost of efficiency, which in principle is increasing with the sampling frequency. Therefore, Zhang *et al.* (2003) and Aït-Sahalia *et al.* (2005) advocate subsampling schemes allowing consistent volatility estimation even if high frequency returns are serially correlated. Hansen and Lunde (2004) recommend to sample at the highest feasible frequency and propose an adjustment of the realized variance that depends on autocovariances up to some lag q.

References

Admati, A. R., Pfleiderer, P. (1988). A theory of intraday patterns: Volume and price variability. *Review of Financial Studies* **1** 3–40.

Aït-Sahalia, Y., Mykland, P. A., Zhang, L. (2005). Ultra high frequency volatility estimation with dependent microstructure noise. National Bureau of Economic Research, Paper in Asset Pricing, Working Paper No. 11380.

ANDERSEN, T. G., BOLLERSLEV, T., DIEBOLD, F. X. (2005). Parametric and nonparametric volatility measurement. In *Handbook of Financial Econometrics* (L. P. Hansen, Y. Aït-Sahalia, eds.), forthcoming. North Holland, Amsterdam.

ANDERSEN, T. G., BOLLERSLEV, T., DIEBOLD, F. X., EBENS, H. (2001). The distribution of realized stock return volatility. *Journal of Financial Economics* **61** 43–76.

ANDERSEN, T. G., BOLLERSLEV, T., DIEBOLD, F. X., LABYS, P. (2001). The distribution of realized exchange rate volatility. *Journal of the American Statistical Association* **96** 42–55.

ANDERSEN, T. G., BOLLERSLEV, T., DIEBOLD, F. X., LABYS, P. (2003). Modeling and forecasting realized volatility. *Econometrica* **71** 579–625.

ANDERSEN, T. G., BOLLERSLEV, T., DIEBOLD, F. X., WU (2004). Realized beta: Persistence and predictability. Northwestern University, Duke University and University of Pennsylvania, Manuscript.

BACK, K. (1991). Asset pricing for general processes. *Journal of Mathematical Economics* **20** 371–395.

BARNDORFF-NIELSEN, O. E., SHEPHARD, N. (2002a). Econometric analysis of realized volatility and its use in estimation stochastic volatility models. *Journal of the Royal Statistical Society, Series B* **64** 253–280.

BARNDORFF-NIELSEN, O. E., SHEPHARD, N. (2002b). Estimating quadratic variation using realized variance. *Journal of Applied Econometrics* **17** 457–477.

BARNDORFF-NIELSEN, O. E., SHEPHARD, N. (2004). Econometric analysis of realized covariation: High frequency based covariance, regression and correlation in financial economics. *Econometrica* **72** 885–925.

BARNDORFF-NIELSEN, O. E., SHEPHARD, N. (2005). How accurate is the asymptotic approximation to the distribution of realized volatility?. In *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg* (D. W. F. Andrews, J. H. Stock, eds.), Cambridge University Press, Cambridge.

BLACK, F. (1976). Studies of stock market volatility changes. *Proceedings of the American Statistical Association, Business and Economic Statistics Section* 177–181.

BOLLERSLEV, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* **31** 307–327.

BOLLERSLEV, T. (1987). A conditional heteroscedastic time series model for speculative prices and rates of return. *Review of Economics and Statistics* **69** 542-547.

DACOROGNA, M. M., MÜLLER, U. A., NAGLER, R. J., OLSEN, R. B., PICTET, O. V. (1993). A geographical model for the daily and weekly seasonal volatility in the foreign exchange market. *Journal of International Money and Finance* **12** 413–438.

DE JONG, F., NIJMAN, T. (1997). High-frequency analysis of lead-lag relationships between financial markets. *Journal of Empirical Finance* **4** 187-212.

DE JONG, F., MAHIEU, R., SCHOTMAN, P. (1998). Price discovery in the foreign exchange market: An empirical analysis of the Yen/Dmark rate. *Journal of International Money and Finance* **17** 5-27.

EASLEY, D., O'HARA, M. (1992). Time and the process of security price adjustment. *Journal of Finance* **19** 69–90.

ENGLE, R. F. (1982). Autoregressive conditional heteroskedasticity with estimates of the variance of U.K. inflation. *Econometrica* **50** 987–1008.

ENGLE, R. F., GONZALEZ-RIVERA, G. (1991). Semiparametric ARCH models. *Journal of Business an Economic Statistics* **9** 435–459.

ENGLE, R. F., GRANGER, C. W. J. (1987). Co–integration and error correction: Representation, estimation and testing. *Econometrica* **55** 251–276.

ENGLE, R. F., RUSSELL, J. R. (1998). Autoregressive conditional duration: A new model for irregularly spaced data. *Econometrica* **66** 1127–1162.

ENGLE, R. F., RUSSELL, J. R. (2005). Analysis of high frequency financial data. In *Handbook of Financial Econometrics* (L. P. Hansen, Y. Aït-Sahalia, eds.), forthcoming. North Holland, Amsterdam.

FRIJNS, B., SCHOTMAN, P. (2003). Price discovery in tick time. Limburg Institute of Financial Economics LIFE, Working Paper 03-024.

GHYSELS, E., GOURIÉROUX, C., JASIAK, J. (1997). Market time and asset price movements: Theory and estimation. In *Statistics in Finance* (D. Hand, S. Jacka, eds.), 307–332. Edward Arnold, London.

GOODHART, C. A. E., O'HARA, M. (1997). High frequency data in financial markets: Issues and applications. *Journal of Empirical Finance* **4** 73–114.

GOODHART, C. A. E., HALL, S. G., HENRY, S. G. B., PESARAN, B. (1993). News effects in a high frequency model of the Sterling-Dollar exchange rate. *Journal of Applied Econometrics* **8** 1-13.

GRAMMIG, J., MELVIN, M., SCHLAG, C. (2005). Internationally cross-listed stock prices during overlapping trading hours: Price discovery and exchange rate effects. *Journal of Empirical Finance* **12** 139–164.

GRANGER, C. W. J. (1980). Testing for causality: A personal viewpoint. *Journal of Economic Dynamics and Control* **2** 329-352.

HANSEN, P. R., LUNDE, A. (2004). An unbiased measure of realized variance. Brown University, Working Paper.

HARRIS, F. H. DEB., SHOESMITH, G. L., MCINISH, T. H., WOOD, R. A. (1995). Cointegration, error correction, and price discovery on informationally linked security markets. *Journal of Financial and Quantitative Analysis* **30** 563-579.

HASBROUCK, J. (1995). One security, many markets, determining the contributions to price discovery. *Journal of Finance* **50** 1175–1199.

HERWARTZ, H. (2001). Investigating the JPY/DEM rate: Arbitrage opportunities and a case for asymmetry. *International Journal of Forecasting* **17** 231-245.

HUANG, R. D. (2002). The quality of ECN and market maker quotes. *Journal of Finance* **57** 1285–1319.

JOHANSEN, S. (1995). *Likelihood-based inference in cointegrated vector autoregressive models.* Oxford University Press, Oxford.

JONES, R. H. (1980). Maximum likelihood fitting of ARMA models to time series with missing observations. *Technometrics* **22** 389–395.

JORDÀ, Ò., MARCELLINO, M. (2003). Modeling high–frequency foreign exchange data dynamics. *Macroeconomic Dynamics* **7** 618–635.

KARPOFF, J. M. (1987). The relation between price changes and trading volume: A survey. *Journal of Financial and Quantitative Analysis* **22** 109–26.

KOHN, R., ANSLEY, C. F. (1986). Estimation, prediction, and interpolation for ARIMA models with missing data. *Journal of the American Statistical Association* **81** 751–761.

LO, M. C., ZIVOT, E. (2001). Threshold cointegration and nonlinear adjustment to the law of one price. *Macroeconomic Dynamics* **5** 533–576.

LÜTKEPOHL, H. (1991). *Introduction to Multiple Time Series Analysis.* Springer, Berlin.

NELSON, D. B. (1990). ARCH models as diffusion approximations. *Journal of Econometrics* **45** 7–39.

MERTON, R. (1973). Theory of rational option pricing. *Bell Journal of Economics and Management Science* **4** 141-183.

OOMEN, R. C. (2003). *Three Essays on the Econometric Analysis of High Frequency Financial Data.* European University Institute, Ph. D. thesis.

PROTTER, P. (1990). *Stochastic Integration and Differential Equations: A New Approach.* Springer, New York.

SCHREIBER, P. S., SCHWARTZ, R. A. (1986). Price discovery in securities markets. *Journal of Portfolio Management* **12** 43-48.

SCHWERT, G. W. (1989). Why does stock market volatility change over time. *Journal of Finance* **44** 1115–1153.

SCHWERT, G. W. (1990). Stock volatility and the crash of '87. *Journal of Financial Studies* **3** 77–102.

TAYLOR, S. J. (1986). *Modeling financial time series.* John Wiley, Chichester.

TERÄSVIRTA, T. (1994). Specification, estimation, and evaluation of smooth transition autoregressive models. *Journal of the American Statistical Association* **89** 208–218.

WHITE, H. (1984). *Asymptotic Theory for Econometricians*. Academic Press, Orlando.

ZHANG, L., MYKLAND, P. A., AÏT-SAHALIA, Y. (2005). A tale of two time scales: Determining integrated volatility with noisy high-frequency data. *Journal of the American Statistical Association* (forthcoming).

8 Using Quantile Regression for Duration Analysis *

Bernd Fitzenberger[1] and Ralf A. Wilke[2]

[1] Department of Economics, J.W. Goethe University Frankfurt
fitzenberger@wiwi.uni-frankfurt.de
[2] Centre for European Economic Research (ZEW)
wilke@zew.de

Summary: Quantile regression methods are emerging as a popular technique in econometrics and biometrics for exploring the distribution of duration data. This paper discusses quantile regression for duration analysis allowing for a flexible specification of the functional relationship and of the error distribution. Censored quantile regression addresses the issue of right censoring of the response variable which is common in duration analysis. We compare quantile regression to standard duration models. Quantile regression does not impose a proportional effect of the covariates on the hazard over the duration time. However, the method cannot take account of time–varying covariates and it has not been extended so far to allow for unobserved heterogeneity and competing risks. We also discuss how hazard rates can be estimated using quantile regression methods.

8.1 Introduction

Duration data are commonly used in applied econometrics and biometrics. There is a variety of readily available estimators for popular models such as the accelerated failure time model and the proportional hazard model, see e. g. Kiefer (1988) and van den Berg (2001) for surveys. Quantile regression is recently emerging as an attractive alternative to these popular models (Koenker and Bilias, 2001; Koenker and Geling, 2001; Portnoy; 2003). By modelling the distribution of the duration in a

*This paper benefitted from the helpful comments by an anonymous referee. Due to space constraints, we had to omit the details of the empirical application. These can be found in the long version of this paper, Fitzenberger and Wilke (2005). We gratefully acknowledge financial support by the German Research Foundation (DFG) through the research project 'Microeconometric modelling of unemployment durations under consideration of the macroeconomic situation'. Thanks are due to Xuan Zhang for excellent research assistance. All errors are our sole responsibility.

flexible semiparametric way, quantile regression does not impose modelling assumptions that may not be empirically valid, e. g. the proportional hazard assumption. Quantile regression models are more flexible than accelerated failure time models or the Cox proportional hazard model because they do not restrict the variation of estimated coefficients over the quantiles. Estimating censored quantile regression allows to take account of right censoring which is present in typical applications of duration analysis (Powell, 1984; Fitzenberger, 1997). However, quantile regression involves three major disadvantages. First, the method is by definition restricted to the case of time–invariant covariates. Second, there is no competing risks framework yet and third, so far quantile regression does not account for unobserved heterogeneity, which is a major ingredient of the mixed proportional hazard rate model.

Quantile regression models the changes of quantiles of the conditional distribution of the duration in response to changes of the covariates. In actual applications of duration analysis, researchers are often interested in the effects on the hazard rate after a certain elapsed duration and how the hazard rate changes with the elapsed duration (duration dependence). Machado and Portugal (2002) and Guimarães *et al.* (2004) have introduced a simple simulation method to obtain the conditional hazard rates implied by the quantile regression estimates. In this paper, we present a slightly modified version of their estimator. The modifications are necessary to overcome difficulties in the case of censored data and to fix a general smoothing problem. Using this method, it is straightforward to analyze duration dependence without having to assume that the pattern estimated for the so-called baseline hazard in proportional hazard rate models applies uniformly to all observations with different covariates.

Section 2 discusses important aspects of quantile regression methods for duration analysis and shows how conditional hazard rates can be obtained from estimated quantile regression coefficients. Section 3 summarizes.

8.2 Quantile Regression and Duration Analysis

This section discusses quantile regression as an econometric tool to estimate duration models and addresses various issues involved. Quantile regression models are contrasted with the popular proportional hazard rate model. Our discussion includes selected results from an empirical application taken from the long version of this paper, Fitzenberger and Wilke (2005).

8.2.1 Quantile Regression and Proportional Hazard Rate Model

Koenker und Bassett (1978) introduced quantile regressions[1] as a regression based method to model the quantiles of the response variable conditional on the covariates. Our focus is on linear quantile regression for duration data involving the estimation of the accelerated failure time model $h(T_i) = x_i'\beta^\theta + \epsilon_i^\theta$ at different quantiles $\theta \in (0,1)$ for the completed duration T_i of spell i, where the θ-quantile of ϵ_i^θ conditional on x_i, $q_\theta(\epsilon_i^\theta|x_i)$, is zero and $h(.)$ is a strictly monotone transformation preserving the ordering of the quantiles. The most popular choice is the log–transformation $h(.) = log(.)$. The transformation can either be chosen a priori (e. g. as being the log–transformation) or it can be estimated by choosing among a class of transformation functions (e. g. among the set of possible Box–Cox–transformations, see e. g. Buchinsky, 1995, or Machado and Mata, 2000) due to the invariance of quantiles under positive monotone transformations. Quantile regression models are not restricted to a linear specification of the conditional quantiles.[2] In fact, quantile regression models the conditional quantile of the response variable $q_\theta(h(T_i)|x_i) = x_i'\beta^\theta$ or, alternatively, due to the equivariance property of quantiles $q_\theta(T_i|x_i) = h^{-1}(x_i'\beta^\theta)$. Modelling conditional quantiles is an indirect way to model the conditional distribution function of $log(T_i)$ given x_i. The linear specification allows for differences in the impact of covariates x_i across the conditional distribution of the response variable. However, the specification imposes that the coefficient is the same for a given quantile θ irrespective of the actual value of $q_\theta(h(T_i)|x_i)$.

We will discuss the asymptotic distribution for linear quantile regression in the next subsection for the case of censored quantile regression which nests the case without censoring. The asymptotic distribution in the case of a smooth transformation function $h(.)$ depending on unknown parameters to be estimated can be found in Powell (1991), Chamberlain (1994), or Fitzenberger *et al.* (2004), who treat the special case of Box–Cox transformation. The asymptotic results generalize in a straight forward manner to other smooth transformation functions.

A possible problem of quantile regression is the possibility that the coefficient estimates can be quite noisy (even more so for censored quantile regression) and often non-monotonic across quantiles. To mitigate this problem, it is possible to obtain smoothed estimates through a minimum–distance approach. One can investigate, whether a parsimonious relation describes the movement of the coefficients across quantiles by minimizing the quadratic form $(\hat{\beta} - f[\delta])'\hat{\Psi}^{-1}(\hat{\beta} - f[\delta])$ with respect to δ, the coefficients of a smooth parametric specification of the coefficients as a funtion of θ. $\hat{\beta}$ is the stacked vector of quantile regression coefficient estimates $\hat{\beta}^\theta$ at

[1] See Buchinsky (1998) and Koenker and Hallock (2002) for surveys. The collection of papers in Fitzenberger *et al.* (2001) comprises a number of economic applications of quantile regressions, among others, the paper by Koenker and Bilias (2001) using quantile regression for duration analysis.

[2] Since the transformation $h(.)$ is monontone, estimation of a Box–Cox transformation is an attractive alternative. Chamberlain (1994) and Buchinsky (1995) suggest a simple two step procedure to implement Box–Cox quantile regressions. However, this procedure can exhibit numerical problems which are analyzed in Fitzenberger *et al.* (2004). The latter study suggests a modified estimator for Box–Cox quantile regressions.

different quantiles and $\hat{\Psi}$ is the estimated covariance matrix of $\hat{\beta}$, see next subsection for the asymptotic distribution. This approach is not pursued in the application below. We are not aware of any application of this approach in the literature.

The most popular parametric Cox proportional hazard model (PHM), Kiefer (1988), is based on the concept of the hazard rate conditional upon the covariate vector x_i given by

$$\lambda_i(t) = \frac{f_i(t)}{P(T_i \geq t)} = exp(x_i'\tilde{\beta})\lambda_0(t)\,, \tag{8.1}$$

where $f_i(t)$ is the density of T_i at duration t and $\lambda_0(t)$ is the so called baseline hazard rate. The hazard rate is the continuous time version of an instantaneous transition rate, i. e. it represents approximately the conditional probability that the spell i ends during the next period of time after t conditional upon survival up to period t.

There is a one-to-one correspondence between the hazard rate and the survival function, $S_i(t) = P(T_i \geq t)$, of the duration random variable, $S_i(t) = exp\left(-\int_0^t \lambda_i(s)ds\right)$. A prominent example of the parametric[3] proportional hazard model is the Weibull model where $\lambda_0(t) = pt^{p-1}$ with a shape parameter $p > 0$ and the normalizing constant is included in $\tilde{\beta}$. The case $p = 1$ is the special case of an exponential model with a constant hazard rate differentiating the increasing ($p > 1$) and the decreasing ($0 < p < 1$) case. Within the Weibull family, the proportional hazard specification can be reformulated as the accelerated failure time model

$$log(T_i) = x_i'\beta + \epsilon_i\,, \tag{8.2}$$

where $\beta = -p^{-1}\tilde{\beta}$ and the error term ϵ_i follows an extreme value distribution, Kiefer (1988, Sections IV and V).

The main thrust of the above result generalizes to any PHM (8.1). Define the integrated baseline hazard $\Lambda_0(t) = \int_0^t \lambda_0(\tilde{t})d\tilde{t}$, then the following well known generalization of the accelerated failure time model holds

$$log(\Lambda_0(T_i)) = x_i'\beta + \epsilon_i\,, \tag{8.3}$$

with ϵ_i again following an extreme value distribution and $\beta = -\tilde{\beta}$, see Koenker and Bilias (2001) for a discussion contrasting this result to quantile regression. Thus, the proportional hazard rate model (8.1) implies a linear regression model for the a priori unknown transformation $h(T_i) = log(\Lambda_0(T_i))$. This regression model involves an error term with an a priori known distribution of the error term and a constant coefficient vector across quantiles.

>From a quantile regression perspective, it is clear that these properties of the PHM are quite restrictive. Provided the correct transformation is applied, it is possible to investigate whether these restrictions hold by testing for the constancy of the estimated coefficients across quantiles. Testing whether the error term follows an extreme value distribution is conceivable though one has to take account of possible

[3]Cf. Kiefer (1988, Section III.A) for nonparametric estimation of the baseline hazard $\lambda_0(t)$.

shifts and normalizations implied by the transformation. However, if a researcher does not apply the correct transformation in (8.3), e. g. the log transformation in (8.2) is used though the baseline hazard is not Weibull, then the implications are weaker. Koenker and Geling (2001, p. 462) show that the quantile regression coefficients must have the same sign if the the data is generated by a PHM.

A strong and quite apparent violation of the proportional hazard assumption occurs, if for two different covariate vectors x_i and x_j, the survival functions $S_i(t)$ and $S_j(t)$, or equivalently the predicted conditional quantiles, do cross. Crossing occurs, if for two quantiles $\theta_1 < \theta_2$ and two values of the covariate vector x_i and x_j, the ranking of the conditional quantiles changes, e. g. if $q_{\theta_1}(T_i|x_i) < q_{\theta_1}(T_j|x_j)$ and $q_{\theta_2}(T_i|x_i) > q_{\theta_2}(T_j|x_j)$. Our empirical application below involves cases with such a finding. This is a valid rejection of the PHM, provided our estimated quantile regression provides a sufficient goodness-of-fit for the conditional quantiles.

There are three major advantages of PHMs compared to quantile regressions as discussed in the literature. PHMs can account for unobserved heterogeneity, for time varying covariates, and for competing risks in a straight forward way (Wooldridge, 2002, Chapter 20). The issue of unobserved heterogeneity will be discussed at some length below. The estimation of competing risks models with quantile regression has not been addressed in the literature. This involves a possible sample selection bias, an issue which has only be analyzed under much simpler circumstances for quantile regression (Buchinsky, 2001). In fact, this is a dynamic selection problem which, also in the case of a PHM, requires fairly strong identifying assumptions.

It is natural to consider time varying covariates when the focus of the analysis is the hazard rate as a proxy for the exit rate during a short time period. This is specified in a PHM as

$$\lambda_i(t) = exp(x'_{i,t}\tilde{\beta})\lambda_0(t) \tag{8.4}$$

and there are readily available estimators for this case. It is not possible anymore to transform this model directly into an accelerated failure time model which could be estimated by regression methods.

Assuming strict exogeneity of the covariates, it is straightforward to estimate proportional hazard models with time varying coefficients (Wooldridge, 2002, Chapter 20). If under strict exogeneity the complete time path of the covariates is known, it is conceivable – though often not practical – to condition the quantile regression on the entire time path to mimick the time varying effect of the covariates. A natural example in the analysis of unemployment durations would be that eligibility for unemployment benefits is exhausted after a certain time period and this is known ex ante. In fact, in such a case quantile regression also naturally allow for anticipation effects which violates specification (8.4). In many cases, the time path of time–varying covariates is only defined during the duration of the spell, which is referred to as internal covariates (Wooldridge, 2002, p. 693). Internal covariates typically violate the strict exogeneity assumption and it is difficult to relax the strict exogeneity assumption when also accounting for unobserved heterogeneity.

The case of time varying coefficients β_t can be interpreted as a special case of time–varying covariates by interacting the covariates with dummy variables for different time periods. However, if the specification of the baseline hazard function

is very flexible then an identification issue can arise. Time varying coefficients β_t are similar in spirit to quantile regressions with changing coefficients across conditional quantiles. While the former involves coefficients changes according to the actual elapsed duration, the latter specifies these changes as a function of the quantile. It depends on the application as to which approach can be better justified.

Summing up the comparison so far, while there are some problems when using the PHM with both unobserved heterogeneity and time-varying covariates, the PHM can take account of these issues in a somewhat better way than quantile regression. Presently, there is also a clear advantage of the PHM regarding the estimation of competing risk models. However, the estimation of a PHM comes at the cost of the proportional hazard assumption which itself might not be justifiable in the context of the application of interest.

8.2.2 Censoring and Censored Quantile Regression

Linear censored quantile regression, introduced by Powell (1984, 1986), allow for semiparametric estimation of quantile regression for a censored regression model in a robust way. A survey on the method can be found in Fitzenberger (1997). Since only fairly weak assumptions on the error terms are required, censored quantile regression (CQR) is robust against misspecification of the error term. Horowitz and Neumann (1987) were the first to use CQR's as a semiparametric method for an accelerated failure time model of employment duration.

Duration data are often censored. Right censoring occurs when we only observe that a spell has survived until a certain duration (e. g. when the period of observation ends) but we do not know exactly when it ends. Left censoring occurs when spells observed in the data did start before the beginning of the period of observation. Spells who started at the same time and who finished before the beginning of the period of observation are not observed. Quantile regression can not be used with left censored data.[4] Left censoring is also difficult to handle for PHMs since strong assumptions have to be invoked to estimate the model. In the following, we only consider the case of right censoring which both PHM and CQR are well suited for. Thus, we can only analyze so–called flow samples (Wooldridge, 2002, Chapter 20) of spells for which the start of the spells lies in the time period of observation.[5]

Let the observed duration be possibly right censored in the flow sample, i. e. the observed completed duration T_i is given by $T_i = min\{T_i^*, yc_i\}$, where T_i^* is the true duration of the spell and yc_i is the spell specific threshold value (censoring point) beyond which the spell cannot be observed. For the PHM, this can be incorporated in maximum likelihood estimation analogous to a censored regression model (Wooldridge, 2002, Chapter 20) and it is not necessary to know the potential censoring points yc_i for uncensored observations. In contrast, CQR requires the knowledge of yc_i irrespective of whether the observation is right censored. CQR

[4] Two–limit censored quantile regression (Fitzenberger, 1997) can be used in the rare situation when all spells are observed which start before the start of the observation period and, in case they end before the start of the observation period, the exact length of the spell is not known.

[5] Analyzing all spells observed at some point of time during the period of observations involves a so–called stock sample also including left–censored observations.

provide consistent estimates of the quantile regression coefficients β^θ in the presence of fairly general forms of fixed censoring.[6] The known censoring points can either be deterministic or stochastic and they should not bunch in a certain way on or around the true quantile regression line, see the discussion in Powell (1984).

Estimating linear CQR involves minimizing the following distance function

$$\sum_{i=1} \rho_\theta(ln(T_i) - min(x_i'\beta^\theta, yc_i)) \tag{8.5}$$

with respect to β^θ, where the so-called 'check function' $\rho_\theta(z) = \theta \cdot |z|$ for $z \geq 0$ and $\rho_\theta(z) = (1-\theta) \cdot |z|$ for $z < 0$ and yc_i denotes the known observation specific censoring points. A quantile regression without censoring is nested as the special case with $yc_i = +\infty$.

Powell (1984, 1986) showed that the CQR estimator $\hat{\beta}^\theta$ is $\sqrt{N}$–consistent and asymptotically normally distributed, see also Fitzenberger (1997) for a detailed discussion of the asymptotic distribution. A crucial feature of this result is that the asymptotic distribution depends only upon those observations where the fitted quantiles are not censored, i. e. $I(x_i'\hat{\beta}^\theta < yc_i) = 1$.

The actual calculation of the CQR-estimator based on individual data is numerically very difficult, since the distance function (8.5) to be minimized is not convex. This is in contrast to quantile regression without censoring. There are a number of procedures suggested in the literature to calculate the CQR-estimator (Buchinsky, 1998; Fitzenberger, 1997, and Fitzenberger and Winker, 2001).[7]

For heteroscedasticity–consistent inference, researchers often resort to bootstrapping, see e. g. Buchinsky (1998) and Fitzenberger (1997, 1998), using the Design–Matrix–Bootstrap (often also denoted as 'pairwise bootstrap'). The covariance of the CQR estimates across quantiles can easily be estimated by basing those estimates on the same resample. Bilias *et al.* (2000) suggest a simplified version of the bootstrap for CQR by showing that it suffices asymptotically to estimate a quantile regression without censoring in the resample based only on those observations for which the fitted quantile is not censored, i. e. $x_i'\hat{\beta}^\theta < yc_i$.

The long version of this paper (Fitzenberger and Wilke, 2005) contains a comprehensive application of censored quantile regression to the duration of unemployment among young West German workers. Figure 8.1 presents estimated quantile regression coefficients for two selected covariates. The confidence bands are obtained by the Bilias *et al.* (2000) bootstrap method. It is apparent that the estimated coefficients change their sign across quantiles and therefore they do not support empirically the proportional hazard model.

[6] Refer to Buchinsky and Hahn (1998) for a semiparametric extension of CQR to the case when the censoring points are not known for the uncensored observations (random censoring).

[7] In light of the numerical difficulties, a number of papers have, in fact, suggested to change the estimation problem to make it convex (Buchinsky and Hahn, 1998; Chernozhukov and Hong, 2002; and Portnoy, 2003).

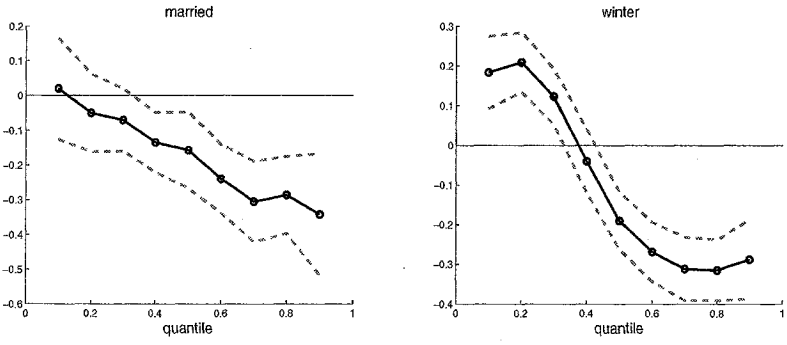


Figure 8.1: *Estimated quantile regression coefficients for the covariates dummy for being married (left) and unemployment spell starting in the winter (right) with* 90% *bootstrap confidence bands (see Fitzenberger and Wilke, 2005, for further explanations).*

8.2.3 Estimating the Hazard Rate Based on Quantile Regression

Applications of duration analysis often focus on the impact of covariates on the hazard rate. Quantile regression estimate the conditional distribution of T_i conditional on covariates and it is possible to infer the estimated conditional hazard rates from the quantile regression estimates.

A direct approach is to construct a density estimate from the fitted conditional quantiles $\hat{q}_\theta(T_i|x_i) = h^{-1}(x_i'\hat{\beta}^\theta)$. A simple estimate for the hazard rate as a linear approximation of the hazard rates between the different θ–quantiles would be

$$\hat{\lambda}_i(t) = \frac{(\theta_2 - \theta_1)}{\left(h^{-1}(x_i'\hat{\beta}^{\theta_2}) - h^{-1}(x_i'\hat{\beta}^{\theta_1})\right)(1 - 0.5(\theta_1 + \theta_2))}, \tag{8.6}$$

where $\hat{\lambda}_i(t)$ approximates the hazard rates between the estimated θ_1–quantile and θ_2–quantile.[8] Two points are noteworthy. First, the estimated conditional quantiles are piecewise constant due to the linear programming nature of quantile regression (Koenker and Bassett, 1978; Fitzenberger, 1997). Second, it is not guaranteed that the estimated conditional quantiles are ordered correctly, i. e. it can occur that $\hat{q}_{\theta_1}(T_i|x_i) > \hat{q}_{\theta_2}(T_i|x_i)$ even though $\theta_1 < \theta_2$. Therefore, θ_1 and θ_2 have to be chosen sufficiently far apart to guarantee an increase in the conditional quantiles.

In order to avoid these problems, Machado and Portugal (2002) and Guimarães *et al.* (2004) suggest a resampling procedure (henceforth denoted as GMP) to obtain the hazard rates implied by the estimated quantile regression. The main idea of

[8] A similar estimator based on the estimate of the sparsity function is described in Machado and Portugal (2002). It shares the same problems discussed for the estimator presented in (8.6).

GMP is to simulate data based on the estimated quantile regressions for the conditional distribution of T_i given the covariates and to estimate the density and the distribution function directly from the simulated data.

In detail, GMP works as follows (see Machado and Portugal, 2002, and Guimarães *et al.*, 2004), possibly only simulating non-extreme quantiles:

1. Generate M independent random draws $\theta_m, m = 1, ..., M$, from a uniform distribution on (θ_l, θ_u), i. e. extreme quantiles with $\theta < \theta_l$ or $\theta > \theta_u$ are not considered here. θ_l and θ_u are chosen in light of the type and the degree of censoring in the data. Additional concerns relate to the fact that quantile regression estimates at extreme quantiles are typically statistically less reliable, and that duration data might exhibit a mass point at zero or other extreme values. The benchmark case with the entire distribution is given by $\theta_l = 0$ and $\theta_u = 1$.
2. For each θ_m, estimate the quantile regression model obtaining M vectors β^{θ_m} (and the transformation $h(.)$ if part of the estimation approach).
3. For a given value of the covariates x_0, the sample of size M with the simulated durations is obtained as $T_m^* \equiv \hat{q}_{\theta_m}(T_i|x_0) = h^{-1}(x_0'\hat{\beta}^{\theta_m})$ with $m = 1, ..., M$.
4. Based on the sample $\{T_m^*, m = 1, ..., M\}$, estimate the conditional density $f^*(t|x_0)$ and the conditional distribution function $F^*(t|x_0)$.
5. We suggest to estimate the hazard rate conditional on x_0 in the interval (θ_l, θ_u) by[9]

$$\hat{\lambda}_0(t) = \frac{(\theta_u - \theta_l) f^*(t|x_0)}{1 - \theta_l - (\theta_u - \theta_l) F^*(t|x_0)}.$$

Simulating the full distribution ($\theta_l = 0$ and $\theta_u = 1$), one obtains the usual expression: $\hat{\lambda}_0(t) = f^*(t|x_0)/[1 - F^*(t|x_0)]$.

This procedure (Step 3) is based on the probability integral transformation theorem from elementary statistics implying $T_m^* = F^{-1}(\theta_m)$ is distributed according to the conditional distribution of T_i given x_0 if $F(.)$ is the conditional distribution function and θ_m is uniformly distributed on $(0, 1)$. Furthermore, the fact is used that the fitted quantile from the quantile regression provides a consistent estimate of the population quantile, provided the quantile regression is correctly specified.

The GMP procedure uses a kernel estimator for the conditional density $f^*(t|x_0) = 1/(M\,h) \sum_{m=1}^{M} K\left((t - T_i^*)/h\right)$, where h is the bandwidth and $K(.)$ the kernel function. Based on this density estimate, the distribution function estimator is

$$F^*(t|x_0) = 1/M \sum_{m=1}^{M} \mathcal{K}\left((t - T_i^*)/h\right) \quad \text{with} \quad \mathcal{K}(u) = \int_a^t K(v)\, dv.$$

[9] $f^*(t|x_0)$ estimates the conditional density in the quantile range (θ_l, θ_u), i. e. $f(t|q_{\theta_l}(T|x_0) < T < q_{\theta_u}(T|x_0), x_0)$, and the probability of the conditioning event is $\theta_u - \theta_l = P(q_{\theta_l}(T|x_0) < T < q_{\theta_u}(T|x_0)|x_0)$. By analogous reasoning, the expression in the denominator corresponds to the unconditional survival function, see Zhang (2004) for further details.

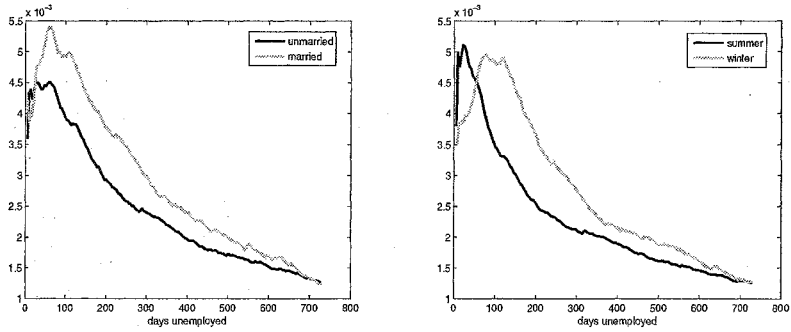


Figure 8.2: *Estimated conditional hazard rates evaluated at sample means of the other regressors. See Figure 8.1 for further explanations.*

Machado and Portugal (2002) and Guimarães *et al.* (2004) suggest to start the integration at zero ($a = 0$), probably because durations are strictly positive. However, the kernel density estimator also puts probability mass into the region of negative durations, which can be sizeable with a large bandwidth, see Silverman (1986, Section 2.10). Using the above procedure directly, it seems more advisable to integrate starting from minus infinity, $a = -\infty$. A better and simple alternative would be to use a kernel density estimator based on log durations. This is possible when observed durations are always positive, there is no mass point at zero. Then, the estimates for density and distribution function for the duration itself can easily be derived from the density estimates for log duration by applying an appropriate transformation.[10]

Figure 8.2 presents estimates of four conditional hazard rates using again the sample of unemployment durations among the young workers in West-Germany. It is apparent that a flexible econometric method can reveal interesting results that would not show up under stronger conditions. Since the estimated hazard rates even cross, a proportional hazard model is inappropriate in this case.

8.2.4 Unobserved Heterogeneity

In duration analysis, unobserved heterogeneity in the form of spell specific, time-invariant location shifts of the hazard rate or the duration distribution play a key role (Wooldridge, 2002, Chapter 20.3.4; van den Berg, 2001). The popular mixed proportional hazard model (MPHM) assumes that the spell specific effect α enters the specification of the hazard rate in a multiplicative fashion: $\lambda_i(t) = exp(x_i'\tilde{\beta})\lambda_0(t)exp(\alpha)$. Under the assumptions of the MPHM, α is a random effect which is distributed independently from the vector of covariates. It is well known that ignoring the presence of the random effect α will lead to misleading evidence on the shape of the baseline hazard $\lambda_0(t)$ inducing spurious duration dependence due to the sorting of spells with respect to α. Spells with a low value of α tend to

[10] Silverman (1986, Section 2.10) discusses further alternatives for this problem.

survive relatively long and, thus, one might conclude that the hazard rate declines with elapsed duration when ignoring the influence of α. In general, ignoring the random effect α also biases the estimated coefficients for the covariates (Lancaster, 1990, p. 65), though the impact is typically small. In the accelerated failure time model (Equations 8.2 and 8.3), the random effect results in another component of the error term which is independent of the covariates. Therefore, with known integrated baseline hazard ($\Lambda_0(.)$), quantile regression (or even OLS in the absence of censoring) can estimate consistently the coefficient estimates.

Quantile regression estimates the conditional quantile $q_\theta(T_i|x_i)$. Clearly, the increase in the conditional quantiles $q_\theta(T_i|x_i)$ for given x_i with increasing θ corresponds to the shape of the hazard rate as a function of elapsed duration. Thus, the increase in the conditional quantile is affected by the presence of unobserved heterogeneity. If the data generating process is an MPHM, then $\partial q_\theta(T_i|x_i)/\partial\theta$ differs from the average increase $E_\alpha\left\{\partial q_{\tilde{\theta}(\alpha)}(T_i|x_i,\alpha)/\partial\theta\right\}$ evaluated at the same duration level $q_\theta(T_i|x_i)$ corresponding to the quantile position $\tilde{\theta}(\alpha)$ for each α. This is due to the well known sorting effects in α ('low α' types tend to survive longer) and, therefore, the former term $\partial q_\theta(T_i|x_i)/\partial\theta$ is typically larger than the latter averaged version across α for small durations and smaller for larger durations. However, for an MPHM, the presence of a random effect typically causes only a small bias on the point estimates of the estimated quantile regression coefficients of the covariates because of the following argument.[11] Applying the implicit function to $S_i(q_\theta(T_i|C_i)|C_i) = 1-\theta$, both for $C_i = (x_i,\alpha)$ and $C_i = x_i$, results in

$$\frac{\partial q_\theta(T_i|C_i)}{\partial x_i} = -\left\{\frac{\partial S_i(t|C_i)}{\partial t}|_{t=q_\theta(T_i|C_i)}\right\}^{-1}\frac{\partial S_i(t|C_i)}{\partial x_i}|_{t=q_\theta(T_i|C_i)}\,. \tag{8.7}$$

Since

$$S(t|x_i,\alpha) = exp\{-exp(x_i'\tilde{\beta})exp(\alpha)\Lambda_0(t)\}E_\alpha\{S(t|x_i,\alpha)\}\,,$$

it follows that

$$\frac{\partial q_\theta(T_i|C_i)}{\partial x_i} = \frac{-\tilde{\beta}\Lambda_0(t)}{\lambda_0(t)} \tag{8.8}$$

for $t = q_\theta(T_i|C_i)$ and $C_i = (x_i,\alpha)$ or $C_i = x_i$. This argument applies in an analogous way using a smooth monotonic transformation of the response variable. Thus, the estimated quantile regression coefficients only depend upon the coefficients $\tilde{\beta}$ and the shape of the baseline hazard. The quantile regression coefficients conditional upon α are the same for the same elapsed duration t irrespective of its rank θ, i. e. the estimated quantile regression coefficients for some unconditional quantile of the elapsed duration reflect the sensitivity of the respective quantile lying at this elapsed duration conditional upon the random effect. Put differently, some fixed duration $\bar{t}$ in general corresponds to two different ranks θ or $\theta(\alpha)$, respectively, when conditioning on $C_i = (x_i,\alpha)$ or $C_i = x_i$, $\bar{t} = q_\theta(T_i|x_i) = q_{\theta(\alpha)}(T_i|x_i,\alpha)$. Then, Equation

[11] See Zhang (2004) for detailed Monte Carlo evidence.

(8.8) implies that the partial effects for the different corresponding quantile regressions at this duration $\bar{t}$ are the same, i. e. $\partial q_\theta(T_i|x_i)/\partial x_i = \partial q_{\theta(\alpha)}(T_i|x_i,\alpha)/\partial x_i$. In this sense, a quantile regression on x_i provides meaningful estimates of partial effects, although the data are generated by an MPHM.

Evidence based on quantile regression can also be informative about the validity of the MPHM. Analogous to the PHM, a finding of crossings of conditional quantiles constitutes a rejection of the MPHM. If $x_i'\tilde{\beta} < x_j'\tilde{\beta}$ for a pair (x_i, x_j), then the hazard is higher for x_j than for x_i for all α and therefore $S(t|x_i,\alpha) > S(t|x_j,\alpha)$, see line before Equation (8.8). Integrating out the distribution of α, one obtains the inequality $S(t|x_i) > S(t|x_j)$ for all t. Thus, $q_\theta(T_i|x_i) > q_\theta(T_j|x_j)$ for all θ and therefore the MPHM implies that there should not be any crossings when just conditioning on the observed covariates. Intuitively, the independence between α and x_i implies that a change in covariates changes the hazard rate in the same direction for all α's. Therefore all quantiles conditional on (x_i, α) move into the opposite direction. The latter implies that the quantile conditional on only x_i must also move into that direction.

Instead of assuming that the random effect shifts the hazard rate by a constant factor as in the MPHM, a quantile regression with random effects for log durations could be specified as the following extension of the accelerated failure time model in equation (8.2)[12]

$$log(T_i) = x_i'\beta^\theta + \alpha + \epsilon_i^\theta\,, \tag{8.9}$$

where the random effect α enters at all quantiles. The entire distribution of log durations is shifted horizontally by a constant α, i. e. $log(T_i) - \alpha$ exhibits the same distribution conditional on x_i. α is assumed independent of x_i and ϵ_i^θ. The latter is defined as $\epsilon_i^\theta = log(T_i) - q_\theta(log(T_i)|x_i,\alpha)$.[13] The regression coefficients β^θ now represent the partial effect of x_i also conditioning upon the random effect α. Such a quantile regression model with random effects has so far not been considered in the literature. It most likely requires strong identifying assumptions when applied to single spell data. Here we use the model in (8.9) purely as point of reference.

How are the estimated quantile regression coefficients affected by the presence of α, when just conditioning on observed covariates x_i? Using $S(log(t)|x_i) = E_\alpha\{S(log(t)|x_i,\alpha)\}$ and result (8.7), it follows that[14]

$$\frac{\partial q_\theta(log(T_i)|x_i)}{\partial x_i} = \int \frac{f(\bar{t}\,|x_i)|x_i,\alpha)}{f(\bar{t}\,|x_i)|x_i)}\,\beta^{\tilde{\theta}(\alpha)}\,dG(\alpha) \tag{8.10}$$

for $\bar{t} = q_\theta(log(T_i)|x_i)$, where $f(.)$ and $F(.)$ are the pdf and the cumulative of the duration distribution, respectively, $G(.)$ is the distribution of α, and $\tilde{\theta}(\alpha) = F(q_\theta(log(T_i)|x_i)|x_i,\alpha)$. All expressions are evaluated at the duration $\bar{t}$ corresponding to the θ-quantile of the duration distribution conditioning only upon x_i. Hence,

[12] The following line of arguments applies analogously to the case with a general transformation $h(.)$.

[13] If ϵ_i^θ is independent of (x_i,α), then all coefficients, except for the intercept, can be estimated consistently by a quantile regression on just x_i. Also in this case, all slope coefficients are constant across quantiles.

[14] After submitting this paper, we found out that the result (8.10) is basically a special case of Theorem 2.1 in Hoderlein and Mammen (2005).

$f(q_\theta(log(T_i)|x_i)|x_i, \alpha)$ is the pdf conditional on both x_i and α, $f(q_\theta(log(T_i)|x_i)|x_i)$ the pdf just conditional on x_i both evaluated at the value of the quantile conditional on x_i. For the derivation of (8.10), note that

$$f(q_\theta(log(T_i)|x_i)|x_i) = \int f(q_\theta(log(T_i)|x_i)|x_i, \alpha)\, dG(\alpha).$$

For the value $\bar{t} = q_\theta(log(T_i)|x_i)$, $\tilde{\theta}(\alpha)$ is the corresponding quantile position (rank) in the distribution conditioning both upon x_i and α. According to Equation (8.10), the quantile regression coefficients conditioning only on x_i estimate in fact a weighted average of the β^θ in Equation (8.9) where the weight is given by the density ratio for the duration $q_\theta(T_i|x_i)$ conditioning on both x_i and α and only on x_i, respectively. Since these weights integrate up to unity, the quantile regression estimate conditioning on x_i correspond to a weighted average of the true underlying coefficients in Equation (8.9).

One can draw a number of interesting conclusions from the above result. First, if β^θ does not change with θ, then the estimated coefficients are valid estimators for the coefficients in Equation (8.9). Second, if β^θ only change monotonically with θ, then the estimated coefficients will move in the same direction, in fact, understating the changes in β^θ. In this case, the random effect results in an attenuation bias regarding the quantile specific differences. Third, if one finds significant variation of the coefficients across quantiles, then this implies that the underlying coefficients in (8.9) exhibit an even stronger variation across quantiles. If the variation in the estimates follows a clear, smooth pattern, then it is most likely that the underlying coefficients in (8.9) exhibit the same pattern in an even stronger way.

Though being very popular in duration analysis, the assumption that the random effect and the covariates are independent, is not credible in many circumstances, for the same reasons as in linear panel data models (Wooldridge, 2002, Chapters 10 and 11). However, fixed effects estimation does not appear feasible with single spell data. Identification is an issue here.

Summing up, though as an estimation method quantile regression with random effects has not yet been developed, it is clear that quantile regression conditioning just on the observed covariates yields meaningful results even in the random effects case. Changing coefficients across quantiles implies that such differences are also present in the underlying model conditional upon the random effect. Such a finding and the even stronger finding of crossing of predicted quantiles constitute a rejection of the mixed proportional hazard model, analogous to the case without random effects as discussed in Section 8.2.1.

8.3 Summary

This survey summarizes recent estimation approaches using quantile regression for (right-censored) duration data. We provide a discussion of the advantages and drawbacks of quantile regression in comparison to popular alternative methods such as the (mixed) proportional hazard model or the accelerated failure time model. We argue that quantile regression methods are robust and flexible in a sense that they can capture a variety of effects at different quantiles of the duration distribution.

Our theoretical considerations suggest that ignoring random effects is likely to have a smaller effect on quantile regression coefficients than on estimated hazard rates of proportional hazard models. Quantile regression does not impose a proportional effect of the covariates on the hazard. The proportional hazard model is rejected empirically when the estimated quantile regression coefficients change sign across quantiles and we show that this holds even in the presence of unobserved heterogeneity. However, in contrast to the proportional hazard model, quantile regression can not take account of time-varying covariates and it has not been extended so far to allow for unobserved heterogeneity and competing risks. We also discuss and slightly modify the simulation approach for the estimation of hazard rates based on quantile regression coefficients, which has been suggested recently by Machado and Portugal (2002) and Guimarães *et al.* (2004).

We motivate our theoretical considerations with selected results of an application to unemployment duration data. It shows that estimated coefficients vary or change sign over the quantiles. Estimated hazard rates indicate that the proportional hazard assumption is violated in the underlying application. A detailed presentation of the results and the data can be found in the long version of this paper (Fitzenberger and Wilke, 2005).

References

BILIAS, Y., CHEN, S., YING, Z. (2000). Simple resampling methods for censored regression quantiles. *Journal of Econometrics* **99** 373–386.

BUCHINSKY, M. (1995). Quantile regression, Box-Cox transformation model, and the U.S. wage structure, 1963-1987. *Journal of Econometrics* **65** 109–154.

BUCHINSKY, M. (1998). Recent advances in quantile regression models: A practical guideline for empirical research. *Journal of Human Resources* **33** 88–126.

BUCHINSKY, M. (2001). Quantile regression with sample selection: Estimating women's return to education in the US. *Empirical Economics* **26** 87–113.

BUCHINSKY, M., HAHN, J. (1998). An alternative estimator for the censored quantile regression model. *Econometrica* **66** 653–672.

CHAMBERLAIN, G. (1994). Quantile regression, censoring, and the structure of wages. In *Advances in Econometrics: Sixth World Congress, Volume 1* (C. Sims, ed.), Econometric Society Monograph No. 23, Cambridge University Press, Cambridge.

CHERNOZHUKOV, V., HONG, H. (2002). Three-step censored quantile regression and extramarital affairs. *Journal of the American Statistical Association* **97** 872–882.

FITZENBERGER, B. (1997). A guide to censored quantile regressions. In *Handbook of Statistics, Volume 15: Robust Inference* (G. S. Maddala, C. R. Rao, eds.), 405–437, North-Holland, Amsterdam.

FITZENBERGER, B. (1998). The moving blocks bootstrap and robust inference for linear least squares and quantile regressions. *Journal of Econometrics* **82** 235–287.

FITZENBERGER, B., WILKE, R. A. (2005). Using quantile regression for duration analysis. ZEW Discussion Paper, 05-58.

FITZENBERGER, B., WILKE, R. A., ZHANG, X. (2004). A note on implementing box-cox regression. ZEW Discussion Paper, 04-61.

FITZENBERGER, B., WINKER, P. (2001). Improving the computation of censored quantile regressions. Discussion Paper, University of Mannheim.

GUIMARÃES, J., MACHADO, J. A. F., PORTUGAL, P. (2004). Has long become longer and short become shorter? Evidence from a censored quantile regression analysis of the changes in the distribution of U.S. unemployment duration. Unpublished discussion paper, Universidade Nova de Lisboa.

HODERLEIN, S., MAMMEN, E. (2005). Partial identification and nonparametric estimation of nonseparable, nonmonotonous functions. Unpublished Discussion Paper, University of Mannheim.

HOROWITZ, J., NEUMANN, G. (1987). Semiparametric estimation of employment duration models. *Econometric Reviews* **6** 5–40.

KIEFER, N. M. (1988). Economic duration data and hazard functions. *Journal of Economic Literature* **26** 649–679.

KOENKER, R., BASSETT, G. (1978). Regression quantiles. *Econometrica* **46** 33–50.

KOENKER, R., BILIAS, Y. (2001). Quantile regression for duration data: A reappraisal of the Pennsylvania reemployment bonus experiments. *Empirical Economics* **26** 199–220.

KOENKER, R., GELING, O. (2001). Reappraising medfly longevity: A quantile regression survival analysis. *Journal of the American Statistical Association* **96** 458–468.

KOENKER, R., HALLOCK, K. (2002). Quantile regression. *The Journal of Economic Perspectives* **15** 143–156.

LANCASTER, T. (1990). *The econometric analysis of transition data.* Econometric Society Monographs No. 17, Cambridge University Press, Cambridge.

MACHADO, J. A. F., PORTUGAL, P. (2002). Exploring transition data through quantile regression methods: An application to U.S. unemployment duration. In *Statistical data analysis based on the L1-norm and related methods – 4th International Conference on the L1-norm and Related Methods* (Y. Dodge, ed.), Birkhäuser, Basel.

PORTNOY, S. (2003). Censored regression quantiles. *Journal of the American Statistical Association* **98** 1001–1012.

POWELL, J. L. (1984). Least absolute deviations estimation for the censored regression model. *Journal of Econometrics* **25** 303–325.

POWELL, J. L. (1986). Censored regression quantiles. *Journal of Econometrics* **32** 143–155.

POWELL, J. L. (1991). Estimation of monotonic regression models under quantile restrictions. In *Nonparametric and semiparametric methods in Econometrics* (W. Barnett, J. Powell, G. Tauchen, eds.), 357–384, Cambridge University Press, New York.

SILVERMAN, B. W. (1986). *Density estimation for statistics and data analysis.* Chapman and Hall, London.

VAN DEN BERG, G .J. (2001). Duration models: Specification, identification and multiple durations. In *Handbook of Econometrics* (J. J. Heckman, E. Leamer, eds.), 3381–3460, Volume 5, Elsevier, Amsterdam.

WOOLDRIDGE, J. M. (2002). *Econometric analysis of cross section and panel data.* MIT Press, Cambridge.

ZHANG, X. (2004). Neuere Entwicklungen in der Analyse von Verweildauermodellen mit Quantilsregressionen als Alternative zum konventionellen Modell der proportionalen Hazardrate. Diploma Thesis, University of Mannheim.

9 Multilevel and Nonlinear Panel Data Models *

Olaf Hübler[1]

[1] Institute of Empirical Economic Research, University of Hannover
huebler@ewifo.uni-hannover.de

Summary: This paper presents a selective survey on panel data methods. The focus is on new developments. In particular, linear multilevel models, specific nonlinear, nonparametric and semiparametric models are at the center of the survey. In contrast to linear models there do not exist unified methods for nonlinear approaches. In this case conditional maximum likelihood methods dominate for fixed effects models. Under random effects assumptions it is sometimes possible to employ conventional maximum likelihood methods using Gaussian quadrature to reduce a T-dimensional integral. Alternatives are generalized methods of moments and simulated estimators. If the nonlinear function is not exactly known, nonparametric or semiparametric methods should be preferred.

9.1 Introduction

Use of panel data regression methods has become popular as the availability of longitudinal data sets has increased. Panel data usually contain a large number of cross section units which are repeatedly observed over time. The advantages of panel data compared to cross section data and aggregated time series data are the large number of observations, the possibility to separate between cohort, period and age effects. Furthermore, we can distinguish between intra- and interindividual effects and we can determine causal effects of policy interventions. New problems with panel data arise in comparison to cross section data by attrition, time-varying sample size and structural changes.

The modelling of panel data approaches distinguishes in the time dependence, in the assumptions of the error term and in the measurement of dependent variables. Due to the specific assumption consequences for the estimation methods follow. Apart from classical methods like least squares and maximum likelihood estimators, we find in panel data econometrics conditional and quasi ML estimators, GEE

*Helpful comments and suggestions from an unknown referee are gratefully acknowledged.

(generalized estimating equations), GMM (generalized methods of moments), simulated, non- and semiparametric estimators. For linear panel data models with predetermined regressors we can apply conventional techniques. The main objective is to determine and to eliminate unobserved heterogeneity. Two situations are distinguished: regressors and unobserved heterogeneity are independent or interact. Much less is known about nonlinear models. In many models not only simple first differences methods, but also conditional likelihood approaches fail to eliminate unobserved heterogeneity. As the specification of nonlinearity is often unknown, non- and semiparametric methods are preferred.

We distinguish between several types of panel data models and proceed from general to more specific models. The endogenous variable y_{it} can be determined by the observed exogenous time invariant ($\tilde{x}_i$) and time-varying (x_{it}) variables, unobserved time invariant regressors (α_{i*}) and a time-varying error term (u_{it}). The term $m(\cdot)$ tells us that the functional relation is unknown, i. e. nonparametric approaches are formulated, which may vary between the periods ($m_t(\cdot)$). If y_{it} is not directly determined by $\tilde{x}_i$, x_{it}, α_{i*} and u_{it}, but across an unobservable variable, we call this a latent model, expressed by $g(\cdot)$. Furthermore, this relation may be time-varying ($g_t(\cdot)$). This generalized time-varying nonparametric latent model can be presented by

$$y_{it} = g_t[m_t(\tilde{x}_i, x_{it}, \alpha_{i*}, u_{it})]\,. \tag{9.1}$$

Simplifications are possible and can lead to conventional linear models with individual effects and time varying coefficients.

9.2 Parametric Linear and Multilevel Models

Standard panel data analysis starts with linear models

$$y_{it} = x'_{it}\beta + u_{it} \qquad i = 1, ..., N \quad t = 1, ..., T\,, \tag{9.2}$$

where y is the dependent variable, x is a K×1 regressor vector, β is a K×1 vector of coefficients and u is the error term. The number of cross section observations is N and these units are repeatedly measured. When the cross sectional data are only pooled over T periods the coefficients can be estimated by OLS under classical assumptions about the error term. If an unobserved time invariant individual term α_i is incorporated, model (9.2) turns into

$$y_{it} = x'_{it}\beta + \alpha_i + \epsilon_{it} =: x'_{it}\beta + u_{it}\,. \tag{9.3}$$

The methods which are developed for this purpose depend on the assumptions of the error term, the regressand, the regressors and the coefficients of the model. Some panel data sets cannot be collected every period due to lack of resources or

cuts in funding. This missing value problem leads to unbalanced panels. Wandsbeek and Kapteyn (1989) or Davis (2002) for example study various methods for this unbalanced model. If we assume under classical conditions of ϵ_{it} that α_i is uncorrelated with ϵ_{it} and the observed regressors x_{it}, we call this a 'random effects model - REM'. This definition follows Wooldridge (2002, p. 252) who argues that conventional discussions about whether the individual effect should be treated as a random variable or as a parameter are wrongheaded for microeconometric panel data applications. The coefficient vector $\beta = (\beta_1, \beta^{*'})$ is determined by OLS of the transformed model

$$y_{it} - \hat{\delta}\bar{y}_i = (1 - \hat{\delta})\beta_1 + (x_{it} - \hat{\delta}\bar{x}_i)'\beta^* + u_{it} - \hat{\delta}\bar{u}_i\,, \tag{9.4}$$

where $\hat{\delta} = 1 - [\hat{\sigma}_\epsilon^2/(\hat{\sigma}_\epsilon^2 + T\hat{\sigma}_\alpha^2)]^{1/2}$. The variance of $\hat{\epsilon}$ can be estimated by the residuals of the within estimator, the OLS estimator of (9.4), and the estimated variance of $\hat{\alpha}$ follows by $\hat{\sigma}_\alpha^2 = [1/(N-K)]\sum_i \hat{u}_i^2 - (1/T)\hat{\sigma}_\epsilon^2$, where the average residuals $\hat{\bar{u}}$ are determined by the between estimator, i. e. the OLS estimator of $\bar{y}_i = \sum_{k=2}^{K} \bar{x}_{ik}\beta_k + \bar{u}_i$. If α_i and the regressors are correlated we denote the approach 'fixed effects model - FEM'. In this case the OLS estimator of the transformed model

$$y_{it} - \bar{y}_i = (x_{it} - \bar{x}_i)'\beta^* + u_{it} - \bar{u}_i \tag{9.5}$$

is determined, where β^* is the coefficient vector without a constant term. The estimated individual effect is

$$\hat{\alpha}_i = (y_{it} - \bar{y}_i) - (x_{it} - \bar{x}_i)'\hat{\beta}^*. \tag{9.6}$$

The significance of individual effects can be tested by an F test. The null hypothesis of model

$$y_{it} - \hat{\delta}\bar{y}_i = (x_{it} - \hat{\delta}\bar{x}_i)'\beta + (x_{it} - \bar{x}_i)'\tilde{\beta}^* + \omega_{it} \tag{9.7}$$

is $H_0 : \tilde{\beta}^* = 0$.

First order serial correlation and heteroscedasticity can be allowed. For example, a four step procedure solves the problem of determining the autocorrelation coefficient, ρ, and β where two transformations are necessary (Hübler, 1990). If lagged dependent or jointly dependent variables or errors in the exogenous variables exist, a GMM estimator suggested by Arellano and Bond (1991) is preferred. Several other alternatives are usually less efficient or even inconsistent. However, Ahn and Schmidt (1995) and Blundell and Bond (1998) have formulated further conditions of moments which lead to better estimators. Hsiao *et al.* (2002) have presented a maximum likelihood estimator which is more efficient than the GMM estimator. If the simple lagged dependent model

$$y_{it} = \gamma y_{i,t-1} + \alpha_i + \varepsilon_{it} = \gamma y_{i,t-1} + u_{it} \tag{9.8}$$

exists, conventional first differences eliminate the individual term. However, such an OLS estimator is biased compared to IV estimators. GMM estimators which incorporate all valid instruments are preferred. Valid instruments are those which

are uncorrelated with the error term. This means that if observations from T waves are available, valid instruments Z fulfill the following orthogonality conditions: $E(Z'u) = 0$. Coefficients are determined by a minimum distance estimator

$$\arg\min_{\gamma}(\xi - \mathrm{E}(\xi_i))'W(\xi - \mathrm{E}(\xi_i)) = \arg\min_{\gamma}\xi'W\xi\,, \tag{9.9}$$

where W is a weighting matrix and estimation of γ is based on the empirical moments. The optimal weighting matrix W is the inverse of the covariance matrix of ξ_i

$$\hat{V}_N^{-1} = \left[\frac{1}{N}\sum_{i=1}^{N}\hat{\xi}_i\hat{\xi}_i'\right]^{-1} = \left[\frac{1}{N}\sum_{i=1}^{N}Z_i'd\hat{u}_i d\hat{u}_i'Z_i\right]^{-1}. \tag{9.10}$$

This procedure can be extended to models with additional regressors. Until now we have assumed that all coefficients including the individual effect are time invariant. The following modification is less restrictive

$$y_{it} = \gamma_t y_{i,t-1} + x_{it}'\beta_t + \psi_t\mu_i + \varepsilon_{it}\,, \tag{9.11}$$

where the individual effect $\psi_t\mu_i$ is time-varying in contrast to conventional panel data models. One can argue that the effect varies e. g. with cyclical ups and downs, although individual characteristics stay the same. Chamberlain (1984, p. 1263) suggests a solution to determine the coefficients in (9.11). This equation in period t-1 is multiplied by $r_t = \psi_t/\psi_{t-1}$ and this expression is subtracted from (9.11), i. e. $y_{it} - r_t y_{i,t-1}$. Then the individual effect disappears. A simultaneous equation system results

$$\begin{aligned}
y_{i3} &= (\gamma_3 + r_3)y_{i2} - r_3\gamma_2 y_{i1} + x_{i3}'\beta_3 - r_3 x_{i2}'\beta_2 + (\varepsilon_{i3} - r_3\varepsilon_{i2})\\
y_{i4} &= (\gamma_4 + r_4)y_{i3} - r_4\gamma_3 y_{i2} + x_{i4}'\beta_4 - r_4 x_{i3}'\beta_3 + (\varepsilon_{i4} - r_4\varepsilon_{i3})\\
&\vdots\\
y_{iT} &= (\gamma_T + r_T)y_{i,T-1} - r_T\gamma_{T-1}y_{i,T-2} + x_{iT}'\beta_T\\
&\quad - r_T x_{i,T-1}'\beta_{T-1} + (\varepsilon_{iT} - r_T\varepsilon_{i,T-1})\,.
\end{aligned} \tag{9.12}$$

Equations y_{i1} and y_{i2} are suppressed as we have no information about y_{i0} and $y_{i,-1}$. As the error terms and the lagged dependent variables are correlated, instrumental variables are used. Therefore, y_{i3} is also eliminated, because not enough instruments can be constructed. Consequently, only the equations of period 4 to T can be used to determine the coefficients.

A critical assumption of the previous models, with exception of the last one, is the general constancy of individual effects. We cannot expect that unobserved individual abilities have the same effects in different situations. One possibility, the time dependence, is described in (9.11). A natural extension is obtained when units are not only separated by one criterion, i=1,...,N, and then repeatedly observed over time (t=1,...,T), but further levels are considered, e. g. establishments and industries. We call this a 'multilevel model'. Recently, especially two-level models, i. e.

linked employer-employee panel data (LEEP) models, have been introduced in the literature (Abowd *et al.*, 2002; Abowd and Kramarz, 1999; Abowd *et al.*, 1999; Goux and Maurin, 1999). In this context the basic model assumes fixed effects and is described by

$$y = X\beta + D\alpha + F\psi + \epsilon, \tag{9.13}$$

where y is a N·T×1 vector, X is a N·T×K matrix. The design matrix for the individual effects D has the order N·T×N containing unit vectors. Analogously, the design matrix for the firm effects F is a N·T×J matrix. The firm effect is expressed by $\psi_{J(it)}$. This means individual i in period t is assigned to one of the j=1,...,J establishments. The conventional technique to estimate β, α and ψ in a partitioned regression (9.13) does not work. The usual way to 'sweep out' the D matrix and then to determine the firm effects, cannot be used in practice. The F matrix is too large a non-patterned matrix due to the large number of firms. Identification of the individual and firm effects in order to estimate using the exact least squares estimator requires finding the conditions under which equation (9.13) can be solved for some subset of the person and firm effects. Abowd *et al.* (2002) present a procedure by applying methods from graph theory to determine groups of connected individuals and firms. Within a connected group identification can be determined using conventional methods from the analysis of covariance. A group contains all workers who have ever worked for any of the firms in the group and all the firms at which any of the workers were ever employed. The algorithm constructs G mutually-exclusive groups of connected observations from the N workers in J firms observed over the sample period. However, usually approximate solutions to (9.13) are employed (Abowd *et al.*, 1999). For this purpose an extended system is formulated, which is more easily manageable under specific restrictions

$$y = X\beta + D\alpha + Z\lambda + M_Z F\psi + \epsilon, \tag{9.14}$$

where $\lambda = (Z'Z)^{-1}Z'F\psi$ denotes an auxiliary parameter, $M_Z = I - Z(Z'Z)^{-1}Z'$. The new matrix Z contains specific columns of X, D and F. The intention behind creating Z is to incorporate all relevant variables which determine interaction effects between X, F and D so that under the condition of Z, orthogonality conditions can be formulated. The selection of this information is a similar problem to the choice of instrumental variables. The following restrictions are imposed: (i) X and D are orthogonal, given Z, (ii) D and F are orthogonal, given Z. Then a four-step procedure can be applied.

If the model is extended by a further level, e. g. by industry effects, we have to consider that each establishment is assigned to only one industry. The gross firm effect $F\psi$ has to be separated into the net firm effect $(F\psi - FA\kappa)$ and an industry effect $(FA\kappa)$ Therefore, model (9.13) passes into

$$y = X\beta + D\alpha + FA\kappa + (F\psi - FA\kappa) + \epsilon. \tag{9.15}$$

Matrix A assigns firms to a specific industry ($a_{jl} = 1$, if firm j belongs to industry l; $a_{jl} = 0$ otherwise).

If firm effects and individual effects are suppressed and if they are effective, the industry effect is biased except when FA and $M_{FA}F$ are orthogonal. The bias is described by a weighted average of the individual and the firm effects. Hildreth and Pudney (1999) discuss issues of non-random missing values and Goux and Maurin (1999) emphasize the use of instrumental variables estimators, which are necessary if the variables are jointly dependent or if errors in variables exist.

While Hausman tests usually reject the random effects model (REM), the fixed effects model (FEM) has the problem that the within transformation of a model wipes out time invariant regressors as well as the individual effect, so that it is not possible to estimate the effects of those regressors on the dependent variable. One way to solve this problem is to replace random individual effects by estimated FE. The basic idea follows the sample selection approach. Heckman (1979) has substituted the conditional expected error term by an estimate, which is employed as an artificial regressor. Let us manage this issue in a two-level model without pure individual effects

$$y_{ijt} = x'_{ijt}\beta + \alpha_j + \alpha_{ij} + \epsilon_{ijt}\,. \tag{9.16}$$

In the first step the general firm effects α_j and the firm specific individual effects α_{ij} are estimated by the within estimator of a FEM

$$\hat{\alpha}_j = (\bar{y}_j - \bar{y}) - (\bar{x}^*_j - \bar{x}^*)'\hat{\beta}^*\,,$$

$$\hat{\alpha}_{ij} = (\bar{y}_{ij} - \bar{y}_j) - (\bar{x}^*_{ij} - \bar{x}^*_j)'\hat{\beta}^*\,, \tag{9.17}$$

where β^* is the coefficient vector without the constant term. The conventional RE estimator is inadequate if firm effects and regressors are correlated. In the second step the firm effects are substituted by the estimates of (14.18). We incorporate the firm effects as linear combinations ($a_1\hat{\alpha}_j$ and $b_1\hat{\alpha}_{ij}$) and expect $\hat{a} = 1, \hat{b} = 1$. OLS estimation of

$$y_{ijt} = x'_{ijt}\beta + a\hat{\alpha}_j + b\hat{\alpha}_{ij} + \epsilon_{ijt} \tag{9.18}$$

leads to new estimates of the firm effects $\hat{\hat{\alpha}}_j = \bar{y}_j - \bar{x}'_j\hat{\beta}$ and $\hat{\hat{\alpha}}_{ij} = \bar{y}_{ij} - \bar{x}'_{ij}\hat{\beta}$.

9.3 Parametric Nonlinear Models

The major problem of nonlinear panel data models is the removal of the individual effect. The main limitation of much of the literature on nonlinear panel data methods is that the explanatory variables are assumed to be strictly exogenous. The discussion focusses on models, in which the parameter that is usually interpreted as an intercept is allowed to be specific of the individual level. Unfortunately, the features of the model that do not depend on α_i tend to be different for the different nonlinear functional forms. Therefore the resulting estimation procedures are

different for different models. This is somewhat unsatisfactory. Many methods with fixed effects rely on the method of conditional ML. Under random effects, one can attempt to employ conventional ML methods. However, the combination of typical nonlinear models with panel structures often yields too complex models to use this procedure. Two alternatives exist. On the one hand, simulated ML estimation is manageable. On the other hand, the GMM approach is a good alternative.

The most popular examples in microeconometric nonlinear models are logit, probit, count data, censored and selection models. This is due to the fact that the majority of panel data on individual and firms level are not continuously measured. In the meanwhile there exist a lot of empirical applications using specific nonlinear approaches (see Hübler, 2005). Specially in labor economics, but also in other fields of applied microeconomics many problems are based on qualitative, discrete or censored variables and subgroups are considered. The paper concentrates on methods of binary dependent variables. Others are discussed in Hübler (2003, 2005). Conditional maximum likelihood estimators are right for logit models with fixed effects (Hsiao, 2004). The basic model is

$$y_{it}^{*} = x_{it}'\beta + \alpha_i + \epsilon_{it} = x_{it}'\beta + u_{it}\,,$$

$$y_{it} = \begin{cases} 1 \text{ , if } \quad y_{it}^{*} \geq 0 \\ 0 \text{ , otherwise}\,, \end{cases} \tag{9.19}$$

and the probability is

$$P(y_{it} = 1) = \frac{\exp\,(x_{it}'\beta + \alpha_i)}{1 + \exp\,(x_{it}'\beta + \alpha_i))}\,, \tag{9.20}$$

where $i = 1, ..., N; t = 1, ..., T$. A simple ML estimator is inconsistent, as FEM's allow a correlation between x and α. The literature has developed alternative estimation strategies. The general idea is that, although the model does not have features that are linear in the α_i's, it is nonetheless sometimes possible to find features of the model that do not depend on α_i. One way is to use the conditional maximum likelihood estimation (CML).

A generalization of the standard logit panel data model is presented by Revelt and Train (1998). They analyze a multinomial panel model and allow the parameters associated with each observed variable to vary randomly across individuals. Conditional on β_i, the probability that person i has the observed sequence of choices is the product of standard logits ($S_i(\beta_i) = \Pi_t P_{it}(\theta^*)$). Exact maximum likelihood estimation is not possible since the integral cannot be calculated analytically. Instead, it is possible to approximate the probability by simulation and maximize the simulated log likelihood. The average of the replicated results is taken as the estimated choice probability. The estimated parameters of the simulated log likelihood function are consistent and asymptotically normal under regularity conditions. These mixed logit approaches do not require the independence of irrelevant alternatives and a general pattern of correlation over alternatives and time are allowed.

As not so much is known about how to deal with fixed effects, it is often appealing to make assumptions on the distribution of individual effects. We cannot find simple functions for the parameters of interest that are independent of the nuisance parameter α_i for probit models. To obtain the ML estimator we must evaluate T-dimensional integrals. Butler and Moffitt (1982) simplify the computation by

$$P(Y_{i1} = y_{i1}, ..., Y_{iT} = y_{iT}) = \int_{-\infty}^{\infty} f(\alpha_i) \prod_{t=1}^{T} [F(\infty|\alpha_i) - F(-x'_{it}\beta|\alpha_i)] d\alpha_i \,. \quad (9.21)$$

A more efficient alternative suggested by Chamberlain (1984) yields the minimum distance estimator which avoids numerical integration.

In the pure random effects model one can also estimate the model by a pseudo-maximum likelihood method that ignores the panel structure altogether. The basic idea can be described as follows: the time correlation structure is seen as 'nuisance' only with subordinated interest. Due to possible misspecification of this correlation structure, the application of the ML method is not completely valid. Therefore, this approach is called a quasi- or pseudo-ML estimation (QML) in the literature. The objective is to minimize

$$S = \sum_{i=1}^{N} (y_i - F_i(x_i.\beta))' \Omega_i^{-1} (y_i - F_i(x'_i.\beta)) \,, \quad (9.22)$$

where

$$y_i = (y_{i1}, \cdots, y_{iT})' \,, \quad F_i(\cdot) = (F(x'_{i1}\beta), \ldots, F(x'_{iT}\beta))' \,.$$

If Ω is known, the LS estimator fulfills the equation

$$\sum_{i=1}^{N} \frac{\partial F_i(\cdot)}{\partial \beta} \Omega_i^{-1} (y_i - F_i(\cdot)) = 0 \,. \quad (9.23)$$

If Ω is unknown, a 'working correlation matrix-$\tilde{\Omega}$' is employed, which is usually misspecified, e. g. $R(\delta) = I$ or it is assumed that the correlations outside the main diagonals are equal. The equations are then called 'generalized estimating equations - GEE' and the solution is in accordance with a QML estimation. If the specification of $F(x'_{it}\beta)$ is correct the QML estimator is consistent and asymptotically normally distributed, provided that the estimation of the covariance matrix $\hat{\beta}$ is robust.

If we consider a multinomial probit panel data model, the CML method fails. The Butler-Moffitt approach is restricted because of the underlying multidimensional integral. As mentioned in the introduction of nonlinear models, two alternatives exist: simulated estimation methods and GMM approaches.

GMM estimators are based on the orthogonality conditions implied by the single equation conditional mean functions

$$E(y_{it} - F(x'_{it}\beta|X_i)) = 0 \,, \quad (9.24)$$

where $F(\cdot)$ denotes the CDF of the univariate normal distribution. By combining classical estimation methods and simulators, several approaches were developed. For example, simulated maximum likelihood methods (SMLM), including the GHK estimator, can be used (Geweke *et al.*, 1997). Keane (1994) derived a computationally practical simulator for the panel probit model. Simulation methods replace the intractable integrals by unbiased Monte Carlo probability simulators. Further possibilities are the method of simulated moments, simulated scores and Markov chain Monte Carlo including Gibbs and Metropolis-Hastings algorithm. Geweke *et al.* (1997) find that Gibbs sampling, simulated moments and maximum likelihood method using the GHK estimator, all perform reasonably well in point estimation of parameters in a three alternative 10-period probit model. Monte Carlo studies of nonlinear panel data models (Bertschek and Lechner, 1998; Breitung and Lechner, 1999) reveal that among different GMM estimators the ranking is not so obvious, while MLE performs best followed by the GMM estimator based on the optimal instruments derived from the conditional mean restrictions. Greene (2004) also find that the GMM estimator performs fairly well compared to the ML estimation.

9.4 Non- and Semiparametric Models

In many cases economic theory cannot help to specify an exact nonlinear relationship. Usually, one can only argue that saturation, production and growth processes or other mechanisms suggest a nonlinear approach without predetermined functional form. Nonparametric models should be a natural consequence. Nevertheless so far, the number of applications in this field is restricted (Hübler, 2005). One reason is that the estimation methods are less known than parametric approaches among empirical researchers and that these methods are not implemented in conventional software packages. Especially, we need more spread and further developments of non- and semiparametric methods for panel data. In a general formulation the causal dependence of the dependent variable y_{it} on independent variables and the error term is typically described by

$$y_{it} = g[m(x_{it}) + u_{it}], \tag{9.25}$$

where $g(\cdot)$ calls a mapping which induces the variable y_{it}. Possibly, y_{it} depends on an unobserved endogenous variable $y_{it}^* = m(x_{it} + u_{it})$. If y_{it} is directly created by x_{it} and u_{it} the relation can be simplified by $y_{it} = m(x_{it}) + u_{it}$ and the linear model results if $m(x_{it}) = x_{it}'\beta$. While parametric models assume a known structural relation under unknown parameters and an error term, a simple nonparametric panel data model formulates a mean regression

$$y_{it} = E(y_{it}|x_{it}) + u_{it} = m(x_{it}) + u_{it}\,. \tag{9.26}$$

The higher degree of flexibility motivates to use nonparametric rather than parametric methods. The basic problem is the enormous amount of calculations, especially if the number of regressors increase. Furthermore, it is difficult to interpret the estimation. This is called 'the curse of dimension'. Two possibilities exist to

solve this problem. Additive or partial linear models are assumed. The former are discussed in Hastie and Tibshirani (1997). We focus on the presentation of partial linear models.

If a pooled estimation of panel data is employed, the same procedures as with cross section data can be used. It is necessary to test whether the pooled procedure is convenient. Parametric procedures are described by Baltagi (2001). Baltagi *et al.* (1996) present a nonparametric test. Li and Hsiao (1998) test whether individual effects exist. If the test does not reject the null hypothesis of poolability, the individual effects can be neglected. Starting point is a partial linear approach of panel data

$$y_{it} = z_{it}'\gamma + m(x_{it}) + u_{it}\,. \tag{9.27}$$

The basic idea is to eliminate the nonparametric part. Then the linear term can be estimated separately following Robinson (1988). In other words, the new regressand is the difference between y and the conditional expected value, which is induced by the nonparametric regressors. Due to the identity $E[m(x)|x] = m(x)$, the nonparametric term vanishes by first differences

$$y - E(y|x) = (z - E(z|x))'\gamma + u\,. \tag{9.28}$$

Before we can estimate γ, the conditional expected value has to be determined by a nonparametric procedure (Pagan and Ullah, 1999, p. 199)

$$\hat{y}_{it} = \hat{E}(y_{it}|x_{it}) = \frac{1}{NT \cdot h} \sum_{j=1}^{N} \sum_{t=1}^{T} \frac{K\left(\frac{x_{it}-x_{jt}}{h}\right)}{\hat{f}_{it}} y_{jt}\,, \tag{9.29}$$

where $\hat{f}_{it} = \frac{1}{N \cdot h} \sum_{j=1}^{N} \sum_{t=1}^{T} K(\frac{x_{it}-x_{jt}}{h})$ is the kernel density estimator and h is the bandwidth. If we do not only consider a univariate nonparametric term, a multivariate kernel is necessary. A simplified form can be assumed in this case, namely the product of the univariate kernels, i.e. $K(x_{it}) = \Pi_{d=1}^{D} K(x_{dit})$. The differences model (9.28) has to be weighted

$$\hat{f}_{it}(y_{it} - \hat{y}_{it}) = \hat{f}_{it}(z_{it} - \hat{z}_{it})'\gamma + \hat{f}_{it}u_{it}\,. \tag{9.30}$$

The least squares estimator of γ follows

$$\hat{\gamma} = (\sum_{i=1}^{N} \sum_{t=1}^{T} (z_{it} - \hat{z}_{it})(z_{it} - \hat{z}_{it})' \hat{f}_{it}^2)^{-1} (\sum_{i=1}^{N} \sum_{t=1}^{T} (z_{it} - \hat{z}_{it})(y_{it} - \hat{y}_{it}) \hat{f}_{it}^2)\,. \tag{9.31}$$

This OLS estimator is consistent. However, a GLS estimator presented by Li and Ullah (1998) achieves the semiparametric efficiency border and is therefore superior. In the second step we obtain the multivariate nonparametric term $m(x)$ under a Taylor series approximation

$$y_{it} - z_{it}'\hat{\gamma} = m(x_{it}) + \beta(x)'(x_{it} - x) + R(x_{it}, x) + z_{it}'(\gamma - \hat{\gamma}) + u_{it}$$

$$= m(x_{it}) + \beta(x)'(x_{it} - x) + \tilde{u}_{x,it}\,. \tag{9.32}$$

The parameter vector $\tilde{\beta}(x)' = (m(x), \beta(x)')$ results from

$$\hat{\tilde{\beta}}(x) = argmin \frac{1}{NT \cdot h} \sum_{i=1}^{N} \sum_{t=1}^{T} K_{it}(y_{it} - z'_{it}\hat{\gamma}$$

$$-m(x) - (x_{it} - x)'\beta(x))^2, \tag{9.33}$$

where $K_{it} = K(\frac{x_{it}-x}{h})$. Now the local linear least squares estimator is determined.

If the following simple nonparametric model is extended by individual effects α_i, all conditional moment procedures can be used to estimate the nonparametric term $m_t(x_{it})$, if a random effects model is convenient. A known distribution of the individual effects is assumed and α_i are identical, independent distributed. Furthermore, the individual effects are independent of the regressors. Under unspecific time dependence of $E(y_{it}|x_{it})$ it is possible to estimate the parameters separately for each wave. If time invariance is assumed, i.e. $m_t(x_{it}) = m(x_{it})$, the pooled procedure can be employed to determine the nonparametric term. A local linear approach is possible. In order to determine the nonparametric terms $m(x)$ and $\beta(x)$, we can choose Ullah and Roy's (1998) GLS approach. This corresponds to the conventional within transformation,where α is eliminated. Under semiparametric partial linear panel data models

$$y_{it} = m(x_{it}) + z'_{it}\gamma + \alpha_i + \epsilon_{it} \tag{9.34}$$

we can follow Li and Stengos (1996) analogously to pooled models.

The estimator of the time invariant nonparametric term of a balanced panel can be assigned to an unbalanced panel (König, 1997). An extension to models with time variable nonparametric models is also possible (König, 2002). In this case a wave specific procedure is suggested. An alternative to Li and Stengos is developed by König (2002, p. 176ff). It is also possible to model a time variable nonparametric term. Conventional first differences and within estimators, well-known from pure linear models, can be applied. This approach is independent from the sensitivity of bandwidth.

The partial linear model may be interpreted as a simple linear model with fixed effects if nonparametric regressors are time invariant ($x_{it} = x_i$). The parameter vector γ is determined by first differences or within estimators if the linear regressors are strictly exogenous. Biased estimators result from a direct application to time variable regressors. The bias can be reduced significantly, if we follow a suggestion by König (2002, p. 182).

In contrast to random effects models there exists an additional problem in nonparametric panel data models with fixed effects. Due to the allowed correlation between α_i and x_{it} the conditional expected value of y_{it} differs from the nonparametric term. Instead we obtain

$$E(y_{it}|x_{it}) = m(x_{it}) + E(\alpha_i|x_{it})\,. \tag{9.35}$$

Therefore, it is not possible to determine the nonparametric part by the conditional moment approach. The conventional solution by first differences or within estimators breaks down. The individual effect is eliminated but not identified by this procedure. Ullah and Roy (1998) suggest a Taylor series of the nonparametric expression as a starting point

$$y_{it} = m(x) + (x_{it} - x)' \frac{\partial m(x_{it})}{\partial x_{it}}|_{x_{it}=x} + \frac{1}{2}(x_{it} - x)' \frac{\partial^2 m(\tilde{x})}{\partial x \partial x'}(x_{it} - x) + \alpha_i + \epsilon_{it}$$

$$=: m(x) + (x_{it} - x)'\beta(x) + R_2(x_{it}, x) + \alpha_i + \epsilon_{it}$$

$$=: m(x) + (x_{it} - x)'\beta(x) + \alpha_i + \tilde{\epsilon}_{x,it}\,, \tag{9.36}$$

where $\tilde{x}$ is assumed to be within the range x and x_{it}. It is intended to estimate $\beta(x)$ of this local linear model. A local within estimator with simple kernel function weights gives biased and inconsistent estimates due to residual terms $E(\tilde{R}_2(x_{i,x}|x_{it} = x) \neq 0)$. The same problem follows under analogous first differences estimators. However, a double weighting of first differences eliminates the bias (König 2002, p. 61ff). We define the product kernel from period t and t-1

$$K(\frac{x_{it} - x}{h}, \frac{x_{i,t-1} - x}{h}) = K(\frac{x_{it} - x}{h})\, K(\frac{x_{i,t-1} - x}{h})$$

$$= : K_{it}\, K_{i,t-1}\,, \tag{9.37}$$

where once again h is the bandwidth. Instead of weighting with $K(\frac{x_{it}-x}{h})$ as in a conventional differences estimator, the weight is the product of the local kernels

$$\hat{\beta}(x)_D = \{\sum_{i=1}^{N}\sum_{t=2}^{T} K_{it}K_{i,t-1}\Delta x_{it}\Delta x_{it}'\}^{-1} \sum_{i=1}^{N}\sum_{t=2}^{T} K_{it}K_{i,t-1}\Delta x_{it}\Delta y_{it}\,, \tag{9.38}$$

where $\Delta x_{it} = x_{it} - x_{i,t-1}$ and $\Delta y_{it} = y_{it} - y_{i,t-1}$. This estimator is not only consistent, but also asymptotically normally distributed with a null vector as the expected value vector and an asymptotic sandwich covariance matrix. A similar weighting is possible in the within model.

Semiparametric partial linear models with fixed individual effects can be described by

$$y_{it} = \tilde{m}(x) + x_{it}'\beta(x) + z_{it}'\gamma + \alpha_i + \tilde{\epsilon}_{it}, \tag{9.39}$$

where $\tilde{\epsilon}_{it} = \epsilon_{it} + R(x_{it}, x)$, $\tilde{m}(x) = m(x) - x'\beta(x)$. The individual term α_i may be correlated with x_{it} and z_{it}. The nonparametric term $m(x_{it})$ is developed by a

Taylor series. In this case the problem of the parameter estimation (γ) also persists in the conditional expected value of y_{it}:

$$E(y_{it}|x_{it}) = \tilde{m}(x) + x_{it}'\beta(x) + E(z_{it}|x_{it})'\gamma + E(\alpha_i|x_{it}). \tag{9.40}$$

Differences, i.e. $y_{it} - E(y_{it}|x_{it})$, eliminate the nonparametric term, but not the individual term. Therefore, it is necessary to remove α_i in the first step. Li and Stengos (1996) employ a differences estimator of γ where the estimator is weighted by the Nadaraya-Watson kernel estimator

$$\hat{\gamma}_D = \{\sum_{i=1}^{N}\sum_{t=2}^{T} \Delta\hat{\tilde{z}}_{it}\Delta\hat{\tilde{z}}_{it}'\hat{\tilde{f}}(x_{it}, x_{i,t-1})^2\}^{-1} \sum_{i=1}^{N}\sum_{t=2}^{T} \Delta\hat{\tilde{z}}_{it}\Delta\hat{\tilde{y}}_{it}\hat{\tilde{f}}(x_{it}, x_{i,t-1})^2, \tag{9.41}$$

where $\hat{\tilde{f}}(x_{it}, x_{i,t-1})$ is the kernel density estimator and

$$\Delta\hat{\tilde{z}}_{it} = z_{it} - z_{i,t-1} - [\hat{E}(z_{it}|x_{it}, x_{i,t-1}) - \hat{E}(z_{i,t-1}|x_{it}, x_{i,t-1})].$$

Analogously, $\Delta\hat{\tilde{y}}_{it}$ is defined. By first differences α_i disappears, but the nonparametric term does not. In order to remove the difference $\tilde{m}(x_{it}) - \tilde{m}(x_{i,t-1})$ we additionally have to subtract the difference of the expected values. If $\Delta\hat{\tilde{z}}_{it} = z_{it} - z_{i,t-1} - [\hat{E}(z_{it}|x_{it}, x_{i,t-1}) - \hat{E}(z_{i,t-1}|x_{it}, x_{i,t-1})]$ and the error terms are correlated, $\Delta\tilde{z}_{it}$ has to be instrumented. König (2002, p. 215) suggests an alternative estimator without kernel weights.

Manski (1975, 1987) has developed nonparametric maximum score estimators for panel data models with fixed effects and dichotomous endogenous variables. Further models and an estimator are presented by Lee (1999) and Honore and Lewbel (2002). A survey on Tobit panel data models with nonparametric components, which include the standard case of censored endogenous variables, selection models and censored multivariate models can be found in Kyriazidou (1995, 1997) and Honore and Kyriazidou (2000). They also develop some new variants which do not require the parametrization of the distribution of the unobservables. However, it is necessary that the explanatory variables are strictly exogenous. Therefore, lagged dependent variables as regressors are excluded. Kyriazidou (1997) obtains values near zero by differences between pairs of observations, because pairs with a large difference obtain small weights. Honore (1992) suggests trimmed least absolute deviation and trimmed least squares estimators for truncated and censored regression models with fixed effects. He exploits the symmetry in the distribution of the latent variables and finds that when the true values of the parameters are known, trimming can transmit the same symmetry in distribution to observed variables. One can define pairs of residuals that depend on the individual effect in exactly the same way so that differencing the residuals eliminates the fixed effects.

9.5 Concluding Remarks

Many new methods to estimate panel data models have been developed in the past. The focus in this paper was directed on multilevel and nonlinear models. As the

functional form of nonlinearity is usually unknown, nonparametric estimates are corollary. Nowadays several methods are implemented in conventional packages, but others still require programming. In contrast to linear fixed effects panel data models, it is more difficult to manage the individual term in combination with a nonparametric term. Conventional differences and within estimators do not help to eliminate the latter. There do not exist uniform methods of nonlinear models. We have only specific estimation methods for several forms of nonlinearity and the results depend on the assumptions. While estimation of random effects panel data models is based on a fully specified model in which one can determine all the quantities of interest, fixed effects panel data models typically result in the estimation of some finite dimensional parameters from which one cannot calculate all the functions of the distribution. Nevertheless, progress can also be observed in the estimation of fixed effects panel data models. Estimates of random effects models are usually more efficient. However, very often the violation of the distributional assumptions yields inconsistent estimates. Fixed effects models make fewer assumptions and they react less sensitive to violations of the assumptions. Random effects models are usually preferable for prediction.

In future we have to analyze the dynamic character of the panel data models more completely. Almost nothing is known about nonlinear models with lagged dependent variables. Furthermore, non- and semiparametric methods should also be applied to multilevel models. In many situations it seems helpful to start with nonparametric estimates. However, the next step would be to derive more fully specified parametric models based on the results of the first step.

References

ABOWD, J. M., CREECY, R. H., KRAMARZ, F. (2002). Computing person and firm effects using linked longitudinal employer-employee data. Cornell University, Working Paper.

ABOWD, J. M., KRAMARZ, F. (1999). The analysis of labor markets using matched employer-employee data. In *Handbook of Labor Economics*, (O. Ashenfelter, D. Card, eds.), 2629–2710. Vol. 3B, Elsevier, Amsterdam.

ABOWD, J. M., KRAMARZ, F., MARGOLIS, D. N. (1999). High wage workers and high wage firms. *Econometrica* **67** 251–333.

AHN, S., SCHMIDT, P. (1995). Efficient estimation of models for dynamic panel data. *Journal of Econometrics* **68** 5–27.

ARELLANO, M. (2003). *Panel data econometrics*. University Press, Oxford.

ARELLANO, M., BOND, S. (1991). Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *Review of Economic Studies* **58** 277–297.

BALTAGI, B. (2001). *Econometric Analysis of Panel Data*. 2nd ed., John Wiley & Sons, Chichester.

BALTAGI, B. H., HIDALGO, J., LI, Q. (1996). A nonparametric test for poolability using panel data. *Journal of Econometrics* **75** 345–367.

BERTSCHEK, I., LECHNER, M. (1998). Convenient estimators for the panel probit model. *Journal of Econometrics* **87(2)** 329–372.

BLUNDELL, R., BOND, S. (1998). Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics* **87** 115–143.

BREITUNG, J., LECHNER, M. (1999). Alternative GMM methods for nonlinear panel data models. In *Generalized Method of Moments Estimation* (L. Matyas, ed.), 248–274. University Press, Cambridge.

BUTLER, J., MOFFITT, R. (1982). A computationally efficient quadrature procedure for the one factor multinomial probit model. *Econometrica* **50** 761–764.

CHAMBERLAIN, G. (1984). Panel data. In *Handbook of Econometrics* (Z. Griliches and M. Intriligator, eds.), 1247–1318. North-Holland, Amsterdam.

DAVIS, P. (2002). Estimating multi-way error components models with unbalanced data structure. *Journal of Econometrics* **106** 67–95.

GEWEKE, J., KEANE, M., RUNKLE, D. (1997). Statistical inference in the multinomial multiperiod probit model. *Journal of Econometrics* **80** 125–165.

GOUX, D., MAURIN, E. (1999). Persistence of interindustry wage differentials: A reexamination using matched worker-firm panel data. *Journal of Labor Economics* **17** 492–533.

GREENE, W. (2004). Convenient estimators for the panel probit model: Further results. *Empirical Economics* **29** 21–47.

HASTIE, T. J. TIBSHIRANI, R. J. (1997). *Generalized Additive Models.* Chapman and Hall, London.

HAUSMAN, J. A. (1978). Specification tests in econometrics. *Econometrica* **46** 1251–1272.

HECKMAN, J. J. (1979). Sample selection bias as a specification error. *Econometrica* **47** 153–161.

HILDRETH, A. K., PUDNEY, S. (1999). Econometric issues in the analysis of linked cross-section employer-worker surveys. In *The Creation and Analysis of Employer-Employee Matched Data* (J. Haltiwanger *et al.*, eds.), 461–488. North-Holland, Amsterdam.

HONORE, B. E. (1992). Trimmed LAD and least squares estimation of truncated and censored regression models with fixed effects. *Econometrica* **60** 533–565.

HONORE, B. E., KYRIAZIDOU, E. (2000). Estimation of Tobit-type models with individual specific effects. *Econometric Reviews* **19** 341–366.

HONORE, B. E., LEWBEL, A. (2002). Semiparametric binary choice panel data models without strictly exogenous regressors. *Econometrica* **70** 2053–2063.

HSIAO, C. (2004). *Analysis of panel data.* 2nd ed., University Press, Cambridge.

HSIAO, C., PESARAN, M., TAHMISCIOGLU, A. K. (2002). Maximum likelihood estimation of fixed effects dynamic panel data models covering short time periods. *Journal of Econometrics* **109** 107–150.

HÜBLER, O. (1990). Lineare Paneldatenmodelle mit alternativer Störgrößenstruktur. In *Neuere Entwicklungen in der Angewandten Ökonometrie* (G. Nakhaeizadeh, K.-H. Vollmer, eds.), 65–99. Physica, Heidelberg.

HÜBLER, O. (2003). Neuere Entwicklungen in der Mikroökonometrie. In *Empirische Wirtschaftsforschung - Methoden und Anwendungen* (W. Franz, H. J. Ramser, M. Stadler, eds.), 1–35. Mohr Siebeck, Tübingen.

HÜBLER, O. (2005). Nichtlineare Paneldatenanalyse. Mimeo.

KEANE, M. (1994). A computationally practical simulation estimator for panel data. *Econometrica* **62** 95–116.

KÖNIG, A. (1997). Schätzen und Testen in semiparametrisch partiell linearen Modellen für die Paneldatenanalyse. University of Hannover, Diskussionspapier Nr. 208.

KÖNIG, A. (2002). *Nichtparametrische und semiparametrische Schätzverfahren für die Paneldatenanalyse.* Lit-Verlag, Münster.

KYRIAZIDOU, E. (1995). *Essays in Estimation and Testing of Econometric Models.* Dissertation, Evanston (Illinois).

KYRIAZIDOU, E. (1997). Estimation of a panel data sample selection model. *Econometrica* **65** 1335–1364.

LEE, M.-J. (1999). Nonparametric estimation and test for quadrant correlation in multivariate binary response models. *Econometric Reviews* **18** 387–415.

LI, Q., HSIAO, C. (1998). Testing serial correlation in semiparametric panel data models. *Journal of Econometrics* **87** 207–237.

LI, Q., STENGOS, T. (1996). Semiparametric estimation of partially linear panel data models. *Journal of Econometrics* **71** 389–397.

LI, Q., ULLAH, A. (1998). Estimating partially linear panel data models with one-way error components. *Econometric Reviews* **17** 145–166.

MANSKI, C. F. (1975). Maximum score estimation of the stochastic utility model of choice. *Journal of Econometrics* **3** 205–228.

MANSKI, C. F. (1987). Semiparametric analysis of random effects linear models from binary panel data. *Econometrica* **55** 357–362.

PAGAN, A., ULLAH, A. (1999). *Nonparametric Analysis.* Cambridge University Press, Cambridge.

REVELT, D., TRAIN, K. (1998). Mixed logit with repeated choices of appliance efficiency levels. *Review of Economics and Statistics* **80** 647–657.

ROBINSON, P. M. (1988). Root-N-consistent semiparametric regression. *Econometrica* **56** 931–954.

ULLAH, A., ROY, N. (1998). Nonparametric and semiparametric econometrics of panel data. In *Handbook of Applied Economics* (A. Ullah, D. E. A. Giles, eds.), 579–604. Marcel Dekker, New York.

WANDSBEEK, T. J., KAPTEYN, A. (1989). Estimation of the error components model with incomplete panels. *Journal of Econometrics* **41** 341–261.

WOOLDRIDGE, J. M. (2002). *Econometric analysis of cross section and panel data.* MIT Press, Cambridge.

10 Nonparametric Models and Their Estimation

Göran Kauermann[1]

[1] Faculty of Business Administration and Economics, University of Bielefeld
gkauermann@wiwi.uni-bielefeld.de

Summary: Nonparametric models have become more and more popular over the last two decades. One reason for their popularity is software availability, which easily allows to fit smooth but otherwise unspecified functions to data. A benefit of the models is that the functional shape of a regression function is not prespecified in advance, but determined by the data. Clearly this allows for more insight which can be interpreted on a substance matter level.

This paper gives an overview of available fitting routines, commonly called smoothing procedures. Moreover, a number of extensions to classical scatterplot smoothing are discussed, with examples supporting the advantages of the routines.

10.1 Introduction

Statistics and Econometrics have been dominated by linear or parametric models over decades. A major reason for this were the numerical possibilities which simply forbid to fit highly structured models with functional and dynamic components. These constraints have disappeared in the last 15 to 20 years with new computer technology occurring and with statistical software developed side by side with more flexible statistical models. In particular models with smooth and nonparametrically specified functions became rather popular in the last years and the models are now easily accessible and can be fitted without deep expert knowledge. A milestone for the propagation of the models with smooth components was the introduction to generalized additive models by Hastie and Tibshirani (1990). The presentation of the models was accompanied by its implementation in the software package Splus. In contrast to linear models, in additive models, a response variable y is modelled to depend additively on a number of covariates and in a smooth but otherwise unspecified manner. In this respect, additive models are a flexible way to estimate in a regression setting the influence of a number of covariates x, say, on a response or outcome variable y. Allowing the outcome variable to be non-normally distributed but distributed according to an exponential family (like binomial, Poisson etc.) leads to generalized additive models. The idea of allowing covariates to have nonparamet-

ric influence was extended to varying coefficient models in Hastie and Tibshirani (1993). Here, interaction effects of covariates, particularly between factorial and metrical quantities, are modelled functionally but nonparametrically. Further contributions were proposed including the wide class of semiparametric models. Here, parts of the covariate effects are modelled parametrically while others are included nonparametrically in the model.

Applications of non- or semiparametric models are versatile and found in nearly all scientific fields. Nonetheless, the classical econometric field is interestingly enough still somewhat dominated by classical parametric models and nonparametric models are not common standard. This is, however, changing rapidly now. Nonparametric ideas for time series data have been recently proposed in Fan and Yao (2003) (see also Härdle *et al.*, 1997). For financial data, Ruppert (2004) shows how nonparametric routines can be used to achieve insight beyond the parametric world. We also refer to Pagan and Ullah (1999), Härdle *et al.* (2004) or Akritas and Politis (2003) for further developments of nonparametric routines in econometrics. In this paper we add some further applications in the economic context, primarily though for demonstrational purpose.

The paper is organized as follows. First we give a sketch of the different scatterplot smoothing methods, like local smoothing, spline smoothing and the new proposal of penalized spline smoothing. In Section 3 we present different smoothing models. Data examples are provided as motivation why nonparametric models are worthwhile to be used. A discussion concludes the paper.

10.2 Scatterplot Smoothing

10.2.1 Sketch of Local Smoothing

An early starting point for smoothing was the local estimate formally proposed by Nadaraya (1964) and Watson (1964). The idea is to estimate a regression function $m(x)$, say, locally as weighted mean. Consider data (x_i, y_i), $i = 1, \ldots, n$, with x as (metrically scaled) covariate and y as response. We assume the regression model

$$y_i = m(x_i) + \varepsilon_i, \quad i = 1, \ldots, n \tag{10.1}$$

with ε_i as independent residuals and $m(\cdot)$ as unknown and not further specified regression function. Postulating smoothness for $m(\cdot)$, that is continuity and sufficient differentiability, one can estimate $m(x)$ at a target point x (within the support of x_i, $i = 1, \ldots, n$) by the locally weighted mean

$$\hat{m}(x) = \sum w_{xi} y_i \,. \tag{10.2}$$

Here w_{xi} are weights summing up to 1 mirroring the local estimation. This means that weights w_{xi} take large values if $|x_i - x|$ is small while w_{xi} gets smaller if $|x_i - x|$ increases. A convenient way to construct such weights is to make use of a so called kernel function $K(\cdot)$, where $K(\cdot)$ is typically chosen as positive, symmetric function

around zero. Convenient choices are the Gaussian shape kernel $K(u) = \exp(-u^2)$ or the Epanechnikov kernel $K(u) = (1-u^2)_+$ with $(u)_+ = u$ for $u > 0$ and 0 otherwise. The weights w_{xi} are then constructed through

$$w_{xi} = \frac{K\left(\frac{x_i - x}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{x_j - x}{h}\right)}$$

with h called the smoothing parameter or bandwidth, respectively. The role of h is to specify the amount of smoothness of the fit. If h is large, then neighbouring observations have weights of similar size and the resulting fit $\hat{m}(\cdot)$ is flat and smooth. In contrast, if h is small, then $\hat{m}(\cdot)$ will be wiggled.

Several practical procedures have been suggested to choose (or estimate) the smoothing parameter from the data at hand. An early discussion is found in Rice (1984). In principle, there are three approaches which have been suggested to choose bandwidth h. First, one can apply cross validation by leaving out one observation at a time and predicting its value with the smooth fit based on the remaining observations (see Stone, 1974). The resulting prediction error is known as cross validation. Let $\hat{m}_{-i,h}(x_i)$ be the estimate of $m(x_i)$ based on data points (x_l, y_l) with $l = 1, \ldots, i-1, i+1, \ldots, n$ and calculated with smoothing parameter h. We define

$$CV(h) = \sum_{i=1}^{n} \{y_i - \hat{m}_{-i,h}(x_i)\}^2 / n$$

as cross validation function. Since y_i and $\hat{m}_{-i,h}(x_i)$ are independent, $CV(h)$ is unbiased for the integrated mean squared error of $\hat{m}_h(x)$. Hence, minimizing $CV(h)$ with respect to h yields a plausible data driven choice of h. Defining with S_h the $n \times n$ smoothing matrix with entries

$$S_{h,ij} = \frac{K\left(\frac{x_i - x_j}{h}\right)}{\sum_{l=1}^{n} K\left(\frac{x_i - x_l}{h}\right)}$$

we obtain $\hat{m}_h(x_i) = \sum_{j=1}^{n} S_{h,ij} y_j$ and $m_{-i,h} = \sum_{j \neq i}^{n} S_{h,ij} y_j / (1 - S_{h,ii})$, which allows to calculate $CV(h)$ quite easily. A second possibility to choose h is using the Akaike (1970) criterion, which balances the goodness of fit with the complexity of the fitted model. This is expressed in the Akaike information function

$$AIC(h) = \log\left(\sum_{i=1}^{n} \{y_i - \hat{m}_h(x_i)\}^2\right) + 2df(h)/n \tag{10.3}$$

with $df(h) = tr\{S_h\}$ as measure for the degree of freedom, that is the complexity of the fitted function. The main idea of the Akaike criterion is, that the complexity

of the fit, contained in the last component in (10.3) is set as counterweight to the goodness of fit, exhibited in the first component in (10.3). The small sample behavior for $AIC(h)$ can be improved by replacing the latter component in (10.3) by the modified term $2\{df(h)+1\}/\{n-df(h)-2\}$ as suggested in Hurvich *et al.* (1998). Again, a suitable choice for h is obtained by minimizing $AIC(h)$. In the same style as the Akaike criterion generalized cross validation has been suggested by Gu and Wahba (1991). The idea is again to choose h by a compromise of goodness of fit and complexity of the fit. Finally, a third method to obtain a bandwidth estimate is to minimize the mean squared error analytically, and then use a plug-in estimate to obtain the optimal bandwidth estimate (see Härdle *et al.*, 1992). Regardless of the method being used, it can be shown theoretically and in simulations, that the convergence of the bandwidth estimate is slow (see also Härdle *et al.*, 1988). As a consequence, one should not blindly accept an automatically selected bandwidth but assess the smoothness of the resulting fit $\hat{m}(\cdot)$ by eye as well. In principle this means, one should play with different bandwidths around the data driven optimal one to validate the sensitivity of the fit on the bandwidth choice. This approach has been developed more formally in Marron and Chaudhuri (1999).

The breakthrough of the local estimation approach came with a simple but practical extension. Instead of estimating locally a constant one can fit locally a polynomial model (see Fan and Gijbels, 1996). This idea proved to behave superior, in particular at boundaries of the support of x, if local linear smoothing is used. Moreover, local linear (or polynomial) smoothing allows not only to estimate the regression function itself but also its derivatives. This results simply as by product of the estimation routine. If locally a linear model is fitted then the fitted local slope serves as estimate for the smooth estimation of the first derivative. Local polynomial smoothing techniques have been en vogue in the late nineties, but have been less focussed the last years. This is, at least partly, due to the numerical hurdles one easily faces with this approach. The local idea says that one locally fits a polynomial model to data. This means, in order to visualize the functional shape of a regression function $m(\cdot)$ one has to estimate $m(x)$ at a number of points x and then connect the resulting estimates. If the local fit is complex and numerically intensive, then local estimation at a number of points can readily lead to the borderline of numerical feasibility. Nonetheless, local estimation is simple in its structure which still justifies the approach. For more information and a general introduction to the ideas of local smoothing we refer to Simonoff (1996) or Loader (1999).

10.2.2 Sketch of Spline Smoothing

Parallel to the kernel and local polynomial smoothing, spline smoothing has been on the market for a while. As standard references we cite here Eubank (1988) and Wahba (1990). These books pursue a more mathematical viewpoint, while Hastie and Tibshirani (1990) or Green and Silverman (1994) present a more practical guideline. The idea leading to spline estimates is that function $m(x)$ in (10.1) is estimated through the penalized criterion

$$\min_{m(\cdot)} \left\{ \sum_{i=1}^{n} \{y_i - m(x_i)\}^2 - h \int \left\{ m''(x) \right\}^2 dx \right\}. \tag{10.4}$$

The main idea behind (10.4) is that the goodness of fit, measured with the first component, is penalized by the complexity of the function, measured by the second order derivative. It is due to Reinsch (1967) who showed that $m = (m(x_1), \ldots, m(x_n))$ as minimizer of (10.4) can be written as $m = C\alpha$, with C as cubic spline basis and α as spline coefficient. Then, minimizing (10.4) is equivalent to minimizing

$$\min_{\alpha} \left\{ (Y - C\alpha)^T (Y - C\alpha) + h\alpha^T D\alpha \right\}, \tag{10.5}$$

where $Y = (y_1, \ldots, y_n)$ and D is a penalty matrix resulting by mathematical theory (see Fahrmeir and Tutz, 2001, for more details). This form easily allows for numerical implementation and like for local smoothing, coefficient h plays the role of the smoothing parameter, also called penalty parameter in this context. The parameter h penalizes the complexity of the function, with $h \to \infty$ leading to a linear fit, since for linear functions $m''(x) = 0$. On the other hand, for $h \to 0$ the fitted curve becomes wiggled and interpolating. A data driven choice of h is available with the same means as for kernel smoothing, that is with cross validation or an Akaike criterion. Additionally, one can comprehend (10.5) as component of a log likelihood in a mixed model with *a priori* normal distribution on coefficient α. In this case, the smoothing parameter h plays the role of a variance ratio which can be estimated within a likelihood framework (see e.g. Efron, 2001, for more details). For spline smoothing a number of numerical adjustments have been suggested. The necessity for this is due to the fact that the cubic spline basis C grows with the same order as the sample size so that for large samples the spline estimate would require the inversion of an $n \times n$ matrix, with n as sample size. This numerical hurdle can be circumvented by reducing the dimension of the basis to achieve numerical feasibility, which has led to the so called pseudo splines (Hastie, 1996, see also Wood, 2003).

10.2.3 Sketch of Penalized Spline (P-Spline) Smoothing

A powerful modification of spline smoothing is available by reformulating (10.5) in the following way. For estimation one sets $m = B\alpha$ with B as high dimensional basis with fixed number of basis functions. For instance one may construct B as B-Spline basis (see de Boor, 1978) with generously chosen number of basis functions (30-60). Unlike spline smoothing, the dimension of the basis is now fixed and does not grow with the sample size n. For fitting, the spline coefficients are penalized like in (10.5) with appropriately chosen penalty matrix D leading to the criterion

$$\min_{\alpha} \left\{ (Y - B\alpha)^T (Y - B\alpha) + h\alpha^T D\alpha \right\}.$$

The approach is nowadays known under the phrase penalized spline smoothing, or shortly P-spline smoothing. Originally introduced by O'Sullivan (1986) the method gained attention with Eilers and Marx (1996) and the recent book by Ruppert *et al.* (2003). Even though the approach is similar to spline smoothing, it is different in so far, that the basis and its fixed dimension is chosen in advance. The benefits

achieved therewith are attractive simple numerical performance as well as practical advantages in smoothing parameter selection by exploiting a methodological link to linear mixed models (see Wand, 2003). In fact, mixed model software can be used to fit smooth nonparametric models and the delicate issue of selecting an appropriate bandwidth is sourced out, since in the linear mixed model the smoothing parameter plays the role of a variance component. This means, maximum likelihood theory can be used for smoothing parameter selection (see Kauermann, 2004).

10.2.4 Software for Smoothing

The parallel development of sophisticated statistical models and available software which allows to fit them to the applicants data brought a wide acceptance of the new statistical technology. Software availability nowadays makes it easy to apply nonparametric smooth models to real data. The leading products are here the public domain package R (see www.r-project.org or Dalgaard (2002) for an introduction) or the commercial origin Splus (see Ripley and Venables, 2002). Implementations of innovative estimation procedures are also available in XploRe (see Härdle *et al.*, 2000).

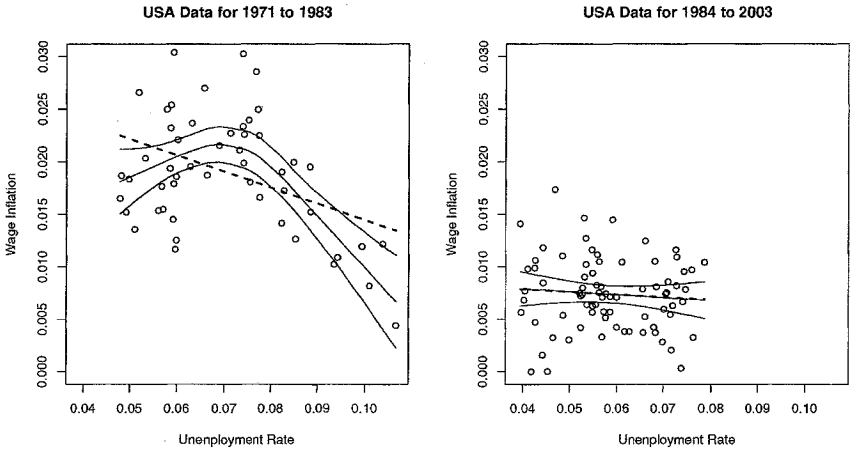


Figure 10.1: *Unemployment rate and wage inflation for USA for two different time intervals. Smooth estimate is plotted with confidence interval, dashed line gives the standard least squares estimate in a linear model.*

10.2.5 Example (Scatterplot Smoothing)

Exemplary for a smoothing model of type (10.1), we consider a simple wage Phillips curve (see for instance Chiarella and Flaschel, 2000) with x as unemployment rate and y as wage inflation. The data shown in Figure 10.2.4 are for the USA for the years 1971 to 1983 quarterly and 1984 to 2003, respectively. Note that the cut of the data at year 1983 is primarily made for presentation reasons and not necessarily

based on economic theory. The traditional paper by Phillips (1958) discusses data for United Kingdom where he advocates a convex shape of the functional relationship between wage inflation and unemployment rate. It is clear that wages do not exclusively depend on the unemployment rate but on further economic quantities and since its original publication, the relationship, now known as Phillips curve, has been extended in numerous ways. Moreover, a discussion about the shape has been focussed by a large number of authors (see Flaschel *et al.*, 2005, for details and references). It is beyond the scope of this paper to contribute to this discussion. Instead we look at the traditional curve to demonstrate non-linearity, in particular concavity for the first time period. This can be seen in Figure 10.2.4 where we plot a local estimate, with smoothing parameter selected by the Akaike information criterion. The estimate is shown as smooth curve with pointwise confidence bands (95% level) included. For comparison we also include a parametric fit by setting

$$m(x) = \beta_0 + x\beta_x \,. \tag{10.6}$$

This is shown as dashed line. Apparently, there is evidence for a concave behavior, that is, that the wage pressure is high but somewhat stable for small up to medium unemployment. If, however, the unemployment rate is large, the pressure on wage decreases with increasing unemployment rate. The picture looks different for data from the mid 80th onwards. Here, the Akaike criterion choses a bandwidth such that the local estimate and the parametric linear fit (10.6) coincide ($\hat{\beta}_0 = 0.008$ ($std = 0.002$), $\hat{\beta}_x = -0.025$ ($std = 0.037$)). Now the pressure on wage depends only slightly on the unemployment rate on a low level.

10.3 Non and Semiparametric Models

10.3.1 Generalized Additive and Varying Coefficient Models

The simple scatterplot smoothing model (10.1) can be extended in various ways. First, the assumption of normality can be generalized. For parametric models this has led to generalized linear models (see McCullagh and Nelder, 1989). The idea is to generalize the linear regression model (10.6) by allowing y for given x to be distributed according to an exponential family distribution with mean structure

$$E(y|x) = g(\beta_0 + x\beta_x) \,, \tag{10.7}$$

where $g(\cdot)$ is a known link function. Clearly (10.7) is a parametric model. If we replace the linear relationship by a nonparametric function we obtain the generalized nonparametric model

$$E(y|x) = g(m(x)) \tag{10.8}$$

with $m(\cdot)$ as an unknown but smooth function in x.

The next extension is achieved by allowing y to depend on more than one (metrically scaled) covariate, e. g. x and u. Modelling the influence of x and u nonparametrically leads to the *Generalized Additive Model* (GAM)

$$E(y|x,u) = g\{m_x(x) + m_u(u)\}\,, \tag{10.9}$$

where $m_x(\cdot)$ and $m_u(\cdot)$ are unknown but smooth functions to be estimated from the data. This model has achieved tremendous popularity with the book by Hastie and Tibshirani (1990) and the Splus software allowing to fit it. The advantage of (10.9) is that estimation is easy by using a backfitting algorithm, i. e. we keep all smooth components except of one as fixed and estimate the remaining one. Circulating over the functions provides the corresponding fit. The benefit of additive modelling is that the so called *curse of dimensionality* is avoided. The latter occurs if high dimensional functions are fitted from data and it says that the required sample size grows exponentially with the dimension of the function fitted. This is avoided if additivity is assumed.

So far we have only considered models with purely metrically scaled covariates. The combination of metrical scale and nominal covariates (like indicator variables) can be modelled flexibly with a so called *Varying Coefficient Model* (Hastie and Tibshirani, 1993). For this purpose, let z be a nominal covariate (i. e. a factor) and x is as before metrical. A nonparametric model is then

$$E(y|x,z) = g\{m_0(x) + z m_z(x)\}\,. \tag{10.10}$$

Now $m_0(x)$ is the influence of x while $m_z(x)$ mirrors the multiplicative smooth interaction effect of x and z. If the covariates effect of z does not interact with x we get $m_z(x) = \beta_z$ as constant function and model (10.10) simplifies to

$$E(y|x,z) = g\{m_0(x) + z\beta_z\}\,. \tag{10.11}$$

This model is parametric and nonparametric at the same time and it is typically referred to as *Semiparametric Model* or *Partly Linear Model* in the literature. From a statistical viewpoint it has been of interest to find efficient ways to fit models (10.9), (10.10) and (10.11), respectively (see e.g. Wood, 2003, Kauermann and Tutz, 2000).

10.3.2 Example (Generalized Additive Models)

As illustrative example we look at data taken from the Munich Founder Study (see Kauermann *et al.*, 2005). We consider data of 1235 firms founded between the years 1985 to 86 in Munich metropolitan area. In particular, we focus on the three year success rate by defining the outcome variable $y_i = 1$ for the i-th firm, if the firm went out of business within its first three years of business while $y_i = 0$ otherwise. As explanatory variables we include $x = age$ of the founder as metrical quantity and the following binary factors coded 0 or 1:

- $z_1 = plan = 1$ if business was planned $> 1/2$ year

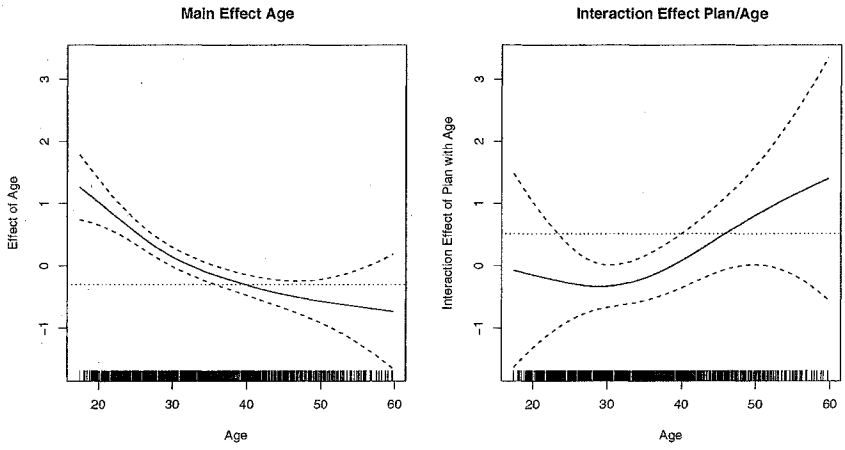


Figure 10.2: *Main effect of age on the three year success rate of newly founded enterprises (left plot) and the interactive effect of plan. Constant effects are shown as horizontal line.*

Table 10.1: *Parametric estimates for Munich founder study.*

variable	estimate	std dev	p-value
(intercept)	0.305	0.174	0.080
plan	-0.512	0.218	0.019
test	-1.047	0.412	0.011
sex	-0.087	0.161	0.587
branch	-0.832	0.147	< 0.001
specialist	-0.338	0.148	0.022
innovation	-0.483	0.156	0.002
school	-0.460	0.155	0.003

- $z_2 = test = 1$ if business had a test phase of $> 1/2$ year
- $z_3 = sex = 1$ for male founder
- $z_4 = branch = 1$ for branch knowledge of founder
- $z_5 = specialist = 1$ for firm aiming for specialized market
- $z_6 = innovation = 1$ for new product produced or sold by the firm
- $z_7 = school = 1$ for high school degree.

For each of the factorial covariates we checked multiplicative interaction with age applying a forward selection. For covariate *plan* the varying coefficient model looks for instance like

$$E(y|x, z_1, \ldots, z_7) = h\{m_0(x) + z_1 m_1(x) - z_2\beta_2 + \ldots + z_7\beta_7\}\,, \quad (10.12)$$

i. e. *plan* and *age* interact. This is also the final model, since no other covariate effects were found to interact with *age* x based on the resulting selected smoothing parameter. The parametric estimates are provided in Table 10.1. For estimation we decompose $m_0(x)$ and $m_1(x)$ to $\beta_0 + m_{00}(x)$ and $\beta_1 + m_{01}(x)$, respectively, with identifiability constraint $\int m_{0l}(x)dx = 0$ for $l = 0, 1$. This means that β_1 is the average plan effect while $m_{01}(x)$ mirrors the orthogonal interaction with age. As can be read from Table 10.1, significant quantities are whether the enterprise was planned and tested, whether the entrepreneur had branch knowledge and whether he or she was aiming for a specialized market or bringing an innovative product to the market. Moreover, school education has a significant effect, all of the effects are risk reducing. The effect of *sex* clearly is not significant. Next we study the fitted interaction effects $m_{0l}(x)$, $l = 0, 1$. As can be seen, elderly entrepreneurs have less risk of failing with their company, while the risk increases more or less linearly for younger founders. If the firm was planned for longer than $1/2$ year, the risk is reduced. This holds particularly for younger founders, as the effect interacts with age. In fact, the smoothing parameter chosen by cross validation indicates a nonparametric interaction effect. Smoothing parameters are chosen by generalized cross validation using the implemented version in the gam$(\cdot)$ procedure in R. Note that $\hat{m}_l(x) = \hat{\beta}_l + \hat{m}_{0l}(x)$, so that $\hat{m}_l(x) = 0$ holds if $-\hat{\beta}_l = \hat{m}_{0l}(x)$. To indicate the area of x-values where the effect vanishes we include $-\hat{\beta}_l$ as dashed horizontal line in the plot. It appears that the risk reducing effect of *plan* is particularly active for entrepreneurs aged 35 to 40 and younger.

10.3.3 Further Models

A potentially useful class of models to include nonparametric effects are duration time models. Here, the Cox (1972) model is playing the dominant role by setting the hazard function $\lambda(t, x)$ for covariates x as

$$\lambda(t, x) = \lambda_0(t) \exp(x\beta)\,, \quad (10.13)$$

where $\lambda_0(t)$ is known as baseline hazard. In particular, in (10.13) one assumes that covariate effects are constant over time. This can be doubtful, in particular for

economic datasets, where duration time can be long and can even exceed years. In this case a more realistic scenario is to allow the covariate effects to change with time. This leads to the model

$$\lambda(t, x) = \lambda_0(t) \exp(xm(t)) \tag{10.14}$$

with $m(t)$ as covariate effect varying smoothly in t. Models of this kind are presented in Grambsch and Therneau (2000), an economic example is found for instance in Kauermann *et al.* (2005).

A second direction to extend nonparametric models is to allow for correlated residuals in a time series framework (Fan and Yao, 2003). Generally, smoothing with correlated errors is more cumbersome as demonstrated in Opsomer *et al.* (2001). The major problem here is that standard smoothing parameter selection fails.

10.3.4 Multivariate and Spatial Smoothing

So far we have discussed models where the smooth function has a univariate argument. This can be extended by assuming that $x = (x_1, \ldots, x_p)$ is multivariate, so that model (10.8) becomes

$$\mathrm{E}(y|x_1, \ldots, x_p) = g(m(x_i, \ldots, x_p)) \,. \tag{10.15}$$

Now $m(\cdot)$ is an unspecified function, smooth in all components. Models of this type are hard to estimate if p is large, which is known under the phrase *curse of dimensionality* (see Hastie and Tibshirani, 1990). The reason for this problem is, that the amount of information necessary to obtain a fit with postulated bounds of error increases exponentially with p. On top of that, the function is nearly impossible to be visualized if $p > 2$. Therefore, multivariate smoothing is usually restricted to bivariate or spatial smoothing where one takes $p = 2$. The multivariate aspect thereby relates to spatial smoothing if x_1 and x_2 are location coordinates, but in principle, x_1 and x_2 can be any continuous measure as the subsequent example shows. Fitting the model is carried out in the same fashion as for univariate smoothing with some minor modifications. In particular, for spatial smoothing kriging is more familiar than spline smoothing, even though the ideas are related (see Nychka, 2000, or Ruppert *et al.*, 2003, Chapter 13).

10.3.5 Example (Bivariate Smoothing)

Exemplary for a bivariate smooth we consider data taken from the German Socio Economic Panel. We consider $n = 4501$ individuals aged 18 to 60 who were domiciled in West Germany and who became unemployed during 1983 and 2001. We model the probability π of returning to professional life within 12 months of unemployment. For individuals in the panel with more than one spell of unemployment we randomly select one of the spells which guarantees independence among our observations. As covariates we consider *age*, *gender* and *nationality*. The model fitted is

$$\mathrm{logit}(\pi) = m(age, start) + nationality\beta_n + gender\beta_g \,,$$

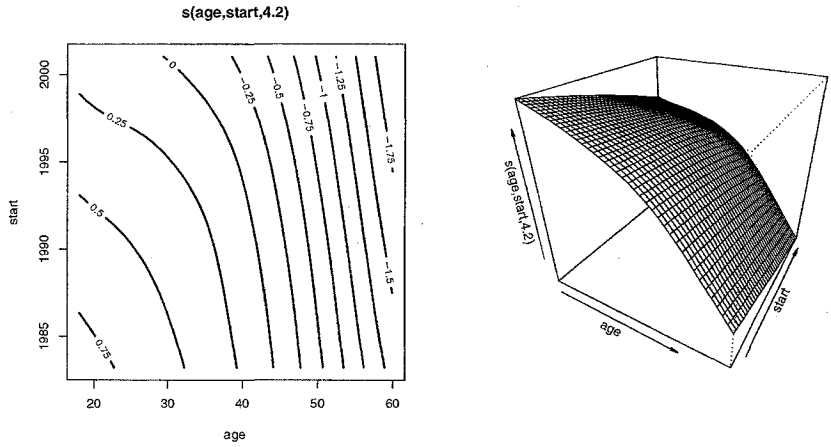


Figure 10.3: *Interactive smooth effect of age and start as calendar time plotted as contour and perspective plot. Estimated degree of freedom is 4.2.*

where *start* is starting time of the unemployment. Figure 10.3.4 displays the fitted curve $m(\cdot,\cdot)$ centered around zero by subtracting the fitted intercept -1.28. As can be seen, the chances of finding work within a year decreases with age in a nonlinear form. Moreover, the chances decrease with calendar time, where younger individuals experienced a stronger descent compared to older individuals. The parametric effects show that Germans have higher chances to find a job within a year ($\beta_n = 0.24$ ($sd = 0.07$) for *nationality* = 1 for Germans, 0 otherwise) and males have higher chances ($\beta_g = 1.03$ ($sd = 0.07$) for *gender* = 1 for males, 0 for females).

10.3.6 Model Diagnostics

Smoothing offers an insightful option for model diagnostics and model checking. In its standard form this boils down to testing the parametric model (10.7) against its smooth generalization (10.8). Due to the unconstrained structure of the alternative model (10.8), a test constructed via smoothing is likely to have power against a wide range of alternatives. The principle idea is to compare the difference in the fit in the two models using an appropriate reference distribution. The latter is non standard, which makes smooth tests a little delicate. A number of authors have contributed to this field proposing different tests based on smoothing (see Härdle and Mammen, 1993, or Kauermann and Tutz, 2001, and references given there). Recently, Crainiceanu *et al.* (2005) propose a likelihood ratio test for penalized spline smoothing showing that the reference distribution for the test statistics collapses to a mixture of χ^2 distributions. Their result can also be applied in smoothing.

Of more exploratory style are the local smoothing ideas published in Bowman and Azzalini (1997).

10.4 Discussion

In this paper we gave a brief overview about smoothing techniques and models with smooth components. We demonstrated the generality of the approach and provided examples as illustration. It should be pointed out that all examples were fitted with R (www.r-project.org) so that the results can not only be easily reproduced, moreover, the reader is invited to "smooth" his or her data as well.

References

AKAIKE, H. (1970). Statistical predictor identification. *Annals of the Institute of Statistical Mathematics* **22** 203–217.

AKRITAS, M., POLITIS, D. (2003). *Recent Advances and Trends in Nonparametric Statistics*. North Holland, Amsterdam.

BOWMAN, A. W., AZZALINI, A. (1997). *Applied Smoothing Techniques for Data Analysis: the Kernel Approach with S-Plus Illustrations*. Oxford University Press, Oxford.

CHIARELLA, C., FLASCHEL, P. (2000). *The Dynamics of Keynesian Monetary Growth: Macro Foundations*. Cambridge University Press, Cambridge.

COX, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Association, Series B* **34** 187–220.

CRAINICEANU, C., RUPPERT, D., CLAESKENS, G., WAND, M. (2005). Exact likelihood ratio test for penalized splines. *Biometrika* **92** 91–103.

DALGAARD, P. (2002). *Introductory Statistics with R*. Springer, New York.

DE BOOR (1978). *A Practical Guide to Splines*. Springer, Berlin.

EFRON, B. (2001). Selection criteria for scatterplot smoothers. *The Annals of Statistics* **29** 470–504.

EILERS, P., MARX, B. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science* **11** 89–121.

EUBANK, R. L. (1988). *Spline Smoothing and Nonparametric Regression*. Dekker, New York.

FAHRMEIR, L., TUTZ, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models*. 2nd ed., Springer, New York.

FAN, J., GIJBELS, I. (1996). *Local Polynomial Modelling and its Applications*. Chapman and Hall, London.

FAN, J., YAO, O. (2003). *Nonlinear Time Series: Nonparametric and Parametric Methods.* Springer, New York.

FLASCHEL, P., KAUERMANN, G., SEMMLER, W. (2005). Testing wage and price Phillips curves for the United States. *Metroeconomica* (to appear).

GRAMBSCH, P. M., THERNEAU, T. M. (2000). *Modelling Survival Data: Extending the Cox Model.* Springer, New York.

GREEN, D. J., SILVERMAN, B. W. (1994). *Nonparametric Regression and Generalized Linear Models.* Chapman and Hall, London.

GU, C., WAHBA, G. (1991). Smoothing spline ANOVA with component-wise Bayesian confidence intervals. *Journal of Computational and Graphical Statistics* **2** 97–117.

HÄRDLE, W., HALL, W., MARRON, J. S. ((). 1. *9* **8** 8.). How far are automatically chosen regression smoothing parameter selectors from their optimum?Journal of the American Statistical Association 83 86–101

HÄRDLE, W., HALL, W., MARRON, J. S. (1992). Regression smoothing parameters that are not far from their optimum. *Journal of the American Statistical Association* **87** 227–233.

HÄRDLE, W., MAMMEN, E. (1993). Comparing nonparametric versus parametric regression fits. *The Annals of Statistics* **21** 1926–1947.

HÄRDLE, W., LÜTKEPOHL, H., CHEN, R. (1997). A review of nonparametric time series analysis. *International Statistical Review* **65** 49–72.

HÄRDLE, W., HLAVKA, Z., KLINKE, S. (2000). *XploRe, Application Guide.* Springer, Berlin.

HÄRDLE, W., MÜLLER, M., SPERLICH, S., WERWATZ, A. (2004). *Nonparametric and Semiparametric Models.* Springer, Berlin.

HASTIE, T. (1996). Pseudosplines. *Journal of the Royal Statistical Society, Series B* **58** 379–396.

HASTIE, T., TIBSHIRANI, R. (1990). *Generalized Additive Models.* Chapman and Hall, London.

HASTIE, T., TIBSHIRANI, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society, Series B* **55** 757–796.

HURVICH, C. M., SIMONOFF, J. S., TSAI, C. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society, Series B* **60** 271–293.

KAUERMANN, G. (2000). Modelling longitudinal data with ordinal response by varying coefficients. *Biometrics* **56** 692–698.

KAUERMANN, G. (2004). A note on smoothing parameter selection for penalized spline smoothing. *Journal of Statistical Planing and Inference* **127** 53–69.

KAUERMANN, G. (2005). Penalized spline fitting in multivariable survival models with varying coefficients. *Computational Statistics and Data Analysis* **49** 169–186.

KAUERMANN, G., TUTZ, G. (2000). Local likelihood estimation in varying-coefficient models including additive bias correction. *Journal of Nonparametric Statistics* **12** 343–371.

KAUERMANN, G., TUTZ, G. (2001). Testing generalized linear and semiparametric models against smooth alternatives. *Journal of the Royal Statistical Society, Series B* **63** 147–166.

KAUERMANN, G., TUTZ, G., BRÜDERL, J. (2005). The survival of newly founded firms: A case study into varying-coefficient models. *Journal of the Royal Statistical Society, Series A* **168** 145–158.

LOADER, C. (1999). *Local Regression and Likelihood.* Springer, Berlin.

MARRON, J. S., CHAUDHURI, P. (1999). SiZer for exploration of structures in curves. *Journal of the American Statistical Association* **94** 807–823.

MCCULLAGH, P., NELDER, J. A. (1989). *Generalized Linear Models.* 2nd ed., Chapman and Hall, New York.

NADARAYA, E. A. (1964). On estimating regression. *Theory of Probability and Application* **9** 141–142.

NYCHKA, D. (2000). Spatial process estimates as smoothers. In *Smoothing and Regression. Approaches, Computation and Application* (Schimek, ed.), Wiley, New York.

O'SULLIVAN, F. (1986). A statistical perspective on ill-posed inverse problems. *Statistical Science* **1** 502–518.

OPSOMER, J. D., WANG, Y., YANG, Y. (2001). Nonparametric regression with correlated errors. *Statistical Science* **16** 134-153.

PAGAN, R., ULLAH, A. (1999). *Nonparametric Econometrics.* Cambridge University Press, Cambridge.

PHILLIPS, A. W. (1958). The relation between unemployment and the rate of change of money wage rates in the United Kingdom, 1861-1957. *Economica* **25** 283–299.

REINSCH, C. H. (1967). Smoothing by spline functions. *Numerical Mathematics* **10** 177–183.

RICE, J. A. (1984). Bandwidth choice for nonparametric regression. *Annals of Statistics* **12** 1215–1230.

RIPLEY, B. D., VENABLES, W. N. (2002). *Modern Applied Statistics with S.* 4th ed., Springer, New York.

RUPPERT, D. (2004). *Statistics and Finance.* Springer, New York.

RUPPERT, R., WAND, M. P., CARROLL, R. J. (2003). *Semiparametric Regression.* Cambridge University Press, Cambridge.

SIMONOFF, J. S. (1996). *Smoothing Methods in Statistics.* Springer, New York.

STONE, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, Series B* **36** 111–147.

WAND, M. P. (2003). Smoothing and mixed models. *Computational Statistics* **18** 223-249.

WAHBA, G. (1990). Regularization and cross validation methods for nonlinear implicit, ill-posed inverse problems.. In *Geophysical Data Inversion Methods and Applications* (A. Vogel, C. Ofoegbu, R. Gorenflo and B. Ursin, eds.), 3-13. Vieweg, Wiesbaden-Braunschweig.

WATSON, G. S. (1964). Smooth regression analysis. *Sankhyā, Series A* **26** 359–372.

WOOD, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society, Series B* **65** 95–114.

11 Microeconometric Models and Anonymized Micro Data *

Gerd Ronning[1]

[1] Wirtschaftswissenschaftliche Fakultät, Universität Tübingen
gerd.ronning@uni-tuebingen.de

Summary: The paper first provides a short review of the most common microeconometric models including logit, probit, discrete choice, duration models, models for count data and Tobit-type models. In the second part we consider the situation that the micro data have undergone some anonymization procedure which has become an important issue since otherwise confidentiality would not be guaranteed. We shortly describe the most important approaches for data protection which also can be seen as creating errors of measurement by purpose. We also consider the possibility of correcting the estimation procedure while taking into account the anonymization procedure. We illustrate this for the case of binary data which are anonymized by 'post-randomization' and which are used in a probit model. We show the effect of 'naive' estimation, i. e. when disregarding the anonymization procedure. We also show that a 'corrected' estimate is available which is satisfactory in statistical terms. This is also true if parameters of the anonymization procedure have to be estimated, too.

11.1 Introduction

Empirical research in economics has for a long time suffered from the unavailability of individual 'micro' data and has forced econometricians to use (aggregate) time series data in order to estimate, for example, a consumption function. On the contrary other disciplines like psychology, sociology and, last not least, biometry have analyzed micro data already for decades. Therefore it is not surprising that most of the by now well-known microeconometric methods have been invented long time ago by biometricians and psychometricians. However, it is the merit of econometricians

*Research in this paper is related to the project "Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten" financed by German Ministry of Research and Technology.

that they have provided the underlying behavioral or structural model. For example, the probit model can be seen as an operational version of a linear model which explains the latent dependent variable describing, say, the unobservable reservation wage. Moreover, the discrete choice model results from the hypothesis that choice among alternatives is steered by maximization of utility of these alternatives.

The software for microeconometric models has created growing demand for micro data in economic research, in particular data describing firm behavior. However, such data are not easily available when collected by the Statistical Office because of confidentiality. On the other hand these data would be very useful for testing microeconomic models. This has been pointed out recently by KVI commission.[1] Therefore, the German Statistical Office initiated research on the question whether it is possible to produce scientific use files from these data which have to be anonymized in a way that re-identification is almost impossible and, at the same time, distributional properties of the data do not change too much. Published work on anonymization procedures and its effects on the estimation of microeconometric models has concentrated on *continuous* variables where a variety of procedures is available. See, for example, Ronning and Gnoss (2003) for such procedures and the contribution by Lechner and Pohlmeier (2003) also for the effects on estimation. Discrete variables, however, mostly have been left aside in this discussion. The only stochastic-based procedure to anonymize discrete variables is post-randomization (PRAM) which switches categories with prescribed probability. In this paper we consider anonymization by PRAM and its effect on the estimation of the microeconometric probit model. Thus, we consider an anonymized binary variable which is used as dependent variable in a probit model whereas the explanatory variables remain in their original form.

In Section 11.2 we describe the most important microeconometric models which by now should be well known so that details on estimation and testing are omitted and only principles of modelling are sketched. Section 11.3 presents anonymization procedures and possible strategies to incorporate them into the estimation of microeconometric models. Finally Section 11.4 will illustrate these general remarks for the special case that the binary dependent variable in a probit model has been anonymized by PRAM. We also consider the case that the user is not informed about details of the anonymization procedure.

11.2 Principles of Microeconometric Modelling

Consider the following linear model:

$$Y^* = \alpha + \beta x + \varepsilon \tag{11.1}$$

with $E[\varepsilon] = 0$ and $V[\varepsilon] = \sigma_\varepsilon^2$. Here the $*$ indicates that the continuous variable Y is latent or unobservable. This model asserts that the conditional expectation of Y^* but not the corresponding conditional variance depends on x. If the dependent

[1] See Kommission zur Verbesserung der statistischen Infrastruktur (2001).

variable Y^* is observable, that is if $Y = Y^*$, then we have the standard simple regression model with just one regressor and unknown parameters α, β and σ_ε^2. For example, Y may be expenditure for a certain good and x may be some indicator of quality of the good. It is reasonable to assume a priori that $\beta > 0$ holds.

11.2.1 Binary Probit (and Logit) Model

Now assume that it is only known whether the good has been purchased or not. If we interpret Y^* as the utility of the good and assume that purchase will be made if utility crosses a certain threshold τ, we get formally

$$Y = \begin{cases} 0 \text{ if } Y^* \leq \tau \\ 1 \text{ else} \end{cases} \tag{11.2}$$

for the observed variable Y. It can be shown that two of the four parameters α, β σ_ε^2 and τ have to be fixed in order to attain identification of the two remaining ones. Usually we set $\tau = 0$ and $\sigma_\varepsilon^2 = 1$ assuming additionally that the error term ε is normally distributed. This is the famous probit model. Note that only the probability of observing $Y = 1$ for a given x can be determined. If we alternatively assume that the error term follows a logistic distribution, we obtain the closely related binary logit model.

11.2.2 Ordinal Probit Model

Some minor modification leads then to the ordinal probit model. Assume that instead of the binary indicator a trichotomous (and ordered) indicator is observed which is assumed to be generated from latent variable Y^* as follows:

$$Y = \begin{cases} 0 \text{ (no purchase)} & \text{if } Y^* \leq \tau_1 \\ 1 \text{ (one unit of the good)} & \text{if } \tau_1 < Y^* \leq \tau_2 \\ 2 \text{ (more than one unit of the good)} & \text{else.} \end{cases}$$

This model has two thresholds instead of only one. Note that the probability is not a monotonic function of x with regard to 'middle' alternatives which is a well-known problem when interpreting estimation results.

11.2.3 Discrete Choice Model

A different situation is met if the indicator variable contains only information from a nominal scale. For example, it may be observed which brand of a certain good is purchased. In this case we get $Y \in \{A, B, C\}$ where A, B and C denote three different brands. Furthermore assume that x_j is the quality indicator for brand j and the linear model now is written as

$$U_j = \alpha + \beta x_j + \varepsilon_j \,, \; j = 1, \ldots, r \,,$$

where U_j denotes utility of brand j and r the number of alternatives. Note that we now have r error terms. Under the random utility maximization hypothesis

we assume that the consumer chooses the alternative which offers maximal utility. However, since U_j is a random variable, we can only obtain the probability of observing alternative j. More formally we assume that

$$P[Y = j|x_1, \ldots, x_r] = P[U_j \geq U_k\,,\, k \neq j]$$

where $P[A]$ denotes the probability of event A. If we assume that the random variables U_j are jointly normally distributed we have the multinomial probit model. If one alternatively assumes the otherwise seldom used extreme value distribution (Gumbel distribution), then we arrive at the multinomial logit model first introduced by Daniel McFadden which is much easier to estimate than the probit version for $r > 3$. However, the latter mentioned model can assume a flexible correlation structure for the error terms mimicking the similarity of pairs of brands. Both distributional versions are known as 'discrete choice model'.

11.2.4 Count Data Models

Consider again the situation described for the ordinal probit model where Y contains information from an ordinal scale. If we switch to the situation where the number of units of the good purchased is known, then we have $Y \in \{0, 1, 2, 3, \ldots\}$, that is, Y is a nonnegative integer. The easiest but also most restrictive model describing such a data set is the Poisson distribution. Note that no longer an underlying latent variable is assumed. However, again we want to estimate the impact of the quality of the good (denoted by x) on the number of units purchased, that is we assume that the conditional expectation of Y given x depends on x. Clearly a distribution with only nonnegative realizations will have a positive expectation. In particular for the Poisson distribution we obtain $E[Y] = V[Y] = \lambda > 0$ where λ is the only parameter of this distribution. Note that the first equation often is termed 'equidispersion'. For the conditional model we use

$$E[Y|x] = \lambda(x) = \exp(\alpha + \beta x),$$

which relates the single parameter λ to the explanatory variable x in a way that the expected value is nonnegative for any x. The resulting model is called the Poisson (regression) model. If the slightly more flexible negative binomial distribution is used, the often observed 'overdispersion' ($V[Y|x] > E[Y|x]$) can be handled more adequately.

11.2.5 Duration Models

Many if not most economic variables are nonnegative. Therefore the probability mass of the corresponding distribution should be restricted to the $\mathbb{R}^+$. However only in duration analysis this aspect has been recognized properly.[2] For example, the duration or 'spell' of unemployment may be described by one of the following distributions: gamma with exponential as a special case, lognormal or Weibull.[3]

[2] The only other model where a nonnegative distribution has been employed is the Gamma distribution when estimating frontier functions. See, for example, Greene (2000, Chapter 9.7).

[3] See, for example, Appendix A in Ronning (1991) for a description of these distributions.

It turns out that Weibull is easiest to handle since its distribution function has a closed form:

$$F(y) = 1 - \exp\left(-\kappa y^{\theta}\right), \quad \kappa, \theta > 0. \tag{11.3}$$

A conditional distribution describing, for example, the impact of age on the unemployment spell can be derived by letting one of the two parameters of this distribution depend on x. For example, we may assume[4]

$$\kappa(x) = \exp(\alpha + \beta x)$$

and then estimate the unknown parameters α, β and θ from this conditional distribution.

However, usually the model is formulated in terms of the hazard rate which is given by

$$\lambda(y) = \frac{f(y)}{1 - F(y)} \tag{11.4}$$

and is also termed 'survivor function' since it describes the 'probability' that an event of duration y will last longer than y. Now it is evident that the Weibull distribution is an attractive candidate since from inserting the distribution function (11.3) and the corresponding density function into (11.4) we obtain the simple expression

$$\lambda(y) = \kappa\theta y^{\theta-1}$$

and for the conditional hazard rate we obtain

$$\lambda(y|x) = \exp(\alpha + \beta x)\theta y^{\theta-1}.$$

Duration analysis faces the special problem of censoring since the true value of a duration or spell may be unknown for two reasons: (i) the start of the spell of unemployment may not be recognized; (ii) the spell is incomplete at the time of observation: All persons unemployed at the time of sampling will - unfortunately - stay longer in this state. The second case called 'right censoring' therefore is the more important case in duration analysis. Formally we have for the observed random variable Y in case of right censoring

$$Y = \begin{cases} Y^* & \text{if } Y^* \leq \tau \\ \tau & \text{if } Y^* > \tau, \end{cases} \tag{11.5}$$

where Y^* is the 'true' duration. For censored observation we only know the probability $P[Y^* > \tau] = 1 - F(\tau)$. Please note that usually τ is known but varies over the sampling units.

[4]See Ronning (1991, Chapter 4.3.4).

11.2.6 Tobit Models

Finally we shortly look at Tobit-type models which are closely related to duration models under censoring as noted, for example, by Amemiya (1985) although in this case we typically have *left* censoring. James Tobin analyzed expenditure data for durables and noted that expenditures could only be observed in case of purchase. Therefore in order to explain expenditures Y by income x he considered the 'latent linear model' (11.1) together with

$$Y = \begin{cases} \tau & \text{if } Y^* \leq \tau \\ Y^* & \text{if } Y^* > \tau, \end{cases} \tag{11.6}$$

which defines left-censoring. However, there is a fundamental difference to the duration model: Usually τ is an unknown parameter not varying over sampling units which has to be fixed a priori in order to estimate the remaining parameters α, β and σ_ε^2.[5] Here $P[Y^* \leq \tau]$ denotes the probability that the good is not purchased. (11.1) together with (11.6) assuming normality for ε is called the censored Tobit model.

Another situation arises if only data for buyers are available. In this case we consider the conditional distribution given the event $Y^* > \tau$. In other words, we consider a distribution which is 'truncated from below' and satisfies

$$P[Y^* > \tau] = 1.$$

A simple transformation of an unrestricted distribution characterized by density function $f(y)$ and distribution function $F(y)$ leads to the desired density

$$g(y|Y^* > \tau) = \frac{f(y)}{1 - F(\tau)}.$$

In case of the normal distribution expectation and variance of this truncated distribution are easily derived.[6] If again we assume normality for ε in (11.1) then we arrive at the 'truncated Tobit model' or just truncated regression model.

In connection with the Tobit model often the 'selectivity bias' is mentioned. In the censored version of the model above this problem only arises if the inappropriate least-squares estimation procedure is applied. The phenomenon of selectivity can arise in any model considered above in Section 11.2. More generally, this problem is termed 'endogenous sampling' which will not be treated in detail in this survey.[7]

11.2.7 Estimation and Testing

We have presented all models in this section without any discussion of estimation. Ronning (1991), for example, presents the maximum likelihood estimation for all these models together with some aspects of testing hypotheses. However, it is widely acknowledged that this estimation principle may be too restrictive in most

[5] See, for example, Ronning (1991, p. 125) for a discussion.
[6] See, for example, Ronning (1991, p. 13).
[7] See, for example, Ronning (1996, p. 86).

applications. More adequate methods are provided today using non-parametric and semi-parametric methods. Verbeek (2000) and Cameron and Trivedi (2005) may be consulted for some of these modern approaches. Additionally, some of the other contributions in this book consider estimation methods and other aspects of microeconometric models: Boes and Winkelmann consider other models for ordinal response. Caliendo and Hujer deal with microeconometric evaluation models. Fitzenberger and Wilke analyze duration models in terms of quantile regression. Hübler extends microeconometric models to the panel case. Kauermann treats estimation of non- and semi-parametric models in general terms. Rässler and Riphan are concerned with non-response which should be incorporated into estimation of microeconometric models if possible.

11.3 Anonymization of Micro Data

11.3.1 General Remarks

As already mentioned in the introduction, many sets of micro data are not available due to confidentiality. For some time now research has been done on the question of how to anonymize these data in such a way that the risk of disclosure is small and, at the same time, the distributional properties of the data are not too much biased. A handbook on anonymization published recently (see Ronning *et al.*, 2005b) gives an overview about procedures used in this field. For continuous variables both microaggregation and addition of noise seem to be good procedures since they both compromise in a satisfactory manner between guaranteeing confidentiality and conserving statistical properties. However, this has been established only for linear models whereas for (nonlinear) microeconometric models so far no general results have been derived. For example, addition of noise leads to the well-known errors-in-variables model for which proper estimation approaches are available in the linear case. See Subsection 11.3.2 for some more details. Since for all microeconometric models given in Section 11.2 the dependent variable is discrete or censored, other anonymization procedures should be applied in these cases. However, so far only anonymization of a binary variable by 'post randomization' (PRAM) has been treated more thoroughly. We describe all three methods in the following subsections. A more detailed exposition is given in Ronning *et al.* (2005b) where also other procedures such as rank swapping are described.

11.3.2 Microaggregation

This procedure assigns each observational unit from a cluster of observations to the arithmetic mean of this cluster. Therefore after anonymization the intra-class variance is zero implying a reduction of variance in general. One distinguishes between 'joint' microaggregation and individual microaggregation. In the first case the same cluster structure is used with regard to all variables whereas in the second case the aggregation is done for each variable individually. Of course, also subsets of variables may be aggregated by the same clustering or some subsets may be left unchanged.

We now demonstrate the effect of microaggregation on the estimation of the linear model[8] which we write in the usual way:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u}. \tag{11.7}$$

Let $\mathbf{D}_y$ be the aggregation matrix applied to $\mathbf{y}$ and $\mathbf{D}_X$ be the aggregation matrix applied to all columns of $\mathbf{X}$. Note that both matrices are symmetric idempotent. When estimating the parameter vector β from the anonymized data, we obtain

$$\hat{\beta} = (\mathbf{X}'\mathbf{D}_X\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}_X\mathbf{D}_y\mathbf{y}.$$

This shows that the estimator will only be unbiased if either $\mathbf{D}_X = \mathbf{D}_y$ (joint microaggregation of all variables) or if $\mathbf{D}_y = \mathbf{I}$ (only the set of regressors is microaggregated). Interestingly the case that only the dependent variable is aggregated creates problems. Schmid and Schneeweiß (2005) discuss this case and show consistency of the estimator under certain conditions. However, no general results are available for nonlinear models such as logit and probit models.

11.3.3 Addition of Noise

Let $\mathbf{e}_y$ be a vector of errors with expectation zero and positive variance corresponding to $\mathbf{y}$ and let $\mathbf{E}_X$ be a matrix of errors corresponding to $\mathbf{X}$ in the linear model (11.7). Addition of noise means that we have to estimate the unknown parameter vector from the model

$$\mathbf{y} + \mathbf{e}_y = (\mathbf{X} + \mathbf{E}_X)\beta + \mathbf{u}.$$

This is the well-known errors-in-variables model for which anonymization of right-hand variables creates estimation problems whereas anonymization of the dependent variable only increases the error variance[9] which should be compared with the case of microaggregation where (separate) anonymization of the dependent variable creates problems. Lechner and Pohlmeier (2005) consider nonparametric regression models where the regressors are anonymized by addition of noise. They show that from the simulation-extrapolation method (SIMEX) reliable estimates can be obtained. However for microeconometric models such as logit and probit models general results regarding the effect of noise addition and the suitability of the SIMEX method are not yet established.

Additive errors have the disadvantage that greater values of a variable are less protected. Take as an example sales of firms. If one firm has sales of 1 million and another sales of 100 million then addition of an error of 1 doubles sales of the first but leaves nearly unchanged sales of the second firm. Therefore research has been done also for the case of multiplicative errors which in this case should have expectation one. Formally this leads to

$$\mathbf{y} \odot \mathbf{e}_y = (\mathbf{X} \odot \mathbf{E}_X)\beta + \mathbf{u},$$

where $\odot$ denotes element-wise multiplication (Hadamard product). For results regarding estimation of this linear model see Ronning *et al.* (2005b).

[8] See Lechner and Pohlmeier (2003) for details.
[9] See Lechner and Pohlmeier (2003) for details.

11.3.4 Randomized Response and Post Randomization

Randomized response originally was introduced to avoid non-response in surveys containing sensitive questions on drug consumption or AIDS disease, see Warner (1965). Särndal *et al.* (1992, p. 573) suggested use of this method 'to protect the anonymity of individuals'. A good description of the difference between the two (formally equivalent) approaches is given by van den Hout and van der Heijden (2002): In the randomized response setting the stochastic model has to be defined in advance of data collection whereas in post randomization this method will be applied to the data already obtained.

Randomization of the binary variable Y can be described as follows: Let Y^m denote the 'masked' variable obtained from post randomization. Then the transition probabilities can be defined by $p_{jk} := P(Y^m = j|Y = k)$ with $j, k \in \{0, 1\}$ and $p_{j0} + p_{j1} = 1$ for $j = 0, 1$. If we define the two probabilities of no change by $p_{00} =: \pi_0$ and $p_{11} =: \pi_1$, respectively, the probability matrix can be written as follows:

$$\mathbf{P}_y = \begin{pmatrix} \pi_0 & 1-\pi_0 \\ 1-\pi_1 & \pi_1 \end{pmatrix}.$$

Since the two probabilities of the post randomization procedure usually are known[10] and there is no argument not to treat the two states symmetrically, in the following we will consider the special case

$$\pi_0 = \pi_1. \tag{11.8}$$

When the variable Y has undergone randomization, we will have a sample with n observations y_i^m where y_i^m is the dichotomous variable obtained from y_i by the randomization procedure.

In the handbook on anonymization (Ronning *et al.*, 2005b) we also discuss the extension of PRAM to more than two categories. If the categories are ordered as, for example, in the case of ordinal variables or count data switching probabilities for adjoining categories should be higher since otherwise the ordering would be totally destroyed. Of course, PRAM could also be extended to joint anonymization of two or more discrete variables.

11.4 The Probit Model under PRAM

11.4.1 Estimation of the Model

We now consider in more detail estimation of an important microeconometric model in case of anonymized data: the binary probit model as defined by (11.1) and (11.2) observing the normalizations mentioned in Subsection 11.2.1.[11] The sample

[10] We discuss the case of unknown probabilities in Subsection 11.4.3.

[11] See also Ronning (2005) and Ronning *et al.* (2005a).

information is given by n pairs (x_i, y_i) where $y_i \in \{0, 1\}$ and x_i is an arbitrary real number. Maximum likelihood estimation of the two unknown parameters α and β is straightforward.[12] In the following we confine ourselves to the case of just one regressor which is assumed to be continuous. The results, however, also apply to the more general case of an arbitrary number of explanatory variables after minor modifications. We consider randomization of the dichotomous variable y which switches its values with some prescribed transition probability (leaving the explanatory variable x in its original form).

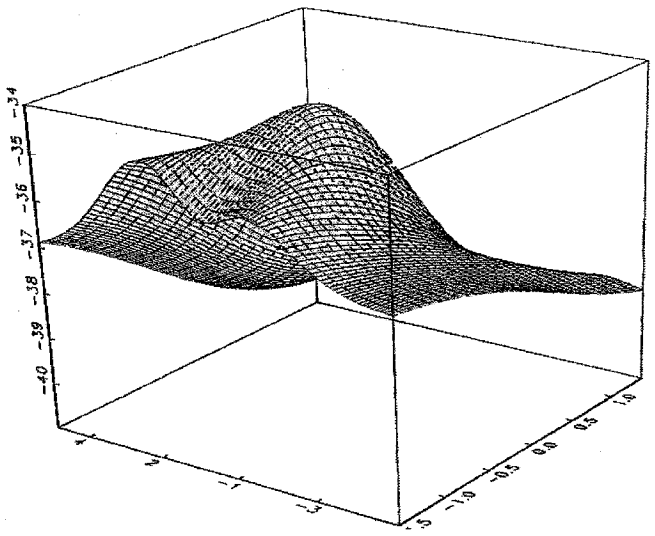


Figure 11.1: *Surface of Loglikelihood Function for Simulated Data with* $\pi = 0.70$.

Under randomization of the dependent observed variable we have the following data generating process:

$$Y_i^m = \begin{cases} 1 \text{ with probability } & \Phi_i \pi + (1 - \Phi_i)(1 - \pi) \\ 0 \text{ with probability } & \Phi_i (1 - \pi) + (1 - \Phi_i) \pi . \end{cases} \tag{11.9}$$

Here Φ_i denotes the conditional probability under the normal distribution that the unmasked dependent variable Y_i takes on the value 1 for given x_i, i. e. $\Phi_i := \Phi(\alpha + \beta x_i) = P(Y_i^* > 0 \mid x_i)$.

From (11.9) we obtain the following likelihood function:

$$\begin{aligned} &\mathcal{L}(\alpha, \beta | (y_i^m, x_i), i = 1, \ldots, n) \\ &= \prod_{i=1}^{n} (\Phi_i \pi + (1 - \Phi_i)(1 - \pi))^{y_i^m} (\Phi_i (1 - \pi) + ((1 - \Phi_i)\pi)^{(1 - y_i^m)} . \end{aligned} \tag{11.10}$$

[12] See some standard text as, for example, Greene (2003) or Ronning (1991).

Global concavity of this function with respect to α and β may be checked by deriving first and second (partial) derivatives of the log-likelihood function. Ronning (2005) derives the Hessian matrix of partial derivatives. Since this matrix is more complex than in the standard case, the proof of negative definiteness used in the standard probit case[13] does not go through here. This is also illustrated by Figure 11.4.1 which shows the surface of loglikelihood function for a simulated data set with $\pi = 0.70$.

A simple formula for the information matrix can be derived from which it is immediately apparent that maximum likelihood estimation under randomization is consistent but implies an efficiency loss which is greatest for values of π near 0.5, see Ronning (2005).

11.4.2 Marginal Effect in Case of the 'Naive' Probit Estimator

It is intuitively clear that the 'naive' probit estimator will be biased if the values of the dependent variable have been randomized although no analytic results are available. However it can be shown quite easily that application of a standard probit estimator will under-estimate the (true) marginal effect which is given by

$$\frac{\partial P(Y=1|x)}{\partial x} = \phi(\alpha+\beta x)\beta\,, \tag{11.11}$$

where we refer to our probit model given in (11.1) and (11.2). It follows from (11.9) that

$$P(Y^m=1|x) = \Phi_i\pi + (1-\Phi_i)(1-\pi) = (2\pi-1)\Phi + (1-\pi)\,, \tag{11.12}$$

where we disregard the observation index i. From this we obtain the marginal effect of the 'naive' estimator as follows:

$$\frac{\partial P(Y^m=1|x)}{\partial x} = (2\pi-1)\phi(\alpha+\beta x)\beta\,,$$

which may be also written as

$$\frac{\partial P(Y^m=1|x)}{\partial x} = (2\pi-1)\frac{\partial P(Y=1|x)}{\partial x}\,. \tag{11.13}$$

Therefore the naive estimator will *under*-estimate the true marginal effect as long as $1/2 < \pi < 1$ and will even reverse the sign of the marginal effect when the PRAM-parameter π satisfies $0 < \pi < 1/2$.

11.4.3 Estimation of Unknown Randomization Probabilities

Until now discussion is not finished on the question whether the details of the anonymization procedure should be made available when offering anonymized data

[13] See, for example, Amemiya (1985, p. 274) or Ronning (1991, p. 46).

sets. We therefore also consider the case that the switching probabilities π_0 and π_1 (defined in Subsection 11.3.4 and restricted by (11.8)) are not known to the user. Fortunately also in this case estimation can be done in a proper way as demonstrated by Hausman *et al.* (1998) in quite another context under the heading 'misclassification'. There it is assumed that a respondent has to answer a two-category question and erroneously chooses the wrong category. For example, employees could be asked whether they have changed their job during the last half year.

First note that the properly estimated probit function employing the 'corrected' likelihood (11.10) has a special property which may be used to estimate the probability π in case this would be unknown. Using (11.12) we obtain the following inequalities:[14]

$$\begin{aligned} 1-\pi \le P(Y^m = 1|x) \le \pi \quad \text{if} \quad \pi > \frac{1}{2} \\ \pi \le P(Y^m = 1|x) \le 1-\pi \quad \text{if} \quad \pi < \frac{1}{2}. \end{aligned} \tag{11.15}$$

Therefore the estimated probit function will have a smaller range if $\pi < 1$. For example, if $\pi = 0.25$ the probit function will only vary within the range $[0.25\,;0.75]$. In a certain way this result mirrors the increasing variance of the maximum likelihood estimator when π moves toward $1/2$: Since the 'middle range' of the probit function is becoming much smaller, the estimation will become more inaccurate.

Hausman *et al* (1998) have shown that it is possible to estimate the unknown probabilities via maximum likelihood jointly with the 'structural' parameters. However, they recommend a more flexible approach which in a first step uses Han's (1987) maximum rank correlation (MRC) estimator to determine the 'index'

$$\hat{y}_{\text{MRC}} := \mathbf{x}'\hat{\beta}_{\text{MRC}}\,.$$

In a second step then the fact is exploited that the response function $F(\hat{y})$ is monotonic with respect to the index $\hat{y}_{\text{MRC}}$. Therefore isotonic regression is used to fit the response function $\hat{F}(\hat{y}_{\text{MRC}})$. Then the unknown misclassification probabilities can be read off from the fitted function using the inequality generalizing (11.14), see, for example, Figure 5b in Hausman *et al.* (1998).

[14] For the general case, i.e. without restriction (11.8), we obtain the inequality

$$1-\pi_0 < P(Y^m = 1|x) < \pi_1 \tag{11.14}$$

assuming $\pi > 0.5$.

References

AMEMIYA, T. (1985). *Advanced Econometrics.* Basil Blackwell, Oxford.

CAMERON, A. C., TRIVEDI, P. (2005). *Microeconometrics. Methods and Applications.* Cambridge University Press, Cambridge.

GREENE, W. H. (2000). *Econometric Analysis.* 4th ed., Prentice Hall, Upper Saddle River.

HAN, A. K. (1987). Non-parametric analysis of a generalized regression model: The maximum rank correlation estimator. *Journal of Econometrics* **35** 303-316.

HAUSMAN, J. A., ABREVAYA, J., SCOTT-MORTON, F. M. (1998). Misclassification of the dependent variable in a discrete-response setting. *Journal of Econometrics* **87** 239-269.

KOMMISSION ZUR VERBESSERUNG DER INFORMATIONELLEN INFRASTRUKTUR (ED.) (2001). Wege zu einer besseren informationellen Infrastruktur. Nomos, Wiesbaden.

KOOIMAN, P., L. WILLENBORG, GOUWELEEUW, J. (1997). PRAM: a method for disclosure limitation of micro data.. http://www.cbs.nl/sdc/ruis.htm.

LECHNER, S., POHLMEIER, W. (2003). Schätzung ökonometrischer Modelle auf der Grundlage anonymisierter Daten. In *Anonymisierung wirtschaftsstatistischer Einzeldaten* (R. Gnoss, G. Ronning, eds.), Forum der Bundesstatistik, **42** 115-137.

LECHNER, S., POHLMEIER, W. (2005). Data masking by noise addition and the estimation of nonparametric regression models. *Jahrbücher für Nationalökonomie und Statistik* **225** 517-528.

POHLMEIER, W., RONNING, G., WAGNER, J. (2005). Econometrics of anonymized micro data. *Jahrbücher für Nationalökonomie und Statistik* **225** Sonderband.

RONNING, G. (1991). *Mikroökonometrie.* Springer, Berlin.

RONNING, G. (1996). Ökonometrie. In *Springers Handbuch der Volkswirtschaftslehre* (A. Börsch-Supan, J. v. Hagen, P. J. J. Welfens, eds.), 78-134. Springer, Berlin.

RONNING, G. (2005). Randomized response and the binary probit model. *Economics Letters* **86** 221-228.

RONNING, G., GNOSS, R. (2003). Anonymisierung wirtschaftsstatistischer Einzeldaten. Statistisches Bundesamt. Forum der Bundesstatistik, Band 42, Wiesbaden.

RONNING, G, ROSEMANN, M., STROTMANN, H. (2005a). Post-randomization under test: Estimation of the probit model. *Jahrbücher für Nationalökonomie und Statistik* **225** 544-566.

RONNING, G., STURM, R., HÖHNE, J., LENZ, J., ROSEMANN, M., SCHEFFLER, M., VORGRIMLER, D. (2005b). Handbuch zur Anonymisierung wirtschaftsstatistischer Mikrodaten. Statistisches Bundesamt, Wiesbaden, Reihe „Statistik und Wissenschaft", Band 4, 2005.

SÄRNDAL, C.-E., SWENSSON, B., WRETMAN, J. (1992). *Model Assisted Survey Sampling*. Springer, New York.

SCHMID, M., SCHNEEWEISS, H. (2005). The effect of microaggregation procedures on the estimation of linear models. *Jahrbücher für Nationalökonomie und Statistik* **225** 529-543.

VAN DEN HOUT, A., VAN DER HEIJDEN, P. G. M. (2002). Randomized response, statistical disclosure control and misclassification: A review. *International Statistical Review* **70** 2-69.

VERBEEK, M. (2000). *A Guide to Modern Econometrics*. Wiley, Chichester.

WARNER, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association* **57** 622-627.

12 Ordered Response Models *

Stefan Boes[1] and Rainer Winkelmann[2]

[1] Socioeconomic Institute, University of Zurich
boes@sts.unizh.ch
[2] Socioeconomic Institute, University of Zurich
winkelmann@sts.unizh.ch

Summary: We discuss regression models for ordered responses, such as ratings of bonds, schooling attainment, or measures of subjective well-being. Commonly used models in this context are the ordered logit and ordered probit regression models. They are based on an underlying latent model with single index function and constant thresholds. We argue that these approaches are overly restrictive and preclude a flexible estimation of the effect of regressors on the discrete outcome probabilities. For example, the signs of the marginal probability effects can only change once when moving from the smallest category to the largest one. We then discuss several alternative models that overcome these limitations. An application illustrates the benefit of these alternatives.

12.1 Introduction

Regression models for ordered responses, i. e. statistical models in which the outcome of an ordered dependent variable is explained by a number of arbitrarily scaled independent variables, have their origin in the biometrics literature. Aitchison and Silvey (1957) proposed the ordered probit model to analyze experiments in which the responses of subjects to various doses of stimulus are divided into ordinally ranked classes. Snell (1964) suggested the use of the logistic instead of the normal distribution as an approximation for mathematical simplification. The first comprehensive treatment of ordered response models in the social sciences appeared with the work of McKelvey and Zavoina (1975) who generalized the model of Aitchison and Silvey to more than one independent variable. Their basic idea was to assume the existence of an underlying continuous latent variable – related to a single index of explanatory variables and an error term – and to obtain the observed categorical outcome by discretizing the real line into a finite number of intervals.

McCullagh (1980) developed independently the so-called *cumulative model* in the statistics literature. He directly modelled the cumulative probabilities of the ordered

*We are grateful to an anonymous referee for valuable comments.

outcome as a monotonic increasing transformation of a linear predictor onto the unit interval, assuming a logit or probit link function. This specification yields the same probability function as the model of McKelvey and Zavoina, and is therefore observationally equivalent. Both papers spurred a large literature on how to model ordered dependent variables, the former mostly in the social sciences, the latter predominantly in the medical and biostatistics literature.

On the one hand, a number of parametric generalizations have been proposed. These include alternative link functions, prominent examples being the log-log or the complementary log-log function (McCullagh, 1980), generalized predictor functions that include, for example, quadratic terms or interactions, or dispersion parameters (Cox, 1995). Olsson (1979) and Ronning and Kukuk (1996) discuss estimation of models in which both dependent and independent variables are ordered in the context of multivariate latent structural models, i. e. an adaptation of log-linear models to ordinal data. On the other hand, semi- and nonparametric approaches replace the distributional assumptions of the standard model, or the predictor function, by flexible semi- or nonparametric functional forms. General surveys of the parametric as well as the semi- and nonparametric literature are given, for example, in Agresti (1999), Barnhart and Sampson (1994), Clogg and Shihadeh (1994), Winship and Mare (1984), Bellemare *et al.* (2002), and Stewart (2004), the two latter references in particular for the semi- and nonparametric treatments of ordered data.

When thinking about the usefulness of these alternative models, it is inevitable to make up one's mind on the ultimate objective of the analysis. It is our perception that in most applications of ordered response models the parameters of the latent model do not have direct interpretation *per se*. Rather, the interest lies in the shift of the predicted discrete ordered outcome distribution as one or more of the regressors change, i. e. the *marginal probability effects*. Perhaps surprisingly, standard ordered response models are not very well suited to analyze these marginal probability effects, because the answer is to a large extent predetermined by the rigid parametric structure of the model. Therefore, we consider a number of generalizations that allow for flexible analyses of marginal probability effects. In addition to the generalized threshold model (Maddala, 1983; Terza, 1985; Brant, 1990) and the sequential model (Fienberg, 1980; Tutz, 1990, 1991), we show how additional flexibility can be gained by modeling individual heterogeneity either by means of a random coefficients model or as a finite mixture/latent class model.

The remainder of the paper is organized as follows. In the next section we provide a short review of the standard model, before turning to the generalizations in Section 3. In Section 4 we illustrate the methods with an analysis of the relationship between income and happiness using data from the German Socio-Economic Panel. Our results show that marginal probability effects in the generalized alternatives are substantially different from those in the standard model. For example, the standard model implies that the probability of being *completely satisfied* increases on average by about 0.017 percentage points by a one-percentage increase in income, while it is decreasing or constant in the generalized models. Section 5 concludes.

12.2 Standard Ordered Response Models

Consider the following examples. In a survey, respondents have been asked about their life-satisfaction, or their change in health status. Answer categories might range from 0 to 10 where 0 means *completely dissatisfied* and 10 means *completely satisfied*, or from 1 to 5, where 1 means *greatly deteriorated* and 5 means *greatly improved*, respectively. The objective is to model these ordered responses as functions of explanatory variables.

Formally, let the ordered categorical outcome y be coded, without loss of generality, in a rank preserving manner, i. e. $y \in \{1, 2, \ldots, J\}$ where J denotes the total number of distinct categories. Furthermore, suppose that a $(k \times 1)$-dimensional vector x of covariates is available. In standard ordered response models the cumulative probabilities of the discrete outcome are related to a single index of explanatory variables in the following way

$$\Pr[y \leq j|x] = F(\kappa_j - x'\beta) \qquad j = 1, \ldots, J, \tag{12.1}$$

where κ_j and $\beta_{(k\times 1)}$ denote unknown model parameters, and F can be any monotonic increasing function mapping the real line onto the unit interval. Although no further restrictions are imposed *a priori* on the transformation F it is standard to replace F by a distribution function, the most commonly used ones being the standard normal (which yields the ordered probit) and the logistic distribution (associated with the ordered logit model), and we assume in what follows that F represents either the standard normal or logistic distribution. In order to ensure well-defined probabilities, we require that $\kappa_j > \kappa_{j-1}$, $\forall j$, and it is understood that $\kappa_J = \infty$ such that $F(\infty) = 1$ as well as $\kappa_0 = -\infty$ such that $F(-\infty) = 0$.

Ordered response models are usually motivated by an underlying continuous but latent process $y^\star$ together with a response mechanism of the form

$$y = j \quad \text{if and only if} \qquad \kappa_{j-1} \leq y^\star = x'\beta + u < \kappa_j \qquad j = 1, \ldots, J,$$

where $\kappa_0, \ldots, \kappa_J$ are introduced as threshold parameters, discretizing the real line, represented by $y^\star$, into J categories. The latent variable $y^\star$ is related linearly to observable and unobservable factors and the latter have a fully specified distribution function $F(u)$ with zero mean and constant variance.

The cumulative model (12.1) can be postulated without assuming the existence of a latent part and a threshold mechanism, though. Moreover, since $y^\star$ cannot be observed and is purely artificial, its interpretation is not of interest. The main focus in the analysis of ordered data should be put on the conditional cell probabilities given by

$$\Pr[y = j|x] = F(\kappa_j - x'\beta) - F(\kappa_{j-1} - x'\beta)\,. \tag{12.2}$$

In order to identify the parameters of the model we have to fix location and scale of the argument in F, the former by assuming that x does not contain a constant term, the latter by normalizing the variance of the distribution function F. Then, equation (12.2) represents a well-defined probability function which allows for straightforward application of maximum likelihood methods for a random sample of size n of pairs (y, x).

The most natural way to interpret ordered response models (and discrete probability models in general) is to determine how a marginal change in one regressor changes the distribution of the outcome variable, i. e. all the outcome probabilities. These marginal probability effects can be calculated as

$$MPE_{jl}(x) = \frac{\partial \Pr[y = j|x]}{\partial x_l} = \Big[f(\kappa_{j-1} - x'\beta) - f(\kappa_j - x'\beta)\Big]\beta_l\,, \quad (12.3)$$

where $f(z) = \mathrm{d}F(z)/\mathrm{d}z$ and x_l denotes the l-th (continuous) element in x. With respect to a discrete valued regressor it is more appropriate to calculate the change in the probabilities before and after the discrete change Δx_l,

$$\Delta \Pr[y = j|x] = \Pr[y = j|x + \Delta x_l] - \Pr[y = j|x]\,. \quad (12.4)$$

In general, the magnitude of these probability changes depends on the specific values of the ith observation's covariates. After taking expectation with respect to x we obtain average marginal probability effects, which can be estimated consistently by replacing the true parameters by their corresponding maximum likelihood estimates and taking the average over all observations.

However, if we take a closer look at (12.3) and (12.4) it becomes apparent that marginal probability effects in standard ordered response models have two restrictive properties that limit the usefulness of these models in practice. First, the ratio of marginal probability effects of two distinct continuous covariates on the same outcome, i. e. *relative* marginal probability effects, are constant across individuals and the outcome distribution, because from (12.3) we have that

$$\frac{MPE_{jl}(x)}{MPE_{jm}(x)} = \frac{\beta_l}{\beta_m}\,,$$

which does not depend on i and j. Second, marginal probability effects change their sign exactly once when moving from the smallest to the largest outcome. More precisely, if we move stepwise from the lowest category $y = 1$ to the highest category $y = J$, the effects are either first negative and then positive ($\beta_l > 0$), or first positive and then negative ($\beta_l < 0$). This 'single crossing property' follows directly from the bell-shaped density functions of the standard normal and the logistic distribution. Therefore, if we are interested in the effect of a covariate on the outcome probabilities, i. e. if we turn our attention to the effects on the full distribution of outcomes, the standard models preclude a flexible analysis of marginal probability effects by design.

12.3 Generalized Ordered Response Models

Three assumptions of the standard model are responsible for its limitations in analyzing marginal probability effects: First, the single index assumption, second, the constant threshold assumption, and third, the distributional assumption which does not allow for additional individual heterogeneity between individual realizations. While relaxing these assumptions we want to retain the possibility of interpreting the model in terms of marginal probability effects. Therefore, we need to search for a richer class of parametric models that does not impose restrictions such as constant relative effects or single crossing. In this section we present four such alternatives.

12.3.1 Generalized Threshold Model

The first model we consider relaxes the single index assumption and allows for different indices across outcomes. This model was introduced by Maddala (1983) and Terza (1985) who proposed to generalize the threshold parameters by making them dependent on covariates

$$\kappa_j = \tilde{\kappa}_j + x'\gamma_j \,,$$

where γ_j is a $k \times 1$-dimensional vector of response specific parameters. Plugging this into (12.1) we get the cumulative probabilities in the generalized threshold model

$$\Pr[y \leq j|x] = F(\tilde{\kappa}_j + x'\gamma_j - x'\beta) = F(\tilde{\kappa}_j - x'\beta_j) \qquad j = 1, \ldots, J \,, \tag{12.5}$$

where it is understood that $\tilde{\kappa}_0 = -\infty$ and $\tilde{\kappa}_J = \infty$, as before. The last equality in (12.5) follows because γ_j and β cannot be identified separately with the same x entering the index function and the generalized thresholds, and we define $\beta_j := \beta - \gamma_j$. The cumulative probabilities define a probability density function in the same manner as in (12.2) and parameters can be estimated directly by maximum likelihood. A non-linear specification can be used to ensure that $\tilde{\kappa}_{j-1} - x'\beta_{j-1} < \tilde{\kappa}_j - x'\beta_j$ for all $\tilde{\kappa}$, $\tilde{\beta}$ and x (e. g. Ronning, 1990). We observe that the generalized threshold model nests the standard model under the restrictions $\beta_1 = \ldots = \beta_{J-1}$ and therefore both models can be tested against each other by performing a likelihood ratio (LR) test.

The generalized threshold model provides a framework in which marginal probability effects can be analyzed with much more flexibility than in the standard model, since

$$MPE_{jl}(x) = f(\tilde{\kappa}_{j-1} - x'\beta_{j-1})\beta_{j-1l} - f(\tilde{\kappa}_j - x'\beta_j)\beta_{jl} \tag{12.6}$$

does not rely anymore on a single crossing property or constant relative effects. Nevertheless, this generalization comes at a cost. The model now contains $(J-2)k$ parameters more than before which reduces the degrees of freedom considerably, in particular when J is large.

12.3.2 Random Coefficients Model

As a second alternative we discuss the class of random coefficients models. The basic idea is to randomize the parameters of interest by adding an error term that is correlated with the unobserved factors in u. Thus, we translate individual heterogeneity into parameter heterogeneity, writing the vector of slopes as

$$\beta = \tilde{\beta} + \varepsilon \,,$$

where ε is an individual specific $(k \times 1)$-dimensional vector of error terms. Moreover, we assume for the joint error term $\gamma := (\varepsilon' \;\; u)'$ that

$$\mathrm{E}[\gamma|x] = 0 \quad \text{and} \quad E[\gamma\gamma'|x] = \Sigma \;\; \text{with} \;\; \Sigma = \begin{pmatrix} \Omega & \psi \\ \psi' & 1 \end{pmatrix} \,,$$

where Ω is the $(k \times k)$-dimensional covariance matrix of ε, ψ is the $(k \times 1)$-dimensional covariance vector between the slope parameters and u, and $\mathrm{Var}[u|x] = 1$, as before. The consequences of this modification are easiest seen from the latent variable representation, where we now have $y^\star = x'\tilde{\beta} + \tilde{u}$ with 'new' error term $\tilde{u} := x'\varepsilon + u$, such that

$$\mathrm{E}[\tilde{u}|x] = 0 \quad \text{and} \quad \mathrm{E}[\tilde{u}\tilde{u}'|x] = x'\Omega x + 2x'\psi + 1 \;=: \sigma_{\tilde{u}}^2 ,$$

and $\tilde{u}/\sigma_{\tilde{u}}$ is distributed with distribution function F. If ε and u are jointly normal with covariance structure given by Σ, we obtain an ordered probit model with unobserved heterogeneity. However, in principle, we do not need to know the distributions of ε or u, as long as F is a well-defined distribution function. In this case, we can express the cumulative probabilities in the random coefficients model as

$$\Pr[y \leq j|x] = F\left(\frac{\kappa_j - x'\tilde{\beta}}{\sigma_{\tilde{u}}}\right) \;=: \tilde{F}_j(x) , \tag{12.7}$$

where $\sigma_{\tilde{u}} = \sqrt{x'\Omega x + 2x'\psi + 1}$ can be seen as dispersion parameter. The standard model is a special case of the random coefficients model under the assumption $\Omega = 0$ and $\psi = 0$. Thus, a simple LR test can be used to test for parameter heterogeneity.

The probability density function of y is obtained in the same way as in (12.2), and one can calculate marginal probability effects in the random coefficients model as

$$\begin{aligned} MPE_{jl}(x) &= \left[\tilde{f}_{j-1}(x) - \tilde{f}_j(x)\right] \frac{\tilde{\beta}_l}{\sigma_{\tilde{u}}} \\ &+ \left[\tilde{f}_{j-1}(x)\left(\kappa_{j-1} - x'\tilde{\beta}\right) - \tilde{f}_j(x)\left(\kappa_j - x'\tilde{\beta}\right)\right] \frac{x'\Omega_l + \psi_l}{\sigma_{\tilde{u}}^3} \end{aligned} \tag{12.8}$$

by using product and chain rules. In (12.8), Ω_l denotes the l-th column in Ω and ψ_l the l-th element in ψ, respectively, and $\tilde{f}(z) = \mathrm{d}\tilde{F}(z)/\mathrm{d}z$. The first term in (12.8) corresponds to the marginal probability effects in the standard model corrected for the standard deviation of the disturbance $\tilde{u}$. The second term arises because we assume a specific form of heteroscedasticity which makes the error term dependent on x. Consequently, marginal probability effects in the random coefficient model are more flexible than those in the standard model since the sign of the second term is indeterminate.

The random coefficients model can be estimated directly by the method of maximum likelihood with heteroscedasticity corrected index function. However, some caution is required in running the optimization routines. Although the parameters of the model are identified by functional form, the specific structure of the model might cause problems in some datasets. Specifically, certain values of Ω, ψ and x can drive $\sigma_{\tilde{u}}^2$ to be negative or its square root to be almost linear in the parameters, such that the argument in F gets complex or is not identified, respectively. Nevertheless, if the data support the model, we should find reasonable estimates of the elements in Ω and ψ.

12.3.3 Finite Mixture Model

The third approach is a finite mixture model for ordered data (Everitt, 1988; Everitt and Merette, 1990; Uebersax, 1999) which provides a very flexible way of modeling heterogeneity among groups of individuals. It is supposed that the population is split into C distinct latent classes and each class has its own data-generating process, i. e. we relax the distributional assumption of the standard model and its implied homogeneity. To fix ideas, let $c = 1, \ldots, C$ denote the index of classes and write the cumulative probabilities for class c as

$$\Pr[y_c \leq j|x] = F(\kappa_{cj} - x'\beta_c) \quad =: F_{cj}(x)\,.$$

However, individual class membership is not observable and we assume that each individual belongs to a certain class c with probability π_c. Thus, we can write the cumulative probabilities of the observed outcomes as a mixture of class specific cumulative probabilities

$$\Pr[y \leq j|x] = \sum_{c=1}^{C} \pi_c F_{cj}(x)\,, \tag{12.9}$$

where the π_c's sum up to unity. The probability density function of the ordered outcome is given by $\Pr[y = j|x] = \sum_c \pi_c \Big(F_{cj}(x) - F_{cj-1}(x)\Big)$ and marginal probability effects can be obtained, as before, by taking the first order derivative with respect to x_l

$$MPE_{jl}(x) = \sum_{c=1}^{C} \pi_c \Big(f_{cj-1}(x) - f_{cj}(x)\Big)\beta_{cl}\,. \tag{12.10}$$

Again, the sign of marginal probability effects is indeterminate because of the dependence on π_c and β_{cl} which might differ in magnitude and sign among classes. The statistical significance of these differences can be tested by conducting a LR test with restrictions $\pi_1 = \ldots = \pi_C$ and $\beta_1 = \ldots = \beta_C$, that is, a total number of $(C-1)(k+1)$ restrictions. Uebersax (1999) gives conditions for identification of class specific thresholds and slope parameters.

The parameters of the finite mixture model can be estimated directly via maximum likelihood. This requires maximization of a (in general multimodal) log-likelihood function of the form

$$\ln L(\theta, \pi|y, x, z) = \sum_{i=1}^{n}\sum_{j=1}^{J} y_{ij} \ln \left\{ \sum_{c=1}^{C} \pi_c \Big(F_{cj}(x_i) - F_{cj-1}(x_i)\Big) \right\},$$

where θ and π is shorthand notation for the vectors of class specific parameters θ_c (which include thresholds and slopes) and probabilities π_c, respectively, and y_j is a binary variable indicating whether $y = j$. The multimodality of the log-likelihood function and the large number of parameters for increasing C might cause the optimization routines to be slow in finding the global maximum. Furthermore, although the probability function of the complete mixture might be well-defined, the probabilities in a subset of classes can turn negative. An alternative approach of getting the maximum likelihood estimates that circumvents these problems is to

formulate the model as an incomplete data problem and to apply the EM algorithm of Dempster *et al.* (1977).

To be more specific, let m_c denote a binary variable indicating individual class membership which can be interpreted as independent realizations of a C-component multinomial distribution with component probabilities π_c, the prior probability of belonging to class c. The (complete-data) log-likelihood function for a random sample of size n conditional on observed class membership m can be written as

$$\ln L(\theta, \pi | y, x, m) = \sum_{i=1}^{n} \sum_{j=1}^{J} y_{ij} \sum_{c=1}^{C} m_{ci} \Big\{ \ln \pi_c + \ln \Big(F_{cj}(x_i) - F_{cj-1}(x_i) \Big) \Big\}. \tag{12.11}$$

Since we cannot observe individual class membership, that is the data are incomplete, we cannot maximize this log-likelihood function directly.

The EM algorithm proceeds iteratively in two steps, based on an E-step in which the expectation of (12.11) is taken with respect to m given the observed data and the current fit of θ and π, and an M-step in which the log-likelihood function (12.11) is maximized with respect to θ and π given expected individual class membership. The linearity of the complete-data log-likelihood in m allows for direct calculation of the expected individual class membership given the observed data and the parameters obtained in the q-th iteration step. This expectation corresponds to the probability of the ith entity belonging to class c, henceforth called posterior probability τ_c. From the assumptions above or simply by Bayes' theorem it can be shown that

$$\tau_c \Big(y, x; \theta^{(q)}, \pi^{(q)} \Big) = \frac{\pi_c^{(q)} \Big(F_{cj}^{(q)}(x) - F_{cj-1}^{(q)}(x) \Big)}{\sum_{c=1}^{C} \pi_c^{(q)} \Big(F_{cj}^{(q)}(x) - F_{cj-1}^{(q)}(x) \Big)}, \tag{12.12}$$

where $F_{cj}^{(q)}$ denotes the value of F evaluated at the parameters obtained in the q-th iteration step. These probabilities can be used to anaylze the characteristics of each class, i. e. we can assign each individual to the class for which its probability is the highest and then derive descriptive statistics or marginal probability effects per class.

The M-step replaces m_c in (12.11) by its expectation, τ_c, and therefore considers the expected log-likelihood to be maximized. Again, the linearity in (12.11) provides a substantial simplification of the optimization routine. First, updated estimates of $\pi_c^{(q+1)}$ can be obtained directly by taking the sample average $n^{-1} \sum_i \tau_c(.)$ where $0 \leq \tau_c(.) \leq 1$ (see (12.12)). Secondly, each class can be maximized separately with respect to θ_c to get updated estimates $\theta_c^{(q+1)}$ taking into account the multiplicative factor τ_c. In other words, we can estimate C simple ordered probits or logits while weighting the data appropriately and alter the E- and M-steps repeatedly until the change in the difference between the log-likelihood values is sufficiently small.

12.3.4 Sequential Model

The last alternative for a flexible ordered response model adopts methods from the literature on discrete time duration data. In this literature, the main quantity of

interest is the conditional exit probability (or 'hazard rate') $\Pr[y = j|y \geq j, x]$, where y is the duration of the spell and j is the time of exit. The key insight is that such discrete time hazard rate models can be used for any ordered response y. Once the conditional transition probabilities are determined, the unconditional probabilities are obtained from the recursive relationship

$$\Pr[y = j|x] = \Pr[y = j|y \geq j, x] \Pr[y \geq j|x] \qquad j = 1, \ldots, J\,, \quad (12.13)$$

where

$$\begin{aligned} \Pr[y \geq 1|x] &= 1\,, \\ \Pr[y \geq j|x] &= \prod_{r=1}^{j-1} \Big\{1 - \Pr[y = r|y \geq r, x]\Big\} \qquad j = 2, \ldots, J\,, \quad (12.14) \end{aligned}$$

and it is understood that $\Pr[y = J|y \geq J, x] = 1$. Using (12.13) and (12.14) the whole probability function of y can be expressed in terms of conditionals, or more precisely, as a sequence of binary choice models where each decision is made for a specific category j conditional on refusing all categories smaller than j. This kind of model can be motivated by a sequential response mechanism where each of the J outcomes can be reached only step-by-step, starting with the lowest category, and therefore the model is referred to as *sequential model*. This model implicitly accounts for the ordering information in y without assuming any cardinality in the threshold mechanism.

To complete the model we specify the conditional transition probabilities as

$$\Pr[y = j|y \geq j, x] = F(\alpha_j + x'\beta_j) = F_j(x) \qquad j = 1, \ldots, J\,, \quad (12.15)$$

where α_j is a category specific constant, β_j is a category specific slope parameter, and it is understood that $\alpha_J = \infty$ such that $F_J(\infty) = 1$. Therefore, in contrast to previously discussed models, we do not parameterize the cumulative probabilities but rather the conditional transition probabilities. The parameters can be estimated by running j consecutive binary choice models where the dependent variable is the binary indicator y_j defined in the previous section, and only observations with $y \geq j$ are included. Therefore, estimation is simplified considerably compared to the generalized threshold and the random coefficients model since no further restrictions on the parameter space are required. The downside is that computation of the marginal probability effects is now more complicated. It can be shown that

$$\begin{aligned} MPE_{1l}(x) &= f_1(x)\beta_{1l}\,, \\ MPE_{jl}(x) &= f_j(x)\beta_{jl} \Pr[y \geq j|x] - F_j(x) \sum_{r=1}^{j-1} MPE_{rl}(x)\,, \quad j = 2, \ldots, J\,. \end{aligned} \quad (12.16)$$

Clearly, these effects are very flexible, as they can vary by category and do not rely on a single crossing property or constant relative effects. The sequential model and the standard model are nonnested models and one may use information based measures like the *Akaike Information Criterion* (AIC) as a model selection criterion. Moreover, for the problem of choosing among the generalized alternatives the same strategy is advisable.

Table 12.1: Model selection.

	Ordered Probit	Generalized Threshold	Sequential Probit	Random Coefficients	Finite Mixture
No. of param.	[13]	[49]	[49]	[15]	[26]
$\ln L$	-3040.58	-2999.59	-2999.12	-3035.88	-3024.65
AIC	6107.16	6097.18	6096.24	6101.76	6101.30
No. of obs.	1735				

Notes: The data were drawn from the 1997 wave of the German Socio-Economic Panel, the dependent variable *happiness* with originally eleven categories (0-10) was recoded to avoid cells with low frequency; we subsumed categories 0-2 in j=1, categories 3/4 in j=2, the remaining in ascending order up to j=8.

12.4 Empirical Illustration

In order to illustrate the benefit of the generalized ordered response models we analyze the effect of income on happiness using data from the German Socio-Economic Panel (GSOEP; see also Boes and Winkelmann, 2004). The relationship between income and happiness was studied before in a number of papers (see, for example, Easterlin, 1973, 1974; Scitkovsky, 1975; Frey and Stutzer, 2000, 2002; Shields and Wheatley Price, 2005, and the references therein) and has gained renewed interest in the recent literature because of its use for valuation of public goods or intangibles (see, for example, Winkelmann and Winkelmann, 1998; Frey *et al.*, 2004; van Praag and Baarsma, 2005).

We used data from the 1997 wave of the GSOEP and selected a sample of 1735 men aged between 25 and 65. The dependent variable *happiness* with originally 11 categories was recoded to avoid cells with low frequency and, after merging the lower categories 0/1/2 and 3/4, we retained a total of $J = 8$ ordered response categories. We included among the regressors logarithmic family income and logarithmic household size as well as a quadratic form in age, and two dummy variables indicating good health status as well as unemployment.

In our regression analysis, we assumed that F is the cumulative density function of the standard normal distribution. The random coefficients model was simplified by restricting Ω and ψ such that $\sigma_u^2 = \Omega_{ll} x_l^2 + 2\psi_l x_l + 1$, where x_l is assumed to be logarithmic income, Ω_{ll} denotes the l-th diagonal element in Ω and ψ_l the l-th element in ψ. Thus, we confine our analysis to parameter heterogeneity in the income coefficient, with all other parameters being deterministic. In the finite mixture model, we considered only two latent classes ($C = 2$). The following discussion proceeds in two steps: First, we evaluate the models by means of likelihood ratio tests and selection criteria, and second, we examine the implications for interpretion in terms of marginal probability effects.

The first question we address is whether one of the models presented above uses the information inherent in the data optimally. For this purpose, we perform likelihood ratio tests or AIC comparisons, depending on the situation. For example,

Table 12.2: *Marginal probability effects of income on happiness.*

	Ordered Probit	Generalized Threshold	Sequential Probit	Random Coeff.	Finite Mixture	
					Class 1	Class 2
$j = 1$	-0.0076	-0.0098	-0.0083	-0.0165	-1.3e-07	-0.0245
$j = 2$	-0.0228	-0.0096	-0.0155	-0.0391	-0.0076	-0.0092
$j = 3$	-0.0223	-0.0352	-0.0338	-0.0297	-0.0024	-0.0565
$j = 4$	-0.0160	-0.0444	-0.0410	-0.0140	-0.0026	-0.0285
$j = 5$	-0.0090	0.0039	0.0095	0.0135	-0.0030	0.0198
$j = 6$	0.0328	0.0680	0.0697	0.0589	0.0028	0.0920
$j = 7$	0.0275	0.0403	0.0334	0.0234	0.0073	0.0069
$j = 8$	0.0173	-0.0133	-0.0140	0.0035	0.0056	5.7e-08

Notes: The table reports average marginal probability effects of logarithmic income on happiness responses, $AMPE_{j,\ln(income)}$. For example, in the ordered probit model $AMPE_{6,\ln(income)} = 0.0328$ means that the probability of $j = 6$ increases by about 0.0328 percentage points given an increase in logarithmic income by 0.01 (which corresponds to an increase in income by about 1 percent).

the differences between the generalized threshold and the standard ordered probit model are statistically significant if we can reject the null hypothesis of no category specific parameters. This can be investigated by running a likelihood ratio test with minus two times the difference between the log-likelihoods of the standard and the generalized model as appropriate test statistic, showing a value of 79.98. The test statistic is asymptotically χ^2-distributed with 36 degrees of freedom. Thus, we can reject the null hypothesis, and thereby the standard orderd probit model. Likewise, we can compare the random coefficients model as well as the finite mixture model with the ordered probit, the latter being rejected in both cases. The sequential model and the standard ordered probit are nonnested models which rules out the application of an LR test. Instead, we may calculate the AIC for each model, showing values of 6 107.96 and 6 096.24 for the ordered probit and the sequential probit, respectively. A smaller value indicates a better fit while penalizing for the proliferation of parameters, and, although 36 parameters more, we favor the sequential probit to the ordered probit model. Furthermore, among the generalized alternatives the generalized threshold and the sequential model have the smallest AIC values, followed by the finite mixture model and the random coefficients model.

We now turn our attention to average marginal probability effects of income on happiness. The MPE's of the ordered probit model are reported in the first column of Table 2. Our results show a positive coefficient of logarithmic income, implying a negative sign of the MPE's for low happiness responses, switching into the positive for $j \geq 6$. The interpretation of, for example, $MPE_6 = 0.0328$ is that a one-percent increase in income raises the probability of $happiness = 6$ by approximately 0.0328

percentage points. Compared to the standard model, the generalized threshold and the sequential model yield substantially different effects (see Columns 2 and 3). First, the sign of MPE_5 changes, indicating a positive effect also for the fifth category. Second, the magnitude of some MPE's are clearly underestimated by the standard model. For example, the estimated MPE_6 in the generalized ordered response models is more than twice as large as in the ordered probit. Third, and probably most important, the sign of the marginal probability effect in the utmost right part of the outcome distribution turns out to be negative, violating the single crossing requirement of the simple model. This means that an increase in income actually *reduces* the probability of being very happy, a result consistent with the view that 'money does not buy happiness'.

The results of the random coefficients model are reported in the fourth column of Table 2. The calculated MPE's tend to support the results of the generalized threshold and the sequential model, although there is no negative effect on the highest happiness response. However, the random coefficient specification provides further insights into the relationship between income and happiness. We estimated $\hat{\Omega}_{ll} = 0.60$ and $\hat{\psi}_l = -0.77$, the latter implying that unobservables in the happiness equation are negatively correlated with the random coefficient. This can be interpreted as follows: If unobservables in the happiness equation tend to increase the probability of higher responses, then the effect of income is lower for these individuals.

In the finite mixture model we can make use of the posterior probabilities to obtain marginal probability effects per class (see Columns 5 and 6). The results indicate that the effect of income on happiness can be neglected for one class (the relatively happy class with average happiness of 5.71) whereas for the class of relatively unhappy people (average happiness of 4.25) income plays a much more important role.

12.5 Concluding Remarks

In this paper we argued that the standard ordered probit and ordered logit models, while commonly used in applied work, are characterized by some restrictive and therefore non-desirable properties. We then discussed four generalized models, namely the generalized threshold, the random coefficients, the finite mixture, and the sequential model. All of them are substantially more flexible in analyzing marginal probability effects since they do not rely on constant relative effects or a single crossing property.

An illustrative application with data from the 1997 wave of the GSOEP dealt with the relationship between income and happiness. We asked how a one-percent increase in income is predicted to change the happiness distribution, ceteris paribus. The analysis showed that the estimated marginal probability effects differed markedly between the standard ordered probit model and the probit-specified alternatives. For example, a negative marginal effect for the highest answer category (as predicted by the generalized threshold model) is ruled out by assumption in the standard model.

As is not uncommon with such generalizations, they can be computationally burdensome due to the larger number of parameters, restrictions on the parameter space, or a multimodality of the likelihood function. Nevertheless, the greater flexibility and enhanced interpretation possibilities should render these alternative models indispensable tools in all research situations, where an accurate estimation of the marginal probability effects over the entire range of the outcome distribution is of interest.

References

AITCHISON, J., SILVEY, S. D. (1957). The generalization of probit analysis to the case of multiple repsonses. *Biometrika* **44** 131–140.

AGRESTI, A. (1999). Modelling ordered categorical data: Recent advances and future challenges. *Statistics in Medicine* **18** 2191–2207.

ANDERSON, J. A. (1984). Regression and ordered categorical variables. *Journal of the Royal Statistical Society, Series B* **46** 1–30.

BARNHART, H. X., SAMPSON, A. R. (1994). Overview of multinomial models for ordered data. *Communications in Statistics - Theory and Methods* **23** 3395–3416.

BELLEMARE C., MELENBERG, B., VAN SOEST, A. (2002). Semi-parametric models for satisfaction with income. *Portuguese Economic Journal* **1** 181–203.

BOES, S., WINKELMANN, R. (2004). Income and happiness: New results from generalized threshold and sequential models. IZA Discussion Paper No. 1175, SOI Working Paper No. 0407, Bonn.

BRANT, R. (1990). Assessing proportionality in the proportional odds model for ordered logistic regression. *Biometrics* **46** 1171–1178.

CLOGG, C. C., SHIHADEH, E. S. (1994). *Statistical Models for Ordered Variables.* Sage Publications, Thousand Oaks.

COX, C. (1995). Location-scale cumulative odds models for ordered data: A generalized non-linear model approach. *Statistics in Medicine* **14** 1191–1203.

DEMPSTER, A. P., LAIRD, N. M., RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* **39** 1–38.

EASTERLIN, R. (1973). Does money buy happiness?. *Public Interest* **30** 3–10.

EASTERLIN, R. (1974). Does economic growth improve the human lot? Some empirical evidence. In *Nations and Households in Economic Growth: Essays in Honor of Moses Abramowitz* (P. David, M. Reder, eds.), 89–125. Academic Press, New York.

EVERITT, B. S. (1988). A finite mixture model for the clustering of mixed-mode data. *Statistics and Probability Letters* **6** 305–309.

EVERITT, B. S., MERETTE, C. (1990). The clustering of mixed-mode data: A comparison of possible approaches. *Journal of Applied Statistics* **17** 283–297.

FIENBERG, S. E. (1980). *The Analysis of Cross-Classified Categorical Data.* MIT Press, Cambridge.

FREY, B. S., LUECHINGER, S., STUTZER, A. (2004). Valuing public goods: The life satisfaction approach. CESifo Working Paper No. 1158, München.

FREY, B. S., STUTZER, A. (2000). Happiness, economy and institutions. *The Economic Journal* **110** 918–938.

FREY, B. S., STUTZER, A. (2002). *Happiness and Economics: How the Economy and Institutions Affect Human Well-Being.* Princeton University Press, Princeton and Oxford.

MADDALA, G. (1983). *Limited-Dependent and Qualitative Variables in Econometrics.* Cambridge University Press, Cambridge.

MCCULLAGH, P. (1980). Regression models for ordered data. *Journal of the Royal Statistical Society, Series B* **42** 109–142.

MCKELVEY, R., ZAVOINA, W. (1975). A statistical model for the analysis of ordered level dependent variables. *Journal of Mathematical Sociology* **4** 103–120.

OLSSON, U. (1979). Maximum-likelihood estimation of the polychoric correlation coefficient. *Psychometrika* **44** 443–460.

RONNING, G. (1990). The informational content of responses from business surveys. In *Microeconometrics. Surveys and Applications* (J. P. Florens, M. Ivaldi, J. J. Laffont, F. Laisney, eds.), 123–144. Basil Blackwell, Oxford.

RONNING, G., KUKUK, M. (1996). Efficient estimation of ordered probit models. *Journal of the American Statistical Association* **91** 1120–1129.

SCITOVSKY, T. (1975). Income and happiness. *Acta Oeconomica* **15** 45–53.

SHIELDS, M., WHEATLEY PRICE, S. (2005). Exploring the economic and social determinants of psychological well-being and perceived social support in England. *Journal of The Royal Statistical Society, Series A* **168** 513–537.

SNELL, E. J. (1964). A scaling procedure for ordered categorical data. *Biometrics* **20** 592–607.

STEWART, M. B. (2004). A comparison of semiparametric estimators for the ordered response model. *Computational Statistics and Data Analysis* **49** 555–573.

TERZA, J. (1985). Ordered probit: A generalization. *Communications in Statistics – Theory and Methods* **14** 1–11.

TUTZ, G. (1990). Sequential item response models with an ordered response. *British Journal of Mathematical and Statistical Psychology* **43** 39–55.

TUTZ, G. (1991). Sequential models in ordered regression. *Computational Statistics and Data Analysis* **11** 275–295.

UEBERSAX, J. S. (1999). Probit latent class analysis with dichotomous or ordered category measures: Conditional independence/dependence models. *Applied Psychological Measurement* **23** 283–297.

VAN PRAAG, B. M. S., BAARSMA, B. E. (2005). Using happiness surveys to value intangibles: The case of airport noise. *The Economic Journal* **115** 224–246.

WINKELMANN, L., WINKELMANN, R. (1998). Why are the unemployed so unhappy? Evidence from panel data. *Economica* **65** 1–15.

WINSHIP, C., MARE, R. D. (1984). Regression models with ordered variables. *American Sociological Review* **49** 512-525.

13 Some Recent Advances in Measurement Error Models and Methods*

Hans Schneeweiß[1] and Thomas Augustin[2]

[1] Department of Statistics, Ludwig-Maximilians-Universität
schneew@stat.uni-muenchen.de
[2] Department of Statistics, Ludwig-Maximilians-Universität
augustin@stat.uni-muenchen.de

Summary: A measurement error model is a regression model with (substantial) measurement errors in the variables. Disregarding these measurement errors in estimating the regression parameters results in asymptotically biased estimators. Several methods have been proposed to eliminate, or at least to reduce, this bias, and the relative efficiency and robustness of these methods have been compared. The paper gives an account of these endeavors. In another context, when data are of a categorical nature, classification errors play a similar role as measurement errors in continuous data. The paper also reviews some recent advances in this field.

13.1 Introduction

A measurement error model is a – linear or non-linear – regression model with (substantial) measurement error in the variables, above all in the regressor variable. Disregarding these measurement errors in estimating the regression parameters (naive estimation) results in asymptotically biased, inconsistent, estimators. This is the motivation for investigating measurement error models. Measurement errors are found in almost all fields of application. A classical example in econometrics is Friedman's (1957) 'permanent income hypothesis'. Another example is the measurement of schooling as a predictor of wage earnings (Card, 2001). In epidemiology, various studies may be cited where the impact of an exposure to noxious substances on the health status of people is studied (e.g., Heid *et al.*, 2002). In engineering, the calibration of measuring instruments deals with measurement errors by definition (Brown, 1982). Many more examples can be found in the literature,

*This work was supported by the Deutsche Forschungsgemeinschaft (DFG) within the frame of the Sonderforschungsbereich SFB 386. We thank two anonymous referees for their helpful comments.

in particular in the monographs by Schneeweiss and Mittag (1986), Fuller (1987), Carroll *et al.* (1995), Cheng and Van Ness (1999), Wansbeek and Meijer (2000). Recently measurement error methods have been applied in the masking of data to assure their anonymity (Brand, 2002). The data are artificially distorted in various ways including through the addition of random errors.

Several estimation methods have been proposed to eliminate, or at least to reduce, the bias of the naive estimation method. The present paper reviews some of these methods and compares their efficiencies.

Section 13.2 introduces the measurement error model. In Section 13.3 we discuss briefly the identification problem. Sections 13.4 to 13.6 deal with various estimation procedures, and Section 13.7 compares their efficiencies. Section 13.8 addresses survival models. A special type of measurement errors, viz., misclassification errors is dealt with in Section 13.9. Section 13.10 has some concluding remarks.

13.2 Measurement Error Models

A measurement error model consists of three parts:

1. A *regression model* relating an unobservable (generally vector-valued, but here for simplicity scalar) regressor variable ξ to a response variable y given by a conditional distribution $f(y|\xi;\theta)$, where θ is an unknown parameter vector. Quite often only the conditional mean function $\mathbb{E}(y|\xi) = m^*(\xi,\beta)$, the regression in the narrower sense, is given, supplemented by a conditional variance function $\mathbb{V}(y|\xi) = v^*(\xi,\beta,\varphi)$, where θ comprises β and φ plus possibly other parameters describing the distribution of y.

Two major examples, that we will often refer to, are the polynomial model (for a survey see Cheng and Schneeweiss, 2002),

$$y = \beta_0 + \beta_1\xi + \cdots + \beta_k\xi^k + \epsilon$$

with $m^*(\xi,\beta) = \beta_0+\beta_1\xi+\cdots+\beta_k\xi^k$ and $v^* = \sigma_\epsilon^2$, and the log-linear Poisson model

$$y|\xi \sim Po(\lambda), \quad \lambda = \exp(\beta_0 + \beta_1\xi)$$

with $m^*(\xi,\beta) = v^*(\xi,\beta) = \lambda$. Survival models are considered separately in Section 13.8.

2. A *measurement model* that relates the unobservable ξ to an observable surrogate variable x, given by a conditional distribution $g(x|\xi;\alpha)$. The so-called non-differentiality property requires that $f(y|\xi,x) = f(y|\xi)$. The classical measurement model assumes an additive random error δ with mean zero, which is independent of ξ and (by non-differentiality) of y

$$x = \xi + \delta\,.$$

An alternative is the so-called Berkson model, where δ is independent of x instead of being independent of ξ (e.g., Küchenhoff *et al.*, 2003). Here we shall only consider the classical model. Typically δ is assumed to be normally distributed: $\delta \sim \mathrm{N}(0, \sigma_\delta^2)$.

3. A *distribution of the latent regressor variable* ξ. The distribution may be specified by a density $h(\xi; \gamma)$ with an unknown parameter vector γ. We then have the *structural variant* of the model. Another possibility is that ξ is not considered a random variable but rather an unknown parameter pertaining to the observation x. In this case, which is called the *functional variant*, the number of parameters ξ grows with the sample size. We do not deal with this case here (but see Cheng and Van Ness, 1999). Instead, following Caroll *et al.* (1995), we distinguish between structural and functional estimation methods. The former use the distribution of ξ, the latter do not, even if such a distribution exists. Estimation of β is based on an i.i.d. sample of data (x_i, y_i), $i = 1, \cdots, n$. For an example of estimation in the context of time series see Nowak (1993).

13.3 Identifiability

Since ξ is latent, the parameters of the model may not be identified. This is the case in the linear model and in the probit model both with normally distributed regressor and error variables. In such cases additional pieces of information are necessary in order to be able to construct consistent estimators for β, for more details see Cheng and Van Ness (1999). But even if the model is identified (as is often the case in non-linear models - for the logistic model see Küchenhoff, 1995; for the quadratic regression model, see Huang and Huwang, 2001), additional information may be of great help to enhance the efficiency of estimation. The most prominent pieces of extra information are knowledge of the error process, in particular the measurement error variance σ_δ^2, and knowledge of instrumental variables. Here we will only deal with the first type of information (for the second, see Schneeweiss and Mittag, 1986, and Wansbeek and Meijer, 2000). Knowledge of σ_δ^2 may come from repeated measurements or from a validation subsample. For an example where knowledge of $h(\xi)$ is used see Hu and Ridder (2005).

13.4 Naive Estimation and Bias Correction

Suppose that a consistent estimator $\hat{\beta}$ for the original, error-free model is available. Simply replacing ξ with x in this estimator gives rise to the so-called *naive* estimator $\hat{\beta}_N$. Simple as it is, this estimator is almost always not consistent.

As an example consider the linear model $y = \alpha + \beta\xi + \epsilon$. The naive estimator of β is the LS estimator $\hat{\beta}_N = s_{xy}/s_x^2$, which has the bias $-(\sigma_\delta^2/\sigma_x^2)\beta$. Note that $|\beta|$ is systematically underestimated by $|\hat{\beta}_N|$ (attenuation effect). This has the undesirable consequence that a strong effect of the covariate ξ on y may not be detectable anymore once the covariate has been corrupted by measurement errors. In the multiple linear model, measurement errors have a more complicated effect (see Schneeweiss and Mittag, 1986). In the quadratic model the attenuation effect is

expressed as a flattening of the curvature at the peak of the parabola (Kuha and Temple, 2003). A segmented linear regression shows a smooth curve connecting the two segments instead of the sharp kink of the error-free model (Küchenhoff and Caroll, 1997).

When the (asymptotic) bias $B = \text{plim}\hat{\beta}_N - \beta$ can be evaluated (typically as a function of β and possibly other parameters), it is sometimes possible to correct the naive estimator such that a consistent estimator results. For instance, the bias of $\hat{\beta}_N$ in the linear model can be easily corrected if σ_δ^2 is known:

$$\hat{\beta}_C = \frac{s_x^2}{s_x^2 - \sigma_\delta^2}\hat{\beta}_N = \frac{s_{xy}}{s_x^2 - \sigma_\delta^2}$$

is a consistent estimator of β. Another example is the bias correction of the naive ML estimator in a logistic model (Küchenhoff, 1992).

13.5 Functional Estimation Methods

Functional estimators do not use the distribution of ξ. They are therefore immune against possible misspecifications of $h(\xi)$ and they are also valid when ξ is non-stochastic. In this latter case the problem of estimating the incidental parameters ξ_i arises (Cheng and Van Ness, 1999). However, one can circumvent this problem and can directly find estimators for the parameter of interest β. We present two such estimators: CS and SIMEX.

13.5.1 Corrected Score (CS) Estimator

Suppose we have a (vector-valued) unbiased estimating (or simply: score) function $\psi(y, \xi; b)$ such that $b = \beta$ is the only solution to the equation $\mathbb{E}[\psi(y, \xi; b)] = 0$. Then the solution $\hat{\beta}$ of $\sum_{i=1}^n \psi(y_i, \xi_i; \hat{\beta}) = 0$, assuming that it exists uniquely, is (under general regularity conditions) a consistent estimator of β. However, as ξ is unobservable, this estimator is not feasible. Therefore, one may try to find a so-called corrected score function $\psi_{CS}(y, x; b)$ such that

$$\mathbb{E}[\psi_{CS}(y, x; b)|y, \xi] = \psi(y, \xi; b)$$

(Nakamura, 1990). With the help of the iterative expectation principle, ψ_{CS} can be seen to be an unbiased estimating function, and so, under mild regularity conditions, $\hat{\beta}_{CS}$ solving

$$\sum_{i=1}^n \psi_{CS}(y_i, x_i; \hat{\beta}_{CS}) = 0$$

is a consistent and asymptotically normal estimator (the CS estimator). Its asymptotic covariance matrix is given by the sandwich formula

$$\Sigma_{CS} = \frac{1}{n} A_{CS}^{-1} B_{CS} A_{CS}^{-\top}, \quad \text{with } A_{CS} = -\mathbb{E}\left(\frac{\partial \psi_{CS}}{\partial \beta}\right), \; B_{CS} = \mathbb{E}(\psi_{CS}\psi_{CS}^T),$$

where $\psi_{CS} = \psi_{CS}(y, x; \beta)$. A common score function of the error-free model is

$$\psi(y, \xi; b) = [y - m^*(\xi, b)]\, v^{*-1} \frac{\partial m^*(\xi, b)}{\partial b}\,.$$

We then need to find functions f_1 and f_2 such that

$$\mathbb{E}\,[f_1(x, b)|\xi] = v^{*-1} m_b^*\,, \qquad \mathbb{E}\,[f_2(x, b)|\xi] = m^* v^{*-1} m_b^*\,,$$

where m_b^* is short for $\partial m^*(\xi, b)/\partial b$. Stefanski (1989) gives conditions for the existence of such functions. If they exist, then $\psi_{CS} = y f_1 - f_2$.

In the polynomial model one can construct polynomials $t_r(x)$ of degree r such that $\mathbb{E}\,[t_r(x)|\xi] = \xi^r$ (Cheng and Schneeweiss, 1998, and Cheng *et al.*, 2000). The corrected score function is then given by

$$\psi_{CS}(y, x; b) = H(x)b - y t(x)\,,$$

where $t(x) = (t_0(x), \cdots, t_k(x))^\top$ and $H(x)$ is a $(k+1)\times(k+1)$ matrix with $H_{rs}(x) = t_{r+s}(x)$, $r, s = 0, \cdots, k$, from which the CS estimator is found as $\hat{\beta}_{CS} = \bar{H}^{-1}\overline{yt}$, where the bar denotes averaging over the sample values (x_i, y_i).

In the Poisson model (see Shklyar and Schneeweiss, 2005), the corrected score function is given by

$$\psi_{CS}\,(y, x; b_0, b_1) = \left(y - \lambda e^{-\frac{1}{2}b_1^2\sigma_\delta^2}\right)(1, x)^\top + \lambda b_1 \sigma_\delta^2 e^{-\frac{1}{2}b_1^2\sigma_\delta^2}\,(0, 1)^\top\,.$$

13.5.2 Simulation-Extrapolation (SIMEX) Estimator

One cannot subtract the measurement error, but one can add a random error to the x_i and thereby study the effect of measurement errors on the estimate of β. This idea gives rise to the following method (Cook and Stefanski, 1994):

1. Compute the naive estimate $\hat{\beta}_N =: \hat{\beta}_{(0)}$.
2. Add random noise to the x_i: $x_i'(a) = x_i + \delta_i'(a)$, $\delta_i'(a) \sim N(0, a\sigma_\delta^2)$ and compute the naive estimate with these artificial data (y_i, x_i').
3. Repeat this step m times with a fixed a and average the m naive estimates to get an estimate $\hat{\beta}_{(a)}$.
4. Do this for a series of a's, $a = 0.1, 0.2, \cdots, 2$. One may plot the resulting points $(a, \hat{\beta}_{(a)})$.
5. Fit a curve through these points by least squares using some convenient function, e. g., a quadratic one.
6. Extrapolate this curve to $a = -1$, which corresponds to the situation of no measurement error. Then $\hat{\beta}_{SIMEX} = \hat{\beta}_{(-1)}$.

This procedure is easy to apply, as it uses only the naive estimation method given from the original error-free model. It is, however, very computer intensive and it only gives a consistent estimator if the correct extrapolation curve has been used (see Carroll *et al.*, 1996). The quadratic curve may be convenient, but it is rarely the correct curve. SIMEX estimators are therefore often biased, but the bias is typically greatly reduced as compared to the bias of the naive estimator (Wolf, 2004).

13.6 Structural Estimation Methods

Structural estimation methods use the information given in the distribution of the regressor variable. Note, however, that this distribution $h(\xi;\gamma)$ contains the unknown (nuisance) parameter vector γ. Typically γ can be estimated from the data x_i alone without recourse to the regression model. For instance, if $\xi \sim N(\mu_\xi, \sigma_\xi^2)$ the nuisance parameters μ_ξ and σ_ξ^2 can be estimated by $\bar{x}$ and $s_x^2 - \sigma_\delta^2$, respectively. For how to estimate γ in a distribution which is a mixture of normals see Thamerus (2003). Replacing γ with a consistent estimate $\hat{\gamma}$ does not alter the consistency property of $\hat{\beta}$, though it does have an effect on the asymptotic variance (cf. Caroll *et al.*, 1995). For simplicity, let us assume in the sequel that γ is known. We will consider three estimators: ML, QS, and RC.

13.6.1 Maximum likelihood (ML) Estimator

The joint density of x and y is given by

$$q(x, y; \theta, \alpha, \gamma) = \int f(y|\xi;\theta) g(x|\xi;\alpha) h(\xi;\gamma)\, d\xi\,.$$

Maximizing it with respect to θ, α, γ gives the ML estimator. Though being the most efficient estimator, it has two drawbacks: it relies on the complete joint distribution of x and y and is therefore sensitive to any kind of misspecification and, due to the integral, it is in most cases extremely difficult to compute, not the least because all the parameters have to be estimated simultaneously. Although the computational burden can be greatly alleviated by using simulation methods (simulated ML, simulated LS, see Wansbeek and Meijer, 2000; Li, 2000, or Hsiao and Wang, 2000), there is still demand for simpler, and more robust, estimation methods. Two of these, QS and RC, will now be discussed.

13.6.2 The Quasi Score (QS) Estimator

The *(structural) quasi score* (QS) estimator is constructed by means of the conditional mean and variance function of y given x:

$$\mathbb{E}(y|x) = m(x;\beta), \qquad \mathbb{V}(y|x) = v(x;\beta,\varphi).$$

These are computed starting from the original mean and variance functions given ξ:

$$m(x;\beta) = \mathbb{E}\left[m^*(\xi;\beta)|x\right], \quad v(x;\beta,\varphi) = \mathbb{V}\left[m^*(\xi;\beta)|x\right] + \mathbb{E}\left[v^*(\xi;\beta,\varphi)|x\right].$$

For these computations we need the conditional distribution of ξ given x. In some cases this distribution may be found directly from validation data. In most other cases it is computed from $g(x|\xi;\alpha)$ and $h(\xi;\gamma)$. Therefore m and v do not only depend on β (and φ) but also on α and γ. Here we assume that α and γ are given. In the classical measurement error model with $\delta \sim \mathrm{N}(0,\sigma_\delta^2)$ and $\xi \sim \mathrm{N}(\mu_\xi,\sigma_\xi^2)$ the conditional distribution of ξ given x is simply given by

$$\xi|x \sim \mathrm{N}(\mu(x),\tau^2) \text{ with } \mu(x) = \mu_x + \left(1 - \frac{\sigma_\delta^2}{\sigma_x^2}\right)(x-\mu_x)\,,\ \tau^2 = \sigma_\delta^2\left(1-\frac{\sigma_\delta^2}{\sigma_x^2}\right).$$

The quasi score function for β then is

$$\psi_{QS}(y,x;b,\varphi) = [y - m(x;b)]\, v^{-1}(x;b,\varphi) m_b(x,b)\,.$$

This should be supplemented by a quasi score function for φ, which we have suppressed for ease of presentation. Given φ, the QS estimator is found as the solution to

$$\sum_{i=1}^{n} \psi_{QS}(y_i,x_i;\hat\beta_{QS},\varphi) = 0\,.$$

As ψ_{QS} is an unbiased estimating function, $\hat\beta_{QS}$ is, under appropriate regularity conditions, a consistent, asymptotically normal estimator with an asymptotic covariance matrix that is again given by a sandwich formula (Kukush and Schneeweiß, 2005).

For the polynomial model, first construct $\mathbb{E}(\xi^r|x) = \mu_r(x)$, which is a polynomial of degree r. The QS estimator is then found from the heteroscedastic regression equations

$$\begin{aligned} y &= \beta_0 + \beta_1\mu_1(x) + \cdots + \beta_k\mu_k(x) + u\,,\\ \sigma_u^2 &= \sigma_\epsilon^2 + \sum_{r=0}^{k}\sum_{s=0}^{k}(\mu_{rs}(x) - \mu_r(x)\mu_s(x))\,\beta_r\beta_s \end{aligned}$$

by applying an iteratively reweighted least squares procedures (Kukush *et al.*, 2001).

For the Poisson model (see Shklyar and Schneeweiss, 2005),

$$\begin{aligned} m(x;\beta) &= \exp(\beta_0 + \beta_1\mu(x) + \frac{1}{2}\beta_1^2\tau^2)\,,\\ v(x;\beta) &= m(x;\beta) + \left[\exp(\beta_1^2\tau^2) - 1\right] m^2(x;\beta)\,. \end{aligned}$$

13.6.3 The Regression Calibration (RC) Estimator

The regression calibration estimator is even simpler to compute than the QS estimator (see Carroll *et al.*, 1995). One replaces the variable x in the naive estimator by $\mu(x)$, which is the best linear predictor of ξ given x (Gleser, 1990).

Thus in the polynomial model, the RC estimator is the LS estimator of the regression

$$y = \beta_0 + \beta_1 \mu(x) + \cdots + \beta_k \mu(x)^k + \epsilon\,.$$

In the Poisson model, the RC estimator is the ML estimator of a Poisson model with $\lambda = \exp\{\beta_0 + \beta_1 \mu(x)\}$.

Unfortunately, the RC estimator is inconsistent in general, an exception being the linear model, where RC = QS = CS. But in most cases the bias is greatly reduced as compared to the naive estimator and often negligible (Wolf, 2004).

13.7 Efficiency Comparison

In this section we compare CS and QS with respect to their relative efficiencies. Various results that have been found in the last years will be summarized (Kukush and Schneeweiß, 2005; Shklyar and Schneeweiss, 2005; Schneeweiss and Cheng, 2006; Shklyar *et al.*, 2005).

We assume that $\delta \sim N(0, \sigma_\delta^2)$ and $\xi \sim N(\mu_\xi, \sigma_\xi^2)$. Thus we are in the structural case. In addition, a very general regression model of the exponential family is assumed:

$$f(y|\xi) = \exp\left(\frac{y\lambda - c(\lambda)}{\varphi} + a(y, \varphi)\right), \quad \text{with} \quad \lambda = \lambda(\xi, \beta)\,.$$

This model comprises the polynomial and the Poisson model as well as other generalized linear models. Note that in this model $m^* = c'(\lambda)$ and $v^* = \varphi c''(\lambda)$, which will be the basis for constructing the CS and QS estimators. Clearly, the ML estimator is the most efficient one. One might speculate that QS is more efficient than CS, as the latter ignores the information inherent in the distribution of ξ. However, this is not at all clear, as QS is not ML. Nevertheless one can, indeed, prove that the presumption is correct, i. e. $\Sigma_{ML} \leq \Sigma_{QS} \leq \Sigma_{CS}$, at least as long as the nuisance parameters μ_ξ and σ_ξ^2 are given and need not be estimated. Thus if ML is avoided because of its complexity, QS seems to be the estimator of ones choice.

But QS depends on the distribution $f(\xi)$ of the latent regressor. If this distribution is misspecified, then $\hat{\beta}_{QS}$ will typically be biased. Suppose that the true distribution is a finite mixture of normals which cluster around a single normal, erroneously assumed to be the true distribution, and suppose the average distance ϑ of the modes (and the variances) of the mixture components is small and tends to zero, then the misspecification bias of $\hat{\beta}_{QS}$ is of the order ϑ^2. Therefore, in most cases, the bias is practically negligible. There are, however, other forms of misspecification which are not that benign. In any case, misspecification of the regressor distribution is a serious problem with QS.

>From that point of view, one might prefer CS as the more robust estimator. Even more so, as for small measurement errors, QS and CS and also ML become almost equally efficient anyway. More precisely:

$$\Sigma_{CS} = \Sigma_{ML} + O(\sigma_\delta^4)\,, \quad \Sigma_{QS} = \Sigma_{ML} + O(\sigma_\delta^4)\,.$$

One can also compare CS and QS to the naive method (N). Of course, N is biased. But according to a general rule of thumb one might surmise that the bias of N is compensated by a smaller covariance matrix. Most often this is true, but there are cases where $\Sigma_{CS} - \Sigma_N$ is indefinite or where $\Sigma_{QS} < \Sigma_N$.

13.8 Survival Analysis

In survival analysis the time until a certain event occurs ('survival time') is considered. The characteristic issue making survival analysis a separate area of research is the problem of censoring: Typically not all survival times T_i, $i = 1, \cdots, n$, can be observed completely; for a subset of the units it is only known that unit i is still alive at some censoring time C_i.

13.8.1 Measurement Error in Cox-type Models

Mainly two classes of regression models have been studied. The first one, which is due to Cox (1972), relates the individual hazard rate $\lambda(t|\xi)$ to the covariates ξ and the regression parameter β according to the relationship $\lambda(t|\xi) = \lambda_0(t) \cdot \exp(\beta\xi)$. The so-called baseline hazard rate $\lambda_0(t)$ characterizes the dynamic development of risk over time, and is assumed not to depend on i, the hazards are proportional to each other. Most often, $\lambda_0(t)$ is seen as an unspecified nuisance function making the model semiparametric. In particular in econometrics, also parametric versions are of interest (Flinn and Heckman, 1982).

There are two classical papers on measurement errors in Cox-type models, namely the work by Prentice (1982) and by Nakamura (1992), both providing - to some extent - negative results. Prentice (1982), who relies on the structural case, has shown that a simple likelihood-based correction along the lines of Section 13.6.1 is not possible (see also Augustin and Schwarz, 2002): The resulting induced relative risk has the form

$$\lambda(t|x) = \lambda_0(t) \cdot \mathbb{E}[\exp(\beta\xi)|x, \{T \geq t\}] \,. \tag{13.1}$$

Via the event $\{T \geq t\}$ appearing in the conditional expectation, the second factor depends on the previous history of the process, and so the characteristic multiplicative form of the Cox model is lost. As a consequence partial likelihood maximization, the usual estimation method for the Cox model, cannot be directly applied anymore.

However, as Prentice also argued, the effect of this time dependence can be expected to be small if the failure intensity is very low. Under this so-called *rare disease assumption* the condition $\{T \geq t\}$ is almost always satisfied, and so (13.1) can be solved analytically for normal measurement errors. Then the resulting estimator for β coincides with that obtained from regression calibration, which moreover turns out to be the same as the naive estimator multiplied by the simple deattenuation factor known from linear regression (cf. Section 13.4). Pepe *et al.* (1989) discuss the accuracy of this approximation (see also Hughes, 1993) and derive further results on handling (13.1) directly.

Further structural approaches are provided by Hu *et al.* (1998). In general, structural approaches appear promising for dealing with Berkson errors, which, for instance, occurs in cohort studies on exposure to risk factors (Bender *et al.*, 2005; Küchenhoff *et al.*, 2003).

The classical paper from the functional point of view is Nakamura (1992), who tries to apply his general method of corrected score function (Nakamura, 1990; see also Section 2) to partial likelihood estimation. However, the partial likelihood has a singularity in the complex plain, and so - according to a general result from Stefanski (1989) - a corrected score function cannot exist. Nakumara (1992) therefore proposes to correct first and second order approximations, instead. The resulting estimators behave not only well in simulation studies, but, surprisingly, the estimator based on first order correction even turned out to be consistent (Kong and Gu, 1999). Moreover, Kong *et al.* (1998) derive a corresponding correction of the cumulative baseline hazard rate $\Lambda_0(t) := \int_0^t \lambda_0(u)du$. Both results are extended in Kong and Gu (1999) to the case of non-normal measurement error. Huang and Wang (2000) suggest a nonparametric variant based on replication data.

A different justification of Nakamura's method for the Cox model and related work is provided by Augustin (2004). He shows that these seemingly approximate corrections are exact corrections, indeed, arising in a straightforward manner when Nakamura's original concept of corrected score function is applied to the so-called Breslow likelihood instead of partial likelihood. This approach immediately extends to those proportional hazards models where the baseline hazard rate is parameterized and to almost arbitrary measurement error distributions.

Alternative functional correction methods include Buzas' (1998) approach and applications of the so-called conditional score principle in longitudinal Cox models (see, in particular, Tsiatis and Davidian, 2004).

13.8.2 Accelerated Failure Time Models

The second class of survival models assumes a linear relationship between the log-survival time and the predictor: $\ln T = \beta_0 + \beta\xi + \sigma\epsilon$. This model provides a superstructure upon the common parametric duration models like the Weibull, log-logistic, log-normal and gamma models, which are obtained by appropriate specification of ϵ. Recently, also the non-parametric variant, where the distribution of ϵ is left unspecified, has experienced a renaissance.

Correction methods for the Weibull model under covariate measurement error have been presented and compared by Gimenez *et al.* (1999). Skinner and Humphreys (1999), Wolff and Augustin (2003), and Augustin and Wolff (2004) discuss Weibull regressions under error-prone or heaped lifetimes.

The simple linear structure in the logarithm of T also suggests to use mean and variance function models. Augustin (2002, Chapter 5f.) derives the corresponding corrected estimating equations to adjust for measurement errors, both from the structural as well as from the functional point of view. The methods obtained allow for a unified treatment of all the commonly used parametric duration models and are the first to handle measurement errors in the covariates and lifetimes simultaneously. Censoring, however, needs additional attention (cf. Augustin, 2002,

Theorem 6.2.2), since the estimation equations do not rely on the likelihood anymore.

13.9 Misclassification

Misclassification of categorical variables is another type of measurement error. As an example, consider a generalized linear model (GLM) for a dichotomous response variable y taking values 0 and 1 with

$$\mathbb{P}(y = 1|x) = G(\kappa), \quad \kappa = x\beta\,,$$

and suppose the response y is occasionally misclassified as y^*. Then using y^* instead of the unknown y in estimating β will produce a bias.

Define the misclassification probabilities

$$\pi_{ij} := \mathbb{P}(y^* = i|y = j, x) = \mathbb{P}(y^* = i|y = j),$$

where the second equality is a consequence of the nondifferentiality postulate. If the π_{ij} are known (as, when misclassification is used as a masquing device to anonymize data, see Ronning, 2005), or if they can be estimated, (through a validation study, see Schuster, 1998), then consistent estimators can be constructed. Just observe that

$$\mathbb{P}(y^* = 1|x) = \pi_{11}G(\kappa) + \pi_{10}(1 - G(\kappa)) =: H(\kappa)$$

is again a GLM and can be estimated by conventional methods. For further details see Hausman *et al.* (1998).

Recently Küchenhoff *et al.* (2005) developed a variant of the SIMEX method (see Section 13.5.2) to be applied to models of the above kind and to more complicated ones. By artificially contorting the data y^* through further misclassification and estimating the resulting models in a naive way, as if the data were not misclassified, one gets an idea of the amount of bias due to misclassification. One can then extrapolate to the state of no misclassification.

13.10 Concluding Remarks

In this survey we restricted our presentation to parametric regression models in explicit form. We should like to mention a few other approaches.

Functional relations between variables ξ_1 and ξ_2, say, can also be given in the implicit form $f(\xi_1, \xi_2; \beta) = 0$. If instead of ξ_1 and ξ_2 we observe surrogates x_1 and x_2 with additive measurement errors: $x_i = \xi_i + \delta_i$, $i = 1, 2$, and if the error variances are known to be equal, then orthogonal, or total, least squares (TLS) is the method of choice. TLS works nicely in linear models (Cheng and Van Ness, 1999),

but leads to biased estimation in nonlinear models. But there is an asymptotic small-σ_δ theory (Fuller, 1987; Amemiya and Fuller, 1988). For the quadratic model, consistent estimators exist (Kukush *et al.*, 2004).

We mentioned that masquing of data can be seen as a method of adding artificial measurement errors to the data. However, these measurement errors are often of a quite different type than those considered in this paper. In particular, microaggregation is such a method, which may lead to biased regression estimators. In order to deal with this bias new methods have been developed (Schmid *et al.*, 2005a,b). A related field, deserving further attention, is the analysis of rounding and heaping errors (Wolff and Augustin, 2003).

References

AMEMIYA, Y., FULLER, W. (1988). Estimation for the nonlinear functional relationship. *Annals of Statistics* **16** 147–160.

AUGUSTIN, T. (2002). *Survival Analysis under Measurement Error.* Habilitationsschrift (post-doctoral thesis). University of Munich.

AUGUSTIN, T. (2004). An exact corrected log-likelihood function for Cox's proportional hazards model under measurement error and some extensions. *Scandinavian Journal of Statistics* **31** 43–50.

AUGUSTIN, T., SCHWARZ, R. (2002). Cox's proportional hazards model under covariate measurement error – A review and comparison of methods. In *Total Least Squares and Errors-in-Variables Modeling: Analysis, Algorithms and Applications* (S. Van Huffel, P. Lemmerling, eds.), 175–184. Kluwer, Dordrecht.

AUGUSTIN, T., WOLFF, J. (2004). A bias analysis of Weibull models under heaped data. *Statistical Papers* **45** 211–229.

BENDER, R., AUGUSTIN, T., BLETTER, M. (2005). Simulating survival times for Cox regression models. *Statistics in Medicine* **24** 1713–1723.

BROWN, P. J. (1982). Multivariate calibration. *Journal of the Royal Statistical Society, Series B* **44** 287–321.

BRAND, R. (2002). Microdata protection through noise addition. In *Inference Control in Statistical Databases - From Theory to Practice.* (J. Doningo-Ferrer ed.), Lecture Notes in Computer Science 2316. Springer, Berlin.

BUZAS, J. S. (1998). Unbiased scores in proportional hazards regression with covariate measurement error. *Journal of Statistical Planning and Inference* **67** 247–257.

CARD, D. (2001). Estimating the return to schooling: Progress on some persistent econometric problems. *Econometrica* **69** 1127–1160.

CARROLL, R. J., RUPPERT, D., STEFANSKI, L. A. (1995). *Measurement Error in Nonlinear Models.* Chapman and Hall, London.

CARROLL, R. J., KÜCHENHOFF, H., LOMBARD, F., STEFANSKI, L. A. (1996). Asymptotics for the Simex estimator in structural measurement error models. *Journal of the American Statistical Association* **91** 242–250.

CHENG, C.-L., SCHNEEWEISS, H. (1998). Polynomial regression with errors in the variables. *Journal of the Royal Statistical Society, Series B* **60** 189–199.

CHENG, C.-L., SCHNEEWEISS, H., THAMERUS, M. (2000). A small sample estimator for a polynomial regression with errors in the variables. *Journal of the Royal Statistical Society, Series B* **62** 699–709.

CHENG, C.-L., SCHNEEWEISS, H. (2002). On the polynomial measurement error model. In *Total Least Squares and Errors-in-Variables Modeling* (S. van Huffel, P. Lemmerling, eds.), 131–143. Kluwer, Dordrecht.

CHENG, C.-L., VAN NESS, J. W. (1999). *Statistical Regression with Measurement Error.* Arnold, London.

COOK, J., STEFANSKI, L. A. (1994). Simulation-extrapolation estimation for parametric measurement error models. *Journal of the American Statistical Association* **89** 1314–1328.

COX, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B* **34** 187–220.

FLINN, C. J., HECKMAN, J. J. (1982). Models for the analysis of labor force dynamics. *Advances in Econometrics* **1** 35–95.

FRIEDMAN, M. (1957). *A Theory of the Consumption Function.* Princeton University Press, Princeton.

FULLER, W. A. (1987). *Measurement Error Models.* Wiley, New York.

GIMENEZ, P., BOLFARINE, H., COLOSIMO, E. A. (1999). Estimation in Weibull regression model with measurement error. *Communications in Statistics – Theory and Methods* **28** 495–510.

GLESER, L. J. (1990). Improvement of the naive estimation in nonlinear errors-in-variables regression. In *Statistical Analysis of Measurement Error Models and Application* (P. J. Brown, W. A. Fuller, eds.), *Contemporary Mathematics* **112** 99–114.

HAUSMAN, J. A., ABREVAYA, J., SCOTT-MORTON, F. M. (1998). Misclassification of the dependent variable in a discrete-response setting. *Journal of Econometrics* **87** 239–269.

HEID, I., KÜCHENHOFF, H., WELLMANN, J., GERKEN, M., KREIENBROCK, L. (2002). On the potential of measurement error to induce differential bias on odds ratio estimates: An example from radon epidemiology. *Statistics in Medicine* **21** 3261–3278.

HSIAO, C., WANG, Q. K. (2000). Estimation of structural nonlinear errors-in-variables models by simulated least-squares method. *International Economic Review* **41** 523–542.

HU, P., TSIATIS A. A., DAVIDIAN M. (1998). Estimating the parameters in the Cox model when covariate variables are measured with error. *Biometrics* **54** 1407–1419.

HU, Y., RIDDER, G. (2005). *Estimating a nonlinear model with measurement error using marginal information.* http://www-rcf.usc.edu/ ~ridder / Wpapers/ EIV-marg-final.pdf.

HUANG, H. S., HUWANG, L. (2001). On the polynomial structural relationship. *The Canadian Journal of Statistics* **29** 493-511.

HUANG, Y., WANG, C. Y. (2000). Cox regression with accurate covariates unascertainable: A nonparametric-correction approach. *Journal of the American Statistical Association* **95** 1209–1219 (Correction: **98** 779).

HUGHES, M. D. (1993). Regression dilution in the proportional hazards model. *Biometrics* **49** 1056–1066.

KONG, F. H., GU, M. (1999). Consistent estimation in Cox proportional hazards model with covariate measurement errors. *Statistica Sinica* **9** 953–969.

KONG, F. H., HUANG, W., LI, X. (1998). Estimating survival curves under proportional hazards model with covariate measurement errors. *Scandinavian Journal of Statistics* **25** 573–587.

KÜCHENHOFF, H., (1992). Estimation in generalized linear models with covariate measurement error using the theory of misspecified models. In *Statistical Modelling* (P. van der Heijden, W. Jansen, B. Francis, G. Seeber, eds.), 185–193. Elsevier, Amsterdam.

KÜCHENHOFF, H. (1995). The identification of logistic regression models with errors in the variables. *Statistical Papers* **36** 41–48.

KÜCHENHOFF, H., BENDER, R., LANGER, I., LENZ-TÖNJES, R. (2003). Effect of Berkson measurement error on parameter estimates in Cox regression models. Discission Paper 346, Sonderforschungsbereich 386, University of Munich.

KÜCHENHOFF, H., CARROLL, R. J. (1997). Segmented regression with errors in predictors: semiparametric and parametric methods. *Statistics in Medicine* **16** 169–188.

KÜCHENHOFF, H., MWALILI, S., LESAFFRE, E. (2005). A general method for dealing with misclassification in regression: The misclassification SIMEX. *Biometrics* (to appear).

KUHA, J. T., TEMPLE, J. (2003). Covariate measurement error in quadratic regression. *International Statistical Review* **71** 131–150.

KUKUSH, A., MARKOVSKY, I., VAN HUFFEL, S. (2004). Consistent estimation in an implicit quadratic measurement error model. *Computational Statistics & Data Analysis* **47** 123–147.

KUKUSH, A., SCHNEEWEISS, H., WOLF, R. (2001). Comparison of three estimators in a polynomial regression with measurement errors. Discussion Paper 233, Sonderforschungsbereich 386, University of Munich.

KUKUSH, A., SCHNEEWEISS, H. (2005). Comparing different estimators in a nonlinear measurement error model. I and II. *Mathematical Methods of Statistics* **14** 53–79 and 203–223.

LI, T. (2000). Estimation of nonlinear errors-in-variables models: A simulated minimum distance estimator. *Statistics and Probability Letters* **47** 243–248.

NAKAMURA, T. (1990). Corrected score functions for errors-in-variables models: Methodology and application to generalized linear models. *Biometrika* **77** 127–137.

NAKAMURA, T. (1992). Proportional hazards model with covariates subject to measurement error. *Biometrics* **48** 829–838.

NOWAK, E. (1993). The identification of multivariate linear dynamic error-in-variables models. *Journal of Econometrics* **59** 213–227.

PEPE, M. S., SELF, M. S., PRENTICE, R. L. (1989). Further results in covariate measurement errors in cohort studies with time to response data. *Statistics in Medicine* **8** 1167–1178.

PRENTICE, R. L. (1982). Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika* **69** 331–342.

RONNING, G. (2005). Randomized response and the binary probit model. *Economics Letters* **86** 221–228.

SCHMID, M., SCHNEEWEISS, H., KÜCHENHOFF, H. (2005a). Consistent estimation of a simple linear model under microaggregation. Discussion Paper 415, Sonderforschungsbereich 386, University of Munich.

SCHMID, M., SCHNEEWEISS, H., KÜCHENHOFF, H. (2005b). Statistical inference in a simple linear model under microaggregation. Discussion Paper 416, Sonderforschungsbereich 386, University of Munich.

SCHNEEWEISS, H., CHENG, C.-L. (2006). Bias of the structural quasi–score estimator of a measurement error model under misspecification of the regressor distribution. *Journal of Multivariate Analysis* **97** 455–473.

SCHNEEWEISS, H., MITTAG, H. J. (1986). *Lineare Modelle mit fehlerbehafteten Daten*. Physica, Heidelberg.

SCHUSTER, G. (1998). ML estimation from binomial data with misclassifications - a comparison: Internal validation versus repeated measurements. In *Econometrics in Theory and Practice* (R. Galata, H. Küchenhoff, eds.), 45–58. Physika, Heidelberg.

SHKLYAR, S., SCHNEEWEISS, H. (2005). A comparison of asymptotic covariance matrices of three consistent estimators in the Poisson regression model with measurement errors. *Journal of Multivariate Analysis* **94** 250–270.

SHKLYAR, S., SCHNEEWEISS, H., KUKUSH, A. (2005). Quasi score is more efficient than corrected score in a polynomial measurement error model. Discussion Paper 445, Sonderforschungsbereich 386, University of Munich.

SKINNER, C. J., HUMPHREYS, K. (1999). Weibull regression for lifetimes measured with error. *Lifetime Data Analysis* **5** 23–37.

STEFANSKI, L. A. (1989). Unbiased estimation of a nonlinear function of a normal mean with application to measurement error models. *Communications in Statistics - Theory and Methods* **18** 4335–4358.

THAMERUS, M. (2003). Fitting a mixture distribution to a variable subject to heteroscedastic measurement errors. *Computational Statistics* **18** 1–17.

TSIATIS A., DAVIDIAN, M. (2004). Joint modeling of longitudinal and time-to-event data: An overview. *Statistica Sinica* **14** 809–834.

WANSBEEK, T., MEIJER, E. (2000). *Measurement Error and Latent Variables in Econometrics*. Elsevier, Amsterdam.

WOLF, R. (2004). *Vergleich von funktionalen und strukturellen Messfehlerverfahren.* Logos Verlag, Berlin.

WOLFF, J., AUGUSTIN, T. (2003). Heaping and its consequences for duration analysis - a simulation study. *Allgemeines Statistisches Archiv – Journal of the German Statistical Society* **87** 1–28.

14 The Microeconometric Estimation of Treatment Effects - An Overview*

Marco Caliendo[1] and Reinhard Hujer[2]

[1] DIW Berlin, Abteilung Staat
mcaliendo@diw.de
[2] Institut für Statistik und Ökonometrie, J.W.Goethe Universität
hujer@wiwi.uni-frankfurt.de

Summary: The need to evaluate the performance of active labour market policies is not questioned any longer. Even though OECD countries spend significant shares of national resources on these measures, unemployment rates remain high or even increase. We focus on microeconometric evaluation which has to solve the fundamental evaluation problem and overcome the possible occurrence of selection bias. When using non-experimental data, different evaluation approaches can be thought of. The aim of this paper is to review the most relevant estimators, discuss their identifying assumptions and their (dis-)advantages. Thereby we will present estimators based on some form of exogeneity (selection on observables) as well as estimators where selection might also occur on unobservable characteristics. Since the possible occurrence of effect heterogeneity has become a major topic in evaluation research in recent years, we will also assess the ability of each estimator to deal with it. Additionally, we will also discuss some recent extensions of the static evaluation framework to allow for dynamic treatment evaluation.

14.1 Introduction

The need to evaluate the performance of active labour market policies (ALMP) is not questioned any longer. Even though OECD countries spend significant shares of national resources on these measures, unemployment rates remain high or even increase. The ideal evaluation process can be looked at as a series of three steps (Fay, 1996): First, the impacts of the programme on the individual should be estimated (MICROECONOMETRIC EVALUATION). Second, it should be examined if the impacts are large enough to yield net social gains (MACROECONOMIC EVALUATION). Third, it

*The authors thank Stephan L. Thomsen, Christopher Zeiss and one anonymous referee for valuable comments. The usual disclaimer applies.

should be answered if this is the best outcome that could have been achieved for the money spent (COST-BENEFIT ANALYSIS). In this paper we focus on the first step. The main question in microeconometric evaluation is if the outcome for an individual is affected by the participation in an ALMP programme or not. We would like to know the difference between the value of the participant's outcome in the actual situation and the value of the outcome if he had not participated in the programme. The fundamental evaluation problem arises because we can never observe both states (participation and non-participation) for the same individual at the same time, i. e. one of the states is counterfactual. Therefore finding an adequate control group and solving the problem of selection bias is necessary to make a comparison possible.

Depending on the data at hand, different evaluation strategies can be thought of. Since in most European countries - unlike in the US - experimental data are not available, researchers have to use non-experimental data. A lot of methodological progress has been made to develop and justify non-experimental evaluation estimators which are based on econometric and statistical methods to solve the fundamental evaluation problem (see e. g. Heckman *et al.*, 1999). The aim of this paper is to give an overview of the most relevant evaluation approaches and provide some guidance on how to choose between them. Thereby we will also discuss the possible occurrence of effect heterogeneity, which has become a major focus of evaluation research in the last years, and the ability of each estimator to deal with it.

Two broad categories of estimators can be distinguished according to the way selection bias is handled. The first category contains approaches that rely on the so-called unconfoundedness or selection on observables assumption. If one believes that the available data is not rich enough to justify this assumption, one has to rely on the second category of estimators which explicitly allows selection on unobservables, too. We will discuss different approaches for both situations in Section 14.3 where we also present some recent extensions of the static evaluation framework to dynamic concepts. Before we do so, we are going to introduce the evaluation framework in Section 14.2, where we especially present the potential outcome approach, discuss parameters of interest, selection bias on observable and on unobservable characteristics as well as heterogeneous treatment effects. Finally, Section 14.4 concludes.

14.2 The Evaluation Framework

14.2.1 Potential Outcome Approach and the Fundamental Evaluation Problem

Inference about the impact of a treatment on the outcome of an individual involves speculation about how this individual would have performed in the labour market, if he had not received the treatment. The framework serving as a guideline for the empirical analysis of this problem is the potential outcome approach, also known as the Roy (1951) – Rubin (1974) – model.

The main pillars of this model are individuals, treatment (participating in a pro-

gramme or not) and potential outcomes, that are also called responses.[1] In the basic model there are two potential outcomes (Y^1, Y^0) for each individual, where Y^1 indicates a situation with treatment and Y^0 without. To complete the notation, we additionally denote variables that are unaffected by treatment by X. Attributes X are exogenous in the sense that their potential values for different treatment states coincide (Holland, 1986). Furthermore we define a binary assignment indicator D, indicating whether an individual actually received treatment ($D = 1$), or not ($D = 0$). The treatment effect for each individual i is then defined as the difference between his potential outcomes:

$$\Delta_i = Y_i^1 - Y_i^0. \tag{14.1}$$

The fundamental problem of evaluating this individual treatment effect arises because the observed outcome for each individual is given by:

$$Y_i = D_i Y_i^1 + (1 - D_i) Y_i^0. \tag{14.2}$$

This means that for those individuals who participated in treatment we observe Y^1 and for those who did not participate we observe Y^0. Unfortunately, we can never observe Y^1 and Y^0 for the same individual simultaneously and therefore we cannot estimate (14.1) directly. The unobservable component in (14.1) is called the counterfactual outcome.

Concentration on a single individual requires that the effect of the intervention on each individual is not affected by the participation decision of any other individual, i. e. the treatment effect Δ_i for each person is independent of the treatment of other individuals. In statistical literature this is referred to as the stable unit treatment value assumption (SUTVA)[2] and guarantees that average treatment effects can be estimated independently of the size and composition of the treatment population. In particular, it excludes peer-effects as well as cross-effects and general equilibrium effects (Sianesi, 2004).

14.2.2 Treatment Effects and Selection Bias

Since there will never be an opportunity to estimate individual effects in (14.1) with confidence, we have to concentrate on population averages of gains from treatment. Two treatment effects are dominantly used in empirical studies. The first one is the (population) average treatment effect (ATE)

$$\Delta_{ATE} = E(\Delta) = E(Y^1) - E(Y^0), \tag{14.3}$$

which answers the question which would be the outcome if individuals in the population were randomly assigned to treatment. The most frequently used parameter is the so called average treatment effect on the treated (ATT) and focusses explicitly on the effects on those for whom the programme is actually intended. It is given by

$$\Delta_{ATT} = E(\Delta \mid D = 1) = E(Y^1 \mid D = 1) - E(Y^0 \mid D = 1). \tag{14.4}$$

[1] It should be clear, that this framework is not restricted to the evaluation of labour market programmes. It applies for every situation where one group of units, e. g. individuals or firms or other entities, receive some form of treatment and others do not.

[2] See Holland (1986) for a further discussion of this concept.

In the sense that this parameter focuses directly on participants, it determines the realised gross gain from the programme and can be compared with its costs, helping to decide whether the programme is successful or not (Heckman *et al.*, 1999). Given Equation (14.4), the problem of selection bias can be straightforwardly seen since the second term on the right hand side is unobservable as it describes the hypothetical outcome without treatment for those individuals who received treatment. Since with non-experimental data the condition $E(Y^0 \mid D = 1) = E(Y^0 \mid D = 0)$ is usually not satisfied, estimating ATT by the difference in subpopulation means of participants $E(Y^1 \mid D = 1)$ and non-participants $E(Y^0 \mid D = 0)$ will lead to a selection bias. This bias arises because participants and non-participants are selected groups that would have different outcomes, even in absence of the programme. It might be caused by observable or unobservable factors.

14.2.3 Potential Outcome Framework and Heterogeneous Treatment Effects

For the further discussion it will be helpful to relate the potential outcome framework to familiar econometric notation. To do so, we follow Blundell and Costa Dias (2002) and define the following outcome equations

$$Y_{it}^1 = g_t^1(X_i) + U_{it}^1 \quad \text{and} \quad Y_{it}^0 = g_t^0(X_i) + U_{it}^0\,, \tag{14.5}$$

where the subscripts i and t index the individual and the time period, respectively. The functions g^0 and g^1 represent the relationship between potential outcomes and the set of observable characteristics. U^0 and U^1 are error terms which have zero mean and are assumed to be uncorrelated with regressors X. For the familiar case of linear regression, the g functions specialise to $g^1(X) = X\beta_1$, and $g^0(X) = X\beta_0$.

Heckman and Robb (1985) note that the decision to participate in treatment may be determined by a prospective participant, by a programme administrator, or both. Whatever the specific content of the rule, it can be described in terms of an index function framework. Let IN_i be an index of benefits to the relevant decision maker from participating in the programme. It is a function of observed (Z_i) and unobserved (V_i) variables. Therefore

$$IN_i = f(Z_i) + V_i\,, \tag{14.6}$$

with enrolment in the programme D_i given by

$$D_i = \begin{cases} 1 & \text{if} \quad IN_i > 0\,, \\ 0 & \text{otherwise}\,. \end{cases}$$

Under this specification and the further assumption that treatment takes place in period k, one can define the individual-specific treatment effect for any X_i as

$$\Delta_{it}(X_i) = Y_{it}^1 - Y_{it}^0 = [g_t^1(X_i) - g_t^0(X_i)] + [U_{it}^1 - U_{it}^0] \quad \text{with } t > k\,. \tag{14.7}$$

The ATT measured in the post-treatment period $t > k$ is then defined as

$$\Delta_{ATT} = E(\Delta_{it} \mid D_i = 1). \tag{14.8}$$

The assignment process to treatment is most probably not random. Consequently, the assignment process will lead to non-zero correlation between enrolment (D_i) and the outcome's error term (U^1, U^0). This may occur because of stochastic dependence between (U^1, U^0) and V_i in (14.6) or because of stochastic dependence between (U^1, U^0) and Z_i. In the former case we have selection on unobservables, whereas in the latter case selection on observables is prevalent (Heckman and Robb, 1985).

We can use this discussion to highlight the problem of heterogeneous treatments, i. e. situations where the impact of a programme differs across individuals, in a common and intuitive way.[3] If treatment impacts vary across individuals this may come systematically through the observables' component or be part of the unobservables and we can re-write equation (14.5) as

$$Y_{it} = g_t^0(X_i) + \Delta_t(X_i)D_{it} + [U_{it}^0 + D_{it}(U_{it}^1 - U_{it}^0)], \tag{14.9}$$

where

$$\Delta_t(X_i) = E[\Delta_{it}(X_i)] = g_t^1(X_i) - g_t^0(X_i) \tag{14.10}$$

is the expected treatment effect at time t for individuals characterised by X_i (Blundell and Costa Dias, 2002). Since abandoning the assumption of homogeneous treatment effects and identifying the individuals that benefit from programmes provides some scope to improve their future efficiency, we will assess for each estimation method that we will present in the following its capability to deal with heterogeneous treatment effects.

14.3 Non-Experimental Evaluation Methods

The discussion in Subsections 14.2.2 and 14.2.3 has made clear that the problem of selection bias is a severe one and cannot be solved with more data, since the fundamental evaluation problem will not disappear. We have a distorted representation of a true population in a sample as a consequence of a sampling rule, which is the essence of the selection problem (Heckman, 2001). Hence, we have to use some identifying assumptions to draw inference about the hypothetical population based on the observed population. In the following subsections we will present several evaluation approaches. Each approach invokes different identifying assumptions to construct the required counterfactual outcome. We will start the discussion with two estimators (matching and regression) that are based on the selection on observables assumption.[4] Following that we introduce three estimators that allow for selection on unobservables, too, namely difference-in-differences, instrumental variables and selection models. Finally, we also briefly discuss regression discontinuity models and the estimation of treatment effects in a dynamic framework.

[3] See e. g. the discussion in Smith (2000).

[4] See Imbens (2004) for an extensive overview of estimating average treatment effects under unconfoundedness.

14.3.1 Matching Estimator

Matching is based on the identifying assumption that conditional on some covariates X, the outcome Y is independent of D.[5] In the notation of Dawid (1979) this is

ASSUMPTION 1 *Unconfoundedness:* $Y^0, Y^1 \amalg D \mid X$,

where $\amalg$ denotes independence. If Assumption 1 is true, then $F(Y^0 \mid X, D = 1) = F(Y^0 \mid X, D = 0)$ and $F(Y^1 \mid X, D = 1) = F(Y^1 \mid X, D = 0)$. This means, that conditionally on X, non-participant outcomes have the same distribution that participants would have experienced if they had not participated in the programme and vice versa (Heckman *et al.*, 1997). Similar to randomisation in a classical experiment, matching balances the distributions of all relevant, pre-treatment characteristics X in the treatment and comparison group.[6] Thus it achieves independence between the potential outcomes and the assignment to treatment.

ASSUMPTION 2 *Overlap:* $0 < P(D = 1 \mid X) < 1$, *for all* X.

This implies that the support of X is equal in both groups, i. e. $S = \text{Support}(X|D = 1) = \text{Support}(X|D = 0)$. Assumption 2 prevents X from being a perfect predictor in the sense that we can find for each participant a counterpart in the non-treated population and vice versa. If there are regions where the support of X does not overlap for the treated and non-treated individuals, matching has to be performed over the common support region only. The estimated effects have then to be redefined as the mean treatment effect for those individuals falling within the common support (Blundell *et al.*, 2004). Rosenbaum and Rubin (1983) call Assumptions 1 and 2 together 'strong ignorability' under which ATT and ATE can be defined for all values of X. If one is interested in ATT only, it is sufficient to assume $Y^0 \amalg D \mid X$ and the weaker overlap Assumption $P(D = 1 \mid X) < 1$. The mean impact of treatment on the treated can be written as

$$\Delta_{ATT}^{MAT} = E(Y^1|X, D = 1) - E_X[E(Y^0|X, D = 0)|D = 1], \qquad (14.11)$$

where the first term can be estimated from the treatment group and the second term from the mean outcomes of the matched comparison group. The outer expectation is taken over the distribution of X in the treated population. The method of matching can also be used to estimate ATT at some points $X = x$, where x is a particular realisation of X. Two things have to be mentioned: First, it should be clear that conditioning on all relevant covariates is limited in case of a high dimensional vector X. For that case Rosenbaum and Rubin (1983) suggest the use of so-called balancing

[5] These are the covariates which also appear in Z as defined in Equation (14.6).

[6] If we say relevant we mean all those covariates that influence the assignment to treatment as well as the potential outcomes.

scores to overcome this dimensionality problem.[7] Second, there are several different matching algorithms suggested in the literature, e. g. kernel or nearest-neighbour matching, and the choice between them is not trivial since it involves a trade-off between bias and variance (see Smith and Todd, 2005, for an overview).

14.3.2 Linear Regression Approach

Even though regression and matching both rely on the unconfoundedness assumption, there are some key differences between both approaches which are worth to be discussed. One key difference is that matching, due to its non-parametric nature, avoids functional form assumptions implicit in linear regression models. The potential outcomes in a linear regression framework can be written as $Y^1 = X\beta_1 + U^1$ and $Y^0 = X\beta_0 + U^0$ and ATT under regression is given by[8]:

$$\Delta_{ATT}^{Reg} = E(Y^1 - Y^0|X, D = 1) = X(\beta_1 - \beta_0) + E(U^1 - U^0|X, D = 1). \tag{14.12}$$

The identifying assumption needed to justify regression under unconfoundedness is analogue to Assumption 1 and can be re-written as:

ASSUMPTION 3 *Unconfoundedness in Regression:* $U^0, U^1 \amalg D \mid X$.

In the matching framework, the goal is to set the bias $B(X) = 0$ which basically only requires that the mean of the error terms in the treatment group given a covariate cell X equals the corresponding mean in the control group, that is $B(X) = E(U^1|X, D = 1) - E(U^0|X, D = 0) = 0$. This means that it is possible to match on variables that are correlated with the error term in the outcome equation (Hui and Smith, 2002). In the regression framework, however, we need to eliminate the dependence between (U^0, U^1) and X, that is $E(U^1|X, D = 1) = E(U^0|X, D = 0) = 0$ (Heckman *et al.*, 1998). Of course, as Smith (2000) notes, the difference between both approaches fades with the inclusion of a sufficient number of higher-order and interaction terms in the regression. However, not only is such an inclusion not very common in practice, it is also not straightforward to choose these terms. Moreover, whereas matching estimators do rely on the common support assumption, regression estimators do not and will produce estimates even in the absence of similar comparison units, since the linear functional form assumption fills in for the missing data (Smith, 2004). Another key difference between regression and matching is the way both approaches handle heterogeneous treatment effects. As Lechner (2002) notes, the non-parametric matching approach leaves the individual causal effect unrestricted and allows individual effect heterogeneity in the population. This is not true for the regression approach which will not recover ATT, although, at times it might provide a close approximation as shown by Angrist (1998) and Blundell *et al.* (2004).

[7] One possible balancing score is the propensity score. See Rosenbaum (2002) or Caliendo and Kopeinig (2005) for an introduction into propensity score matching estimators and some guidance for their implementation.

[8] For notational convenience we drop individual subscript i and time subscript t.

14.3.3 Instrumental Variables Estimator

Let us now turn to estimators that account for selection on unobservables, too. We will start with the method of instrumental variables (IV). Its underlying identification strategy is to find a variable which determines treatment participation but does not influence the outcome equation. The instrumental variable affects the observed outcome only indirectly through the participation decision and hence causal effects can be identified through a variation in this instrumental variable. IV methods are extensively discussed in Imbens and Angrist (1994) and Angrist *et al.* (1996) among others. In terms of the discussion in Subsection 14.2.3, IV requires the existence of at least one regressor to the decision rule, Z^*, that satisfies the following three conditions (Blundell and Costa Dias, 2000):

- Z^* determines programme participation. For that to be true, it has to have a non-zero coefficient in the decision rule in Equation (14.6).
- We can find a transformation, s, such that $s(Z^*)$ is uncorrelated with the error terms (U^1, V) and (U^0, V), given the exogenous variables X.
- Z^* is not completely determined by X.

The variable Z^* is then called the instrument. In providing variation that is correlated with the participation decision but does not affect potential outcomes from treatment directly, it can be used as a source of exogenous variation to approximate randomised trials (Blundell and Costa Dias, 2000).

Clearly, a major problem with this estimator is to find a good instrument. In the treatment evaluation problem it is hard to think of variables that satisfy all three above mentioned assumptions. The difficulty lies mainly in the simultaneous requirement that the variable has to predict participation but does not influence the outcome equation. As pointed out by Blundell and Costa Dias (2000), a second drawback arises when considering the heterogeneous treatment framework. Recall that the error term from Equation (14.9) in Subsection 14.2.3 is given by $[U_{it}^0 + D_{it}(U_{it}^1 - U_{it}^0)]$. Even if Z^* is uncorrelated with U_{it}, the same cannot be true by definition for $U_{it}^0 + D_{it}(U_{it}^1 - U_{it}^0)$ since Z^* determines D_i by assumption. The violation of this assumption invalidates the application of IV methodology in a heterogeneous framework (Blundell and Costa Dias, 2000). However, in this situation it might still be possible to provide a potentially interesting parameter of the IV estimation - called local average treatment effect (LATE) by Imbens and Angrist (1994). This estimator identifies the treatment effect for those individuals (with characteristics X) who are induced to change behaviour because of a change in the instrument.[9] It should be clear that each instrument implies its own LATE, and LATEs for two different instruments may differ substantially depending on the impacts realised by the persons each instrument induces to participate (Hui and Smith, 2002).

[9] Additionally to those assumptions already made, we further have to assume that the instrument has the same directional effect on all those whose behaviour it changes. This assumption rules out the co-existence of defiers and compliers and is known as 'monotonicity assumption' (Imbens and Angrist, 1994).

14.3.4 Selection Model

This method is also known as the Heckman selection estimator (Heckman, 1978). It is more robust than the IV method but also more demanding in the sense that it imposes more assumptions about the structure of the model. Two main assumptions are required (Blundell and Costa Dias, 2000):

- There has to be one additional regressor in the decision rule which has a non-zero coefficient and which is independent of the error term V.
- Additionally, the joint density of the distribution of the errors U_{it} and V_i has to be known or can be estimated.

The basic idea of this estimator is to control directly for the part of the error term in the outcome equation that is correlated with the participation dummy variable. It can be seen as a two-step-procedure. First, the part of the error term U_{it} that is correlated with D_i is estimated. Second, this term is then included in the outcome equation and the effect of the programme is estimated. By construction, the remains of the error term in the outcome equation are not correlated with the participation decision any more (Blundell and Costa Dias, 2000).[10]

The Heckman selection estimator is not without critique, which rests mainly on the following point (see e. g. Puhani, 2000): If there are no exclusion restrictions, the models are identified only by assumptions about functional form and error distributions. This may lead to large standard errors and results that are very sensitive to the particular distributional assumptions invoked. This point of criticism is very closely related to the problem of finding a good instrument as described for the IV method. In fact, in a recent paper Vytlacil (2002) shows that the identifying assumptions for the selection model are equivalent to those invoked by Imbens and Angrist (1994) in the linear instrumental variables context.

14.3.5 Difference-in-Differences Estimator

The difference-in-differences (DID) estimator requires access to longitudinal data and forms simple averages over the group of participants and non-participants between pre-treatment period t' and post-treatment period t, that is, changes in the outcome variable Y for treated individuals are contrasted with the corresponding changes for non-treated individuals (Heckman *et al.*, 1998):

$$\Delta^{DID} = [Y_t^1 - Y_{t'}^0 \mid D = 1] - [Y_t^0 - Y_{t'}^0 \mid D = 0]\,. \tag{14.13}$$

The identifying assumption of this method is

$$E(Y_t^0 - Y_{t'}^0 \mid D = 1) = E(Y_t^0 - Y_{t'}^0 \mid D = 0)\,. \tag{14.14}$$

The DID estimator is based on the assumption of time-invariant linear selection effects, so that differencing the differences between participants and non-participants eliminates the bias (Heckman *et al.*, 1998). To make this point clear, we can re-write

[10] Blundell and Costa Dias (2000) also show that this approach is capable of identifying ATT if effects are assumed to be heterogeneous.

the outcome for an individual i at time t as $Y_{it} = \pi_{it} + D_{it} \cdot Y_{it}^1 + (1 - D_{it}) \cdot Y_{it}^0$, where π_{it} captures the effects of selection on unobservables. The validity of the DID estimator then relies on the assumption $\pi_{it} = \pi_{it'}$, where it is not required that the bias vanishes completely, but that it remains constant (Heckman *et al.*, 1998). One problem when using DID is Ashenfelter's dip, i. e. a situation where shortly before participation in an ALMP programme the employment situation of future participants deteriorates (Ashenfelter, 1978). If the 'dip' is transitory and the dip is eventually restored even in the absence of participation in the programme, the bias will not average out. To allow a more detailed discussion, Blundell and Costa Dias (2002) further decompose π_{it} in three parts: an individual-specific fixed effect, a common macroeconomic effect and a temporary individual-specific effect. Clearly, for the DID to be unbiased it is sufficient that selection into treatment is independent from the temporary individual-specific effect, since the other two effects vanish in the sequential differences. They also discuss the case where the macroeconomic effect has a differential impact across the group of participants and non-participants. This may happen when both groups differ on unobserved characteristics which make them react differently to macroeconomic shocks. To overcome this problem they propose a differential trend adjusted DID estimator (Blundell and Costa Dias, 2002). Heckman *et al.* (1998) combine the DID approach with the already presented matching estimator by comparing the before-after outcome of participants with those of matched non-participants. Smith and Todd (2005) show that this 'conditional DID estimator' is more robust than traditional cross-section matching estimators, as it allows for selection on observables as well as time-invariant selection on unobservables.

14.3.6 Regression Discontinuity Model

The regression discontinuity model (RDM) can be seen as a particular type of instrumental variable identification strategy. It uses discontinuities in the selection process to identify causal effects. In this model, treatment depends on some observed variable, Z, according to a known, deterministic rule, such as $D = 1$ if $Z > \overline{Z}$ and $D = 0$ otherwise (Heckman *et al.*, 1999). The variable Z has direct impact on Y as well as an indirect impact on Y through D. This indirect impact is the causal effect we would like to identify. Frölich (2002) notes that this effect is identified if the direct and indirect impacts of Z on Y can be separated.

There are several things to note about RDM (see e. g. Heckman *et al.*, 1999). First, it is assumed that selection is on observable characteristics only. Second, it should be clear that there is no common support for participants and non-participants making matching impossible. Hence, RDM takes over when there is selection on observables (here: the deterministic rule) but the overlapping support condition required for matching breaks down (with a certain Z you either belong to the participant or the non-participant group). Finally, the selection rule is assumed to be deterministic and known and that variation in the relevant variable Z is exogenous.

14.3.7 Dynamic Evaluation Concepts

Sequential Matching Estimators. What we have discussed so far is basically a static evaluation framework where an individual can participate in one programme (or not). A recent extension of this framework for matching estimators considers the case, where individuals can participate in subsequent treatments. Lechner and Miquel (2002) discuss identifying assumptions for so-called sequential matching estimators. These estimators mimic the matching estimators described above but allow to estimate effects in a dynamic causal model. Their framework can be made clear in a three-periods-two-treatments model. We follow the discussion in Lechner (2004) and present the needed additional notation in the following. First, we introduce a time index $t \in \{0,1,2\}$ and extend the treatment indicator D by this time index, that is $D = (D_0, D_1, D_2)$. It is further assumed that in period 0 everybody is in the same treatment state $D_0 = 0$, whereas from the second period on D_t can take two values. Realisations of D_t are denoted by $d_t \in \{0,1\}$. So in period 1 an individual is observed in exactly one of these two treatments $(0,1)$, whereas in period 2 an individual participates in one of four possible treatment sequences $\{(0,0),(1,0),(0,1),(1,1)\}$. Additionally, the history of variables up to period t are denoted by a bar below a variable, e. g. $\underline{d}_2 = (d_1, d_2)$. The potential outcomes are indexed by treatments and the time period, i. e. $Y^{\underline{s}_t} = (Y_0^{\underline{d}_t}, Y_1^{\underline{d}_t}, Y_2^{\underline{d}_t})$. The observed outcomes are given by the following equation

$$Y_t = D_1 Y_t^1 + (1 - D_1) Y_t^0 = D_1 D_2 Y_t^{1,1} + D_1 (1 - D_2) Y_t^{1,0} + (1 - D_1) D_2 Y_t^{0,1} + (1 - D_1)(1 - D_2) Y_t^{0,0} . \quad (14.15)$$

As in the static model, variables that influence treatment selection and potential outcomes are called attributes and are denoted by X. An important distinction has to be made regarding the exogeneity of these variables. Whereas in the static model exogeneity is assumed, in the dynamic model the X-variables in later periods can be influenced by treatment realisations. Hence, there are potential values of these variables as well: $X^{d_t} = (X_0^{d_t}, X_1^{d_t}, X_2^{d_t})$, where e. g. $X_1^{d_1}$ may contain $Y_1^{d_1}$ or functions of it. The sequential matching framework is a powerful tool and is applicable for situations where individuals can participate more than once in a programme and where it is possible to identify treatment sequences.

Duration Models. Another methodology for modelling dynamically assigned treatments is the application of duration models (Abbring and van den Berg, 2003). In these models not only the information if an individual participates in a programme is considered, but also the timing of the treatment within the unemployment spell. To introduce the notation we normalise the point in time when an individual enters unemployment to zero, denote the duration until the individual enters regular employment with T_e and the duration until the individual enters a programme with T_p (realisations are denoted by t_u and t_p, respectively). Both durations are assumed to vary with observable characteristics x and unobservable characteristics v_e and v_p. Abbring and van den Berg (2003) assume that the realisation t_p affects the distribution of T_e in a deterministic way from t_p onwards. For the specification of the hazard rates a mixed proportional hazard model is used.

Basic feature of this model is that the duration dependence, observable covariates and unobservable components enter the hazard rate multiplicatively:

$$\theta_e(t|t_p, x, v_e) = \lambda_e(t) \exp[x'\beta_e + \mu(t - t_p)I(t > t_p) + v_e]\,. \tag{14.16}$$

The hazard rate for the transition into regular employment θ_e consists of the baseline hazard $\lambda_e(t)$ that determines the duration dependence, the systematic part $\exp(x'\beta_e)$ and the unobserved heterogeneity term $\exp(v_e)$. The treatment effect $\exp[\mu(t - t_p)I(t > t_p)]$ with $I(t > t_p)$ as an indicator function taking the value 1 if $t > t_p$ is specified as a function of the difference $t - t_p$. In general, the treatment effect is allowed to vary over time after the treatment has started and can be interpreted as a shift of the hazard rate by $\exp(\mu(t - t_p))$. The transition rate from unemployment into programmes θ_p is analogously specified as a mixed proportional hazard model:

$$\theta_p(t|x, v_p) = \lambda_p(t) \exp[x'\beta_p + v_p]\,. \tag{14.17}$$

Identifying the treatment effect requires to consider selectivity which is present if individuals with a relatively high transition rate into employment also have a relatively high transition into programme participation (Abbring and van den Berg, 2003). In this case we obviously would observe a positive correlation between v_e and v_p and the joint distribution $G(v_e, v_p)$ has to be specified. Abbring and van den Berg (2003) show that the bivariate model (14.16) and (14.17) and especially the treatment effect is nonparametrically identified, since no parametric assumptions with respect to the baseline hazard and the unobserved heterogeneity distribution are required. Furthermore the identification does not require exclusion restrictions on x which are often hardly to justify from a theoretical point of view.[11]

Matching with Time-Varying Treatment Indicators. An alternative concept of modelling dynamic treatment effects is presented by Fredriksson and Johansson (2004) and Sianesi (2004). They introduce a non-parametric matching estimator that takes the timing of events into account but does not rely on proportionality assumptions. An important topic in this framework is the choice of an appropriate control group. Instead of defining control individuals as those who never participate, Sianesi (2004) defines control individuals as those who did not participate until a certain time period. Fredriksson and Johansson (2004) formalise her approach and argue that the standard way of defining a control group, i. e. those individuals who never participated in a given time interval, might lead to biased results, because the unconfoundedness assumption might be violated as the treatment indicator itself is defined conditional on future outcomes. Following Sianesi (2004), the key choice faced by the unemployed in this framework is not whether to participate at all, but whether to participate in a programme or not now. In the latter case, the individual searches longer in open unemployment. The corresponding parameter of interest in this setting is then defined as the effect of joining a programme now in contrast to waiting longer. The population of interest at time u are those still openly unemployed after u months. Treatment receipt in u is denoted by $D^{(u)} = 1$. The comparison group consists of all persons who do not join at least

[11] It should be noted that anticipatory programme effects are ruled out in the above mentioned specification (Abbring and van den Berg, 2003).

up to u, denoted by $D^{(u)} = 0$. The outcome of interest is defined over time t and is given by $Y_t^{(u)}$. The potential outcome if an individual joins in u is denoted by $Y_t^{1(u)}$ and if he does not join at least up to u by $Y_t^{0(u)}$. For each point of elapsed unemployment duration the parameter of interest is

$$\Delta_u^t = E(Y_t^{1(u)} - Y_t^{0(u)}|D^{(u)} = 1) = E(Y_t^{1(u)}|D^{(u)} = 1) \\ -E(Y_t^{0(u)}|D^{(u)} = 1), \quad \text{for} \quad t = u, u+1, \ldots, T. \tag{14.18}$$

This is the average impact at time t, for those joining a programme in their u^{th} month of unemployment compared to waiting longer in open unemployment. Sianesi (2004) notes that the treatment effects are based on a comparison of individuals who have reached the same elapsed duration of unemployment. Measurement starts at time u, the start of the programme and therefore possible locking-in effects might encounter. The second term on the right hand side of (14.18) is not identified and the CIA needed in that case is given by

$$Y_t^{0(u)} \amalg D^{(u)}|X = x \quad \text{for} \quad t = u, u+1, \ldots, T, \tag{14.19}$$

which means that given a set of observed characteristics X, the counterfactual distribution of $Y_t^{0(u)}$ for individuals joining in u is the same as for those not joining in u and waiting longer. The estimated treatment effect is then the effect for those who participate in a programme at some time in their unemployment spell instead of waiting longer. Even though this is not a standard evaluation parameter of interest, it still shows whether a programme was effective or not.

14.4 Summary - Which Estimator to Choose?

We have presented several different evaluation strategies in this paper. The final question to be answered is: Which strategy to choose when evaluating labour market programmes? Unfortunately, there is no 'one' answer to this question because there is no 'magic bullet' that will solve the evaluation problem in any case. As described above, different strategies invoke different identifying assumptions and also require different kinds of data for their implementation. When those assumptions hold, a given estimator will provide consistent estimates of certain parameters of interest (Smith, 2004). The literature provides a lot of guidance for making the right choice, based either on experimental datasets to benchmark the performance of alternative evaluation estimators or Monte-Carlo simulations.

The different estimators can be classified with respect to two dimensions. The first dimension is the required data for their implementation. Except the DID estimator, the presented methods for the static evaluation framework require only cross-sectional information for the group of participants and non-participants. However, longitudinal information might help to justify the unconfoundedness assumption, enables the researcher to combine e. g. matching with DID estimators and allows an extension to dynamic concepts of treatment evaluation. The second dimension

concerns the handling of selection bias. We have presented three estimators that are based on the unconfoundedness assumption. Clearly, the most crucial point for these estimators is that the identifying assumption is in general a very strong one and they are only as good as the used control variables X (Blundell *et al.*, 2004). If the assumption holds, both, matching and regression, can be used. Since regression analysis ignores the common support problem, imposes a functional form for the outcome equation, and is not as capable as matching of handling effect heterogeneity, matching might be preferred. If there is no common support at all, regression discontinuity models can be applied. For the situation where there is selection on unobservables, too, we have presented three strategies. Whereas selection models try to model the selection process completely, IV methods focus on searching a source of independent variation affecting the participation decision (but not the outcome) and DID methods erase a time-invariant selection effect by differencing outcomes of participants and non-participants before and after treatment took place. The crucial assumption for the latter approach is that the selection bias is time invariant. Finding a suitable and credible instrument and heterogeneous treatment effects are possible drawbacks for the IV method. The latter point is not a problem for selection models, even though this flexibility comes at a price, because a full specification of the assignment rule and stronger assumptions are required. Hence, if the common effect assumption is plausible in a given context, the IV estimator might be preferred (Smith, 2004). Finally, we have also presented some recent extensions of the static evaluation framework to analyse dynamic treatment effects, e. g. to allow for subsequent treatments and to take the timing of events into account.

References

Abbring, J. H., van den Berg, G. J. (2003). The non-parametric identification of treatment effects in duration models. *Econometrica* **71** 1491–1517.

Angrist, J. (1998). Estimating the labor market impact of voluntary military service using social security data on military applicants. *Econometrica* **66** 249–288.

Angrist, J. D., Imbens, G. W., Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* **91** 444–472.

Ashenfelter, O. (1978). Estimating the effects of training programs on earnings. *Review of Economics and Statistics* **60** 47–57.

Blundell, R., Costa Dias, M. (2000). Evaluation methods for non-experimental data. *Fiscal Studies* **21** 427–468.

Blundell, R., Costa Dias, M. (2002). Alternative approaches to evaluation in empirical microeconomics. *Portuguese Economic Journal* **1** 91–115.

Blundell, R., Dearden, L., Sianesi, B. (2004). Evaluating the impact of education on earnings in the UK: Models, methods and results from the NCDS. Working Paper No. 03/20, The Institute of Fiscal Studies, London.

CALIENDO, M., KOPEINIG, S. (2005). Some practical guidance for the implementation of propensity score matching. Discussion Paper No. 1588, IZA, Bonn.

DAWID, A. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society, Series B* **41** 1–31.

FAY, R. (1996). Enhancing the effectiveness of active labor market policies: Evidence from programme evaluations in OECD countries. *Labour Market and Social Policy Occasional Papers,* OECD, Paris.

FREDERIKSSON, P., JOHANSSON, P. (2004). Dynamic treatment assignment - The consequences for evaluations using observational data. Discussion Paper No. 1062, IZA, Bonn.

FRÖLICH, M. (2002). *Programme Evaluation and Treatment Choice.* Lecture Notes in Economics and Mathematical Systems, Springer, Berlin.

HECKMAN, J. (1978). Dummy endogenous variables in a simultaneous equation system. *Econometrica* **46** 931–959.

HECKMAN, J. (2001). Micro data, heterogeneity, and the evaluation of public policy: Nobel lecture. *Journal of Political Economy* **109** 673–748.

HECKMAN, J., ICHIMURA, H., SMITH, J., TODD, P. (1998). Characterizing selection bias using experimental data. *Econometrica* **66** 1017–1098.

HECKMAN, J., ICHIMURA, H., TODD, P. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *Review of Economic Studies* **64** 605–654.

HECKMAN, J., LALONDE, R., SMITH, J. (1999). The economics and econometrics of active labor market programs. In *Handbook of Labor Economics Vol. III* (O. Ashenfelter, D. Card, eds.), 1865–2097. Elsevier, Amsterdam.

HECKMAN, J., ROBB, R. (1985). Alternative methods for evaluating the impact of interventions - An overview. *Journal of Econometrics* **30** 239–267.

HOLLAND, P. (1986). Statistics and causal inference. *Journal of the American Statistical Association* **81** 945–960.

HUI, S., SMITH, J. (2002). The labor market impacts of adult education and training in Canada. Report prepared for the Human Resources Development Canada (HRDC), Quebec.

IMBENS, G. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *The Review of Economics and Statistics* **86** 4–29.

IMBENS, G., ANGRIST, J. (1994). Identification and estimation of local average treatment effects. *Econometrica* **62** 467–475.

LECHNER, M. (2002). Some practical issues in the evaluation of heterogenous labour market programmes by matching methods. *Journal of the Royal Statistical Society, Series A* **165** 59–82.

LECHNER, M. (2004). Sequential matching estimation of dynamic causal models. Discussion Paper No. 1042, IZA, Bonn.

LECHNER, M., MIQUEL, R. (2002). Identification of effects of dynamic treatments by sequential conditional independence assumptions. Working Paper, SIAW, University St. Gallen.

PUHANI, P. A. (2000). The Heckman correction for sample selection and its critique. *Journal of Economic Surveys* **14** 53–68.

ROSENBAUM, P. R. (2002). *Observational Studies*. Springer, New York.

ROSENBAUM, P., RUBIN, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–50.

ROY, A. (1951). Some thoughts on the distribution of earnings. *Oxford Economic Papers* **3** 135–145.

RUBIN, D. (1974). Estimating causal effects to treatments in randomised and non-randomised studies. *Journal of Educational Psychology* **66** 688–701.

SIANESI, B. (2004). An evaluation of the active labour market programmes in Sweden. *The Review of Economics and Statistics* **86** 133–155.

SMITH, J. (2000). A critical survey of empirical methods for evaluating active labour market policies. *Schweizerische Zeitschrift für Volkswirtschaft und Statistik* **136** 1–22.

SMITH, J. (2004). Evaluating local development policies: Theory and practice. Working Paper, University of Maryland.

SMITH, J., TODD, P. (2005). Does matching overcome LaLonde's critique of non-experimental estimators?. *Journal of Econometrics* **125** 305–353.

VYTLACIL, E. (2002). Independence, monotonicity and latent index models: An equivalence result. *Econometrica* **70** 331–341.

15 Survey Item Nonresponse and its Treatment *

Susanne Rässler[1] and Regina T. Riphahn[2]

[1] Kompetenzzentrum für Empirische Methoden, IAB Institut für Arbeitsmarkt- und Berufsforschung
susanne.raessler@iab.de
[2] Lehrstuhl für Statistik und empirische Wirtschaftsforschung, Universität Erlangen-Nürnberg
regina.riphahn@wiso.uni-erlangen.de

Summary: One of the most salient data problems empirical researchers face is the lack of informative responses in survey data. This contribution briefly surveys the literature on item nonresponse behavior and its determinants before it describes four approaches to address item nonresponse problems: Casewise deletion of observations, weighting, imputation, and model-based procedures. We describe the basic approaches, their strengths and weaknesses and illustrate some of their effects using a simulation study. The paper concludes with some recommendations for the applied researcher.

15.1 Introduction

Survey data can be imperfect in various ways. Sampling and noncoverage, unit nonresponse, interviewer error as well as the impact of survey design and administration can affect data quality. For the applied researcher item nonresponse, i. e., missing values among respondents' answers present a regular challenge. This problem receives increasing attention in the literature, where problems of statistical analysis with missing data have been discussed since the early 1970's (e. g., Hartley and Hocking, 1971; Rubin, 1972, 1974; or see Madow *et al.*, 1983).

Even though there exist numerous alternative approaches, most statistical software packages 'solve' the problem of item nonresponse by deleting all observations with incomplete data. This so-called 'complete case analysis' does not only neglect available information but may also yield biased estimates. In their eminent textbook Little and Rubin (1987, 2002) categorize the approaches to deal with missing data

*We are grateful to an anonymous referee who provided helpful comments. Also we like to thank Donald B. Rubin for helpful comments and always motivating discussions as well as Ralf Münnich for inspiring discussions about raking procedures.

in four main groups. Besides complete case analysis there are weighting, imputation, and model-based procedures. Weighting approaches are typically applied to correct for unit nonresponse, i. e., the complete refusal of single respondents to provide information, which may lead to biased estimates as well. The basic idea is to increase the weights of respondents in some subsamples (e. g., among providers of complete data) in order to compensate for missing responses from respondents in other subsamples (e. g., incomplete data providers). Weighting procedures can consider population or sampling weights to align the observable sample with the relevant population.

In contrast, imputation techniques insert values for missing responses and generate an artificially completed dataset. A large number of alternative procedures are applied to choose the values by which missing values are replaced: hot deck imputations use values from other observations in the sample, mean imputation fills missing variables using the mean of appropriate sub-samples, and regression imputation generates predicted values from regression models. Besides these single imputation methods, multiple imputation procedures impute more than one value for each missing value, in order to reflect the uncertainty of missingness and imputation.

Finally, model-based procedures rely on a specified model of the observed data. Inference is based on the likelihood or - in the Bayesian framework - on the posterior distribution under that model. In general, predictions of the missing data are generated based on the respondents' observed characteristics by taking advantage of correlation patterns measured for respondents without missing values. These value substitutions can occur at different levels of complexity.

An evaluation of the properties of the four approaches hinges on the assumptions regarding the nature of the missing values. The crucial role of this missing data mechanism was largely ignored until its concept was formalized by Rubin (1976). Modern statistical literature now distinguishes three cases: missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR).

MCAR refers to missing mechanisms which are unrelated to the survey variables, missing or observed. If, for instance, the probability that income is reported is the same for all individuals, regardless of, e. g., their age or income itself, then the missing income data are said to be MCAR. Data are labeled MAR, if the missing mechanism is dependent on observed but not on unobserved variables. This is the case, e. g., if special socio-economic groups are disproportionately subject to missing values and the missingness can be explained by observed variables. Finally, data are termed NMAR, if the missingness depends on the values of the variables that are actually not observed. This might be the case for income reporting, where individuals with higher incomes tend to be less likely to respond, even conditional on their observed data.

The next section describes the prevalence, determinants, and effects of item nonresponse using the German Socioeconomic Panel Survey (GSOEP) as an example. Section 3 discusses the strengths and weaknesses of the alternative approaches to solve the item nonresponse problem. The paper concludes with recommendations for applied researchers.

15.2 Item Nonresponse in the German Socioeconomic Panel

15.2.1 Prevalence of Item Nonresponse in the GSOEP

The German Socioeconomic Panel is a household panel survey covering a broad range of issues. Its questionnaire has been administered annually since 1984. It now covers over 20,000 individual respondents. The extent of item nonresponse in the GSOEP varies considerably across items. Averaging across the available 19 annual panel waves (1984-2002) we obtain 0.4 percent item nonresponse for subjective health satisfaction, 0.5 percent for political party preference, 8.9 percent for gross monthly labor earnings, and 1.3 percent for the question on whether an individual has disability status.[1]

Riphahn and Serfling (2002, 2005) compared the item nonresponse rates across financial variables in the GSOEP cross-section of 1988. At the individual level item nonresponse rates varied between 2.6 percent e. g., for retirement benefits and 15.3 percent for income from self-employment. Among variables measured at the household level they observe more than 30 percent item nonresponse for questions about interest and annuity payments. In contrast, certain questions on social transfers such as child or welfare benefits yielded nonresponse rates of below one percent.

Schräpler (2004) describes the development of item nonresponse behavior with respect to individual gross labor income. He compares the nonresponse rates of a sample of respondents over the years and finds declining nonresponse rates which differ depending on the method of data collection and respondent characteristics. Other studies confirm that individuals with a low propensity to continue responding to a panel survey are also less likely to disclose their income.

15.2.2 Determinants and Effects of Item Nonresponse

The theoretical literature on item nonresponse mainly applies two explanatory approaches, the cognitive and the rational choice model (see e. g., Schräpler 2004). Extending theoretical approaches from cognitive psychology to the interview situation, the cognitive model conceptualizes individual response behavior as a multi-stage process (Sudman *et al.*, 1996): after hearing a question it must be interpreted and understood. Next, the respondent gathers the relevant information, a stage which is affected by the complexity of the question. Finally, the information is translated to the answer format required by the questionnaire and possibly adjusted based on objectives such as self representation or social desirability.

In contrast, rational choice theory focuses only on this last stage, when respondents evaluate behavioral alternatives based on their expected costs and benefits (Esser, 1984). Benefits of responding consist of supporting a potentially appreciated cause, and of avoiding the negative effects of refusal such as breaking social norms generated by the interview situation or violating courtesy towards the interviewer. Key

[1] We thank Oliver Serfling for generating these figures.

costs of answering a survey consist of the potential negative consequence of providing private information (e. g., from tax authorities or through breach of privacy) as well as of the necessary effort to recall the desired facts.

The hypotheses that can be derived from these theories regarding the determinants of item nonresponse behavior relate to the nature of the question (i. e., cognitive complexity and sensitivity), to the relationship between respondent and interviewer, to the interview situation, and finally to the characteristics of the respondent. Dillman *et al.* (2002) provide a classification of seven causes of item nonresponse (INR):

- Survey Mode: INR is higher in self-administered questionnaires than in face-to-face interviews.
- Interviewers: if the interviewer is able to develop a high level of rapport with respondents, difficult answers may be given willingly. Interviewers' response to unanswered questions affects nonresponse outcomes.
- Question Topic and Structure: certain contents such as finances, drug use, criminal and sexual behavior are notorious for INR. Also, open-ended or multiple-part questions, as well as those with complex branching structures produce more INR.
- Question Difficulty: cognitive difficulty of questions or coverage of long time horizons generate more INR.
- Institutional Policies: sensitive information e. g., sales or investment in business surveys have high INR rates. Offering a 'don't know' answer option also increases INR. NRespondents' Attributes: in many surveys older and less educated people are less likely to respond.

Schräpler (2004), Frick and Grabka (2003), and Riphahn and Serfling (2005) estimated multivariate models of item nonresponse behavior controlling for relevant indicators. The studies differ in their empirical approach, the subsample taken from the GSOEP, the number of items considered, and in the key issues addressed.

Nevertheless some general findings can be summarized as follows: (i) there is significant heterogeneity in the processes determining item nonresponse behavior across items; (ii) the association between interviewer and respondent characteristics does not appear to be influential for item nonresponse behavior; (iii) item nonresponse rates are higher when the interviewer is female and when a new interviewer is assigned to respondents; (iv) item nonresponse on income is higher at low and high income levels; (v) face-to-face interviews yield lower nonresponse rates than self-reporting or computer assisted interviewing; (vi) item nonresponse and 'don't know' answers are determined by different mechanisms.

As item nonresponse behavior appears to affect financial variables most severely, analyses of income and wealth issues may be most subject to biases deriving from missing data. Given that item nonresponse may indeed bias the results of empirical analyses in general, correction methods need to be considered.

15.3 Dealing with Item Nonresponse

This section discusses four frequently applied methods for the analysis of data with missing values due to item nonresponse:[2]

15.3.1 Complete Case Analysis

Software packages often handle incomplete data by deleting all cases with at least one missing item (listwise deletion or complete case analysis, CC). This practice is inefficient and often leads to substantially biased inferences. Listwise deletion can reduce the available data considerably, so that they are no longer representative of the population of interest.

Thus, CC analysis can be wasteful, as informative data are discarded when they belong to records that have missing values on other variables. As an alternative for univariate analyses often all values that are observed for a variable of interest are used independent of missing values on other variables (available case analysis, AC). A major disadvantage of AC analysis is that different analyses from a given dataset will be performed on different samples, depending on which observations have complete data for each analysis. This can lead to inconsistent estimates especially when comparisons are made using estimates from different subsamples. In general, basing inferences only on the complete cases implies the tacit assumption that the missing data are missing completely at random, which is typically not the case. The size of the resulting bias depends on the degree of violation of the MCAR assumption, the share of missing data, and the specifics of the analysis.

15.3.2 Weighting

The most common procedure to correct for (unit) nonresponse in official statistics and survey research is weighting. In general, weighting is applied to address problems of nonresponse and to adjust the sample when unequal probabilities of selection have been used. Therefore, two types of weights for a unit i, the nonresponse or poststratification weights g_i and the inverse-probability or design weights $d_i = 1/\pi_i$, should be distinguished (Gelman and Carlin, 2002). The former are typically used to correct for differences between sample and population and have to be estimated. The latter are usually known in advance, and are needed to generate unbiased estimates for the population target quantity under repeated sampling given a specific sampling design.

There is common agreement that for estimating population totals, means, and ratios, weighted averages are appropriate. An example are Horvitz-Thompson type estimators which are, e. g., for a population total given by

$$\widehat{Y} = \sum_{i=1}^{n} \frac{y_i}{\pi_i} = \sum_{i=1}^{n} d_i y_i \,.$$

In combination with complete case analysis weights may also be used to address

[2] For a discussion of procedures to avoid item nonresponse in advance, such as interviewer training, questionnaire structure, or administration, see e. g., Groves *et al.* (2002).

nonresponse problems. If the probabilities of response for each responding unit were known, then

$$P(\text{selection and response}) = P(\text{selection})P(\text{response}|\text{selection})$$

(Little and Rubin, 2002) and the individual weights w_i for a unit i are given by $w_i = d_i g_i$. In practice, the response probability is unknown and a standard approach, e. g., is to form adjustment cells based on background variables measured for respondents and nonrespondents. The nonresponse weight for individuals in an adjustment cell is then the inverse of the response rate in that cell.

For illustration, let the sample be divided into J homogeneous cells or groups with respect to the assumed response generating process. Let n_j denote the expected or planned sample size in group or cell j, $j = 1, 2, \ldots, J$, e. g., among young working women, and m_j the number of respondents in this group. The individual weight w_i of an observation i within a cell j is computed as $d_i g_i = d_i \frac{n_j}{m_j}$.

If only sample counts are used in the weighting procedure, weighting can be interpreted as a single conditional mean imputation. To illustrate this, consider the so-called weighting-class estimator (Oh and Scheuren, 1983) which is given by

$$\widehat{Y} = \frac{N}{n} \sum_{j=1}^{J} \frac{n_j}{m_j} \sum_{i=1}^{m_j} y_{ij} = \frac{N}{n} \sum_{j=1}^{J} n_j \overline{y}_j^{obs} = \frac{N}{n} \sum_{j=1}^{J} \left(\sum_{i=1}^{m_j} y_{ij} + (n_j - m_j)\overline{y}_j^{obs} \right),$$

where N/n is the sampling fraction. This weighting-class estimator is identical to the estimate derived by single conditional mean imputation. Thus, naive estimates of standard errors and confidence intervals will be biased downwards as it is typically the case with single imputation. The derivation of an unbiased variance estimator is cumbersome.[3]

In practice, the population totals of the cells, one wants to adjust for, are often unknown, but the marginals of different weighting variables are known for the population. In this situation, a set of weighting vectors can be estimated, which satisfies the constraints given by the population margins: this procedure is termed raking. It applies iterated proportional fitting (IPF) to obtain weighted sample counts that match the population on the set of margins. Approaches that make use of auxiliary information comprise regression and ratio estimates; for these and extensions see Deville and Särndal (1992) and Deville *et al.* (1993). To sum up, calibration and raking procedures which include the generalized regression (GREG) estimator and iterative proportional fitting are widely used in the case of unit nonresponse. If, e. g., only a population quantity such as the total is to be estimated, they may also be used in the presence of item nonresponse.

While weighting methods are often relatively easy to implement, they face three major disadvantages: (i) especially in the presence of outliers weighted estimates can have high variances, (ii) variance estimation for weighted estimates can be

[3] Notice that often additional information is available and instead of weighting a multiple imputation procedure (see Section 3.5) can be applied successfully, see Rässler and Schnell (2004).

computationally expensive, if, e. g., linearization or jackknife methods have to be used (see Gelman and Carlin, 2002), and (iii) weighting methods typically do not model the joint distribution of the data as is done by multiple imputation or model-based approaches.

15.3.3 Imputation Techniques

Imputation techniques fill in one or more plausible values for each missing datum so that one or more completed datasets are created (i. e., single vs. multiple imputation). Often it is easier to first impute missing values and to then use standard complete-data methods of analysis than to develop statistical techniques that allow the analysis of incomplete data directly. Imputation allows to use information not available to the analyst. Imputation of survey data can be performed separately from the analysis, which is appealing. The application of standard methods on data with singly imputed values will result in underestimated standard errors, if the uncertainty of the imputation procedure is ignored. Due to its operational convenience, single imputation has long been used, especially by statistical offices. Among the key challenges for single imputation is to preserve the covariance structures in the data and at the same time to appropriately reflect the uncertainty due to the imputation process. Usually this means that for every point estimate based on singly imputed data its frequency valid variance estimate has to be derived separately; see Lee *et al.* (2002).

Multiple imputation (MI), introduced by Rubin (1978) and discussed in detail in Rubin (1987, 2004), retains the advantages of imputation while allowing the data analyst to make valid assessments of uncertainty. Multiple imputation reflects uncertainty in the imputation of the missing values through wider confidence intervals and larger p-values than under single imputation. MI is a Monte Carlo technique that replaces the missing values by $m > 1$ simulated versions, generated according to a probability distribution which indicates how likely the true values are given the observed data. Typically m is small, e. g., $m = 5$, although with increasing computational power m can be 10 or 20. In general, this depends on the amount of missingness and on the distribution of the parameters to be estimated.

To illustrate this, let Y_{obs} denote the observed components of any uni- or multivariate variable Y, and Y_{mis} its missing components. Then, m values are imputed for each missing datum according to some distributional assumptions creating $m > 1$ independent simulated imputations $(Y_{obs}, Y_{mis}^{(1)})$, $(Y_{obs}, Y_{mis}^{(2)})$, ..., $(Y_{obs}, Y_{mis}^{(m)})$. Standard complete-case analysis can be performed for each of the m imputed datasets, enabling us to calculate the imputed data estimate $\widehat{\theta}^{(t)} = \widehat{\theta}(Y_{obs}, Y_{mis}^{(t)})$ along with its estimated variance $\widehat{var}(\widehat{\theta}^{(t)}) = \widehat{var}(\widehat{\theta}(Y_{obs}, Y_{mis}^{(t)}))$, $t = 1, 2, \ldots, m$. The complete-case estimates are combined according to the MI rule that the MI point estimate for θ is simply the average

$$\widehat{\theta}_{MI} = \frac{1}{m} \sum_{t=1}^{m} \widehat{\theta}^{(t)}. \tag{15.1}$$

To obtain a standard error $\sqrt{\widehat{var}(\widehat{\theta}_{MI})}$ for the MI estimate $\widehat{\theta}_{MI}$, we first calculate

the 'between-imputation' variance

$$\widehat{var}(\widehat{\theta})_{between} = B = \frac{1}{m-1}\sum_{t=1}^{m}(\widehat{\theta}^{(t)} - \widehat{\theta}_{MI})^2, \tag{15.2}$$

and then the 'within-imputation' variance

$$\widehat{var}(\widehat{\theta})_{within} = W = \frac{1}{m}\sum_{t=1}^{m}\widehat{var}(\widehat{\theta}^{(t)})\,. \tag{15.3}$$

Finally, the estimated total variance is defined by

$$\widehat{var}(\widehat{\theta}_{MI}) = T = \widehat{var}(\widehat{\theta})_{within} + (1+\frac{1}{m})\widehat{var}(\widehat{\theta})_{between} = W + \frac{m+1}{m}B\,.$$

For large sample sizes, tests and two-sided $(1-\alpha)100\%$ interval estimates can be based on Student's t-distribution

$$(\widehat{\theta}_{MI} - \theta)/\sqrt{T} \sim t_v \quad \text{and} \quad \widehat{\theta}_{MI} \pm t_{v,1-\alpha/2}\sqrt{T} \tag{15.4}$$

with degrees of freedom

$$v = (m-1)\left(1 + \frac{W}{(1+m^{-1})B}\right)^2. \tag{15.5}$$

MI is in general applicable when the complete-data estimates are asymptotically normal (e. g., ML estimates) or t distributed; see Rubin and Schenker (1986), Rubin (1996), Barnard and Rubin (1999), or Little and Rubin (1987, 2002).

The theoretical motivation for multiple imputation is Bayesian, although the resulting multiple imputation inference is usually also valid from a frequentist viewpoint. Basically, MI requires independent random draws from the posterior predictive distribution of the missing data given the observed data. Usually this is performed by a two-step procedure. First, we take random draws of the parameters according to their observed-data posterior distribution. Second, we perform random draws of the missing data according to their conditional predictive distribution. This is done m times. If only one variable has missing values, such a specification is rather straightforward and univariate (Bayesian) regression models may be applied. When the data have a multivariate structure and different missing data patterns, the observed-data posteriors are often not standard distributions from which random numbers can easily be generated. However, with increasing computational power simpler methods have been developed to enable multiple imputation based on Markov Chain Monte Carlo (MCMC) techniques. Common concerns with multiple imputation address the model-based assumptions and the complexity of the Bayesian posterior predictions. Clearly, there is no assumption-free imputation method but multiple imputation explicitly formulates and evaluates these assumptions. For a broad discussion of advantages and disadvantages of imputation procedures see Groves *et al.* (2002, Chapter 22 and 23).

15.3.4 Model-based Procedures

Model-based procedures to adjust for nonresponse simultaneously have to model the distribution of the data Y and the response mechanism R. Without any further assumptions regarding the response mechanism, the joint distribution $f_{Y,R}(y, r; \theta, \xi)$ has to be modelled. In so-called nonignorable nonresponse models this is done in two slightly differing ways. On the one hand, selection models as considered by Heckman (1976), specify $f_{Y,R}(y, r; \theta, \xi)$ as

$$f_{Y,R}(y, r; \theta, \xi) = f_Y(y; \theta) f_{R|Y}(r|y; \xi) \tag{15.6}$$

and have to formulate an explicit model for the distribution of the response missing-data mechanism $f_{R|Y}(r|y; \xi)$ where θ and ξ are the unknown parameters or in the Bayesian context are random variables as well. Keeping the notation simple, with missing data the likelihood of (15.6) is

$$L(\theta, \xi; y, r) = \int f_{Y_{obs}, Y_{mis}}(y_{obs}, y_{mis}; \theta) f_{R|Y_{obs}, Y_{mis}}(r|y_{obs}, y_{mis}; \xi) dy_{mis} . \tag{15.7}$$

On the other hand, pattern-mixture models as discussed by Glynn et al. (1986) factor the joint distribution in a different way:

$$f_{Y,R}(y, r; \theta, \xi) = f_{Y|R}(y|r; \theta) f_R(r; \xi) , \tag{15.8}$$

where the distribution of Y is conditioned on the missing data pattern R. Therefore, the resulting marginal distribution of Y will be a mixture of distributions.

Under the MCAR assumption expressions (15.6) and (15.8) are equivalent. If distributional assumptions are added and the data are not MCAR, these specifications can lead to different models. Maximum-likelihood estimates are found by maximizing the likelihood functions with respect to θ and ξ. In the Bayesian context the posterior distribution is obtained by incorporating a prior distribution and performing the necessary integrations.

In general, either way has its merits and demerits. Specification models usually require the existence of identifying restrictions, are very sensitive to model misspecification, and the results are often claimed to be unstable. Pattern-mixture models are often under-identified and also require identifying restrictions. Typically, pattern-mixture models are suggested to be used for sensitivity analyses, see, e. g., Little (1993).

Since the assumption of MAR cannot be contradicted by the observed data, more often the observed-data likelihood, which is also called the likelihood ignoring the missing data mechanism, is considered:

$$L(\theta; y_{obs}) = \int f_{Y_{obs}, Y_{mis}}(y_{obs}, y_{mis}; \theta) dy_{mis} . \tag{15.9}$$

Inferences about θ can be based on (15.9) rather than on the full likelihood (15.7) if the missing data mechanism is ignorable. Notice that ignorable Bayesian inference would add a prior distribution for θ. Rubin (1976) has shown that an ignorable

missing data mechanism is given when two conditions hold. First, the parameters θ and ξ have to be distinct, i. e., they are not functionally related or - in the Bayesian framework - are a priori independent. Second, the missing data are MAR.

Ignorable ML methods focussing on the estimation of θ have a couple of advantages. Usually the interest is in θ and not in ξ. Then the explicit modeling of the response mechanism can be cumbersome and easily misspecified. Also, often information for the joint estimation of θ and ξ is limited. Thus, estimates assuming MAR data turn out to be more robust in many cases.

However, in many missing data problems, even the observed-data likelihood (15.9) is complicated and explicit expressions for the ML estimate cannot be derived. Here, the Expectation-Maximization (EM) algorithm is a broadly applicable approach to the iterative computation of maximum likelihood estimates. On each iteration of the EM algorithm there are two steps, called the expectation or E-step and the maximization or M-step. The basic idea of the EM algorithm is first (E-step) to fill in the missing data Y_{mis} by their conditional expectation given the observed data and an initial estimate of the parameter θ to achieve a completed likelihood function, and second (M-step) to recalculate the maximum likelihood (ML) estimate of θ given the observed values y_{obs} and the filled-in values of $Y_{mis} = y_{mis}$. The E-step and M-step are iterated until convergence of the estimates is achieved.

More precisely, it is the log likelihood $\ln L(\theta; y)$ of the complete-data problem that is manipulated in the E-step. As it is based partly on unobserved data, it is replaced by its conditional expectation

$$E(\ln L(\theta; Y)|y_{obs}; \theta^{(t)})$$

given the observed data y_{obs} and a current fit $\theta^{(t)}$ for the unknown parameters. Thus the E-step consists of calculating this conditional expectation $E(\ln L(\theta; Y)|y_{obs}; \theta^{(t)})$. The simpler M-step computation can now be applied to this completed data and a new actual value $\theta^{(t+1)}$ for the ML estimate is computed therefrom. Now let $\theta^{(t+1)}$ be the value of θ that maximizes $E(\ln L(\theta; Y)|y_{obs}; \theta^{(t)})$. Dempster *et al.* (1977) have shown that $\theta^{(t+1)}$ then also maximizes the observed-data likelihood $L(\theta; y_{obs})$ in the sense that the observed-data likelihood of $\theta^{(t+1)}$ is at least as high as that of $\theta^{(t)}$, i. e., $L(\theta^{(t+1)}; y_{obs}) \geq L(\theta^{(t)}; y_{obs})$.

Starting from some suitable initial parameter values $\theta^{(0)}$, the E- and the M-steps are repeated until convergence, for instance, until $|\theta^{(t+1)} - \theta^{(t)}| \leq \epsilon$ holds for some fixed $\epsilon > 0$. Not all the problems are well-behaved, however, and sometimes the EM does not converge to a unique global maximum.[4]

15.3.5 Evidence from a Comparison Study

In this section we present a simple simulation study to illustrate the implications of alternative imputation procedures. We compare moments of a random variable (income) when applying multiple imputation (MI), simple single mean imputation

[4]For a detailed description of the EM algorithm and its properties see McLachlan and Krishnan (1997), Schafer (1997), Little and Rubin (2002), and the fundamental paper of Dempster *et al.* (1977).

(SI), single mean imputation within classes (also known as conditional mean imputation and here equivalent to a weighting procedure as shown in Section 3.2) (SI CM), and complete case analysis (CC).

Assume that a randomly drawn variable which we label age (AGE) is normally distributed with mean 40 and standard deviation 10, and another randomly drawn variable labelled income (INC) is normally distributed with mean 1500 and standard deviation 300. Because real income variables do not generally follow a normal distribution, often their log transformation log(INC) is used to achieve approximate normality. Let the correlation between age and income be 0.8, then[5]

$$(AGE, INC) \sim N\left(\begin{pmatrix} 40 \\ 1500 \end{pmatrix}, \begin{pmatrix} 10^2 & 0.8 \cdot 3000 \\ 0.8 \cdot 3000 & 300^2 \end{pmatrix}\right).$$

A sample of n = 2000 is drawn from this universe. After being generated, the AGE variable is recoded into 6 categories, 1 $\leq$ 20 years, 2 = over 20 - 30 years, ..., 6 > 60 years. First, the complete cases are analyzed, the mean income estimate, its standard error (s. e.), and the 95% confidence interval are calculated. Then different missingness mechanisms (MCAR, MAR, NMAR) are applied on income. Under MAR, income is missing with higher probability when age is higher, under NMAR, the probability that income is missing is higher the higher income is itself.

After discarding 30% of the income data, first the complete cases are analyzed, then a simple mean imputation is performed, and, finally, a proper multiple imputation procedure with $m = 5$ is used according to Rubin (1987, p. 167). The whole simulation process of creating the data, applying the missingness, performing the imputations, and analyzing the sample is repeated 1000 times. The coverage (cvg.) is counted, i. e., the number of confidence intervals out of 1000 that cover the true mean value. The average bias, the standard errors, and the usual correlation estimates between age (recoded) and income are given in Table 1.

The results in Table 15.1 show how precision is reduced when only the complete cases are used under MCAR, and how biased the complete case estimate (CC) gets when the missingness is MAR or NMAR.[6] The table also shows how biased a simple mean imputation is and how this bias is corrected when conditional means are imputed instead of the overall mean (cf. the means in Rows 7 and 8 and 11 and 12). However, this conditional mean imputation requires that the missingness depends on the variable conditioned on. The single mean imputation within classes also leads to an overestimation of the correlation between recoded AGE and INC though the simple single imputation underestimates it (see the last column of Table 1). Moreover, with single imputation the standard errors are always too small to get the nominal coverage.

Even if the missingness is MCAR, a simple mean imputation affects standard errors and correlations. Under MAR and even under NMAR, multiple imputation

[5]For robustness checks this study was also run with lower correlation values. However, that did not change the main message. Notice that the lower the correlation the less efficient are the procedures under NMAR.

[6]For the precision compare the standard errors in Row 1 to those of the CC analyses in Rows 2, 6, and 10. For bias compare the means in Rows 2, 6 and 10.

yields results much closer to the true values. Particularly in a NMAR scenario MI borrows strength from the correlation between age and income. Standard errors, correlation, and the nominal coverage are well reproduced by MI. Notice that confidence intervals under MI can be even narrower than confidence intervals based on complete case analysis (CC). This is especially true if the imputed sample is substantially larger than the complete case sample. Therefore, typically, the following comparisons hold for most surveys and most estimates of standard errors:

$$\text{s.e.(SI)} < \text{s.e.(truth)} < \text{s.e.(MI)} < \text{s.e.(CC)}.$$

More elaborate comparisons by simulation studies are provided, e. g., by Schafer (1997), Raghunathan and Rubin (1998), or Münnich and Rässler (2005). The latter are comparing especially GREG and Horvitz-Thompson estimators using nonresponse corrections as well as MI procedures.

Table 15.1: *Results of the simulation study.*

No	Missing	Proc.	Mean(INC)	Bias(INC)	S.e.(INC)	Cvg.	Cor(AGE, INC)
1	None		1500.21	0.21	6.71	0.96	0.77
2	MCAR	CC	1500.14	0.14	8.01	0.95	0.77
3	MCAR	SI	1500.14	0.14	5.61	0.82	0.64
4	MCAR	SI CM	1500.20	0.20	6.28	0.91	0.82
5	MCAR	MI	1500.24	0.24	7.34	0.95	0.77
6	MAR	CC	1470.35	-29.65	7.98	0.04	0.77
7	MAR	SI	1470.35	-29.65	5.58	0.01	0.63
8	MAR	SI CM	1499.90	-0.10	6.28	0.88	0.82
9	MAR	MI	1499.82	-0.18	7.43	0.93	0.77
10	NMAR	CC	1474.29	-25.71	7.99	0.11	0.77
11	NMAR	SI	1474.29	-25.71	5.59	0.03	0.64
12	NMAR	SI CM	1489.33	-10.66	6.26	0.59	0.82
13	NMAR	MI	1489.30	-10.70	7.36	0.71	0.77

15.4 Conclusions and Recommendations

Item nonresponse is a common problem in empirical analyses. Research on the determinants of nonresponse behavior yields a catalogue of relevant factors. The evidence on German data confirms that data collection methods and respondent characteristics affect nonresponse behavior. Extant studies also confirm that dif-

ferent ways of dealing with item nonresponse may affect the results of empirical analyses.

We discuss the strengths and weaknesses of four commonly used approaches to deal with item nonresponse and provide a simulation study. This simulation yields that the most commonly used approach, which considers only observations without missing values, can lead to substantial biases in the estimates. The performance of single imputation procedures depends on whether there are patterns in the missingness of the data and on whether the information is missing (completely) at random. Multiple imputation procedures appear to yield the best coverage of the true value and the best reflection of existing correlation patterns.

Casewise deletion can only be an appropriate procedure if the missing data are missing completely at random. In all other cases it involves biased estimates and other procedures are preferable. Weighting is a first step to correct for nonresponse and disproportionalities. The literature suggests that multiple imputation under MAR often is quite robust against violations of the MAR assumption. Only when NMAR is a serious concern and the share of missing information is substantial it seems necessary to jointly model the data and the missingness using model-based procedures. Since missing values cannot be observed, there is no direct evidence in the data to test a MAR assumption. Therefore, it seems useful to consider alternative models and to explore the sensitivity of resulting inferences. We conclude that a multiple imputation procedure seems to be the best alternative at hand to account for missingness and to exploit all available information. In particular it generates the only format with correct standard errors allowing valid inference from standard complete case analysis.

It is recommendable that empirical researchers step beyond standard complete or available case analysis and investigate the robustness of findings by applying alternative procedures. This is aided by the fact that various single imputation techniques, such as mean imputation, conditional mean imputation, or regression imputation, are now available in commercial statistical software packages. Free programs and routines comprise the stand-alone Windows program NORM or the S-PLUS / R libraries NORM, CAT, MIX, PAN, and MICE which are all basically data augmentation algorithms. NORM uses a normal model for continuous data, CAT a log-linear model for categorical data. MIX relies on a general location model for mixed categorical and continuous data. PAN is created for panel data applying a linear mixed-effects model. Moreover, there are the free SAS-callable application IVEware as well as a STATA packet MVIS which are, like MICE, based on the very flexible sequential regression approach. The SAS procedures PROC MI with PROC MIANALYZE provide a parametric and a nonparametric regression imputation approach, as well as the multivariate normal model. Finally, there is the free Windows or Gauss version AMELIA. With increasing computational power, more and more multiple imputation techniques are now implemented in available statistics software to create multiply-imputed datasets for further analyses.[7]

[7]For links and further details see www.multiple-imputation.com, Horton and Lipsitz (2001), or Rässler *et al.* (2003).

References

BARNARD, J., RUBIN, D. B. (1999). Small-sample degrees of freedom with multiple imputation. *Biometrika* **86** 948–955.

DEMPSTER, A. P., LAIRD, N. M., RUBIN, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B* **39** 1–38.

DEVILLE, J. C., SÄRNDAL, C. E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association* **87** 376–382.

DEVILLE, J. C., SÄRNDAL, C. E., SAUTORY, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association* **88** 1013–1020.

DILLMAN, D. A., ELTINGE, J. L. , GROVES, R. M., LITTLE, R. J. A. (2002). Survey nonresponse in design, data collection, and analysis. In *Survey Nonresponse* (R. M. Groves, D. A. Dillman, J. L. Eltinge, R. J. A. Little, eds.), 3–26. Wiley, New York.

ESSER, H. (1984). Determinanten des Interviewer- und Befragtenverhaltens: Probleme der theoretischen Erklärung und empirischen Untersuchung von Interviewereffekten. In *Allgemeine Bevölkerungsumfrage der Sozialwissenschaften* (K. Mayer, P. Schmidt, eds.), 26-71. Campus, Frankfurt.

FRICK, J. R., GRABKA, M. M. (2003). Missing income data in the German SOEP: Incidence, imputation and its impact on the income distribution. DIW Discussion Papers 376, DIW Berlin.

GELMAN, A., CARLIN, J. B. (2002). Poststratification and weighting adjustment. In *Survey Nonresponse* (R. M. Groves, D. A. Dillman, J. L. Eltinge, R. J. A. Little, eds.), 289–302. Wiley, New York.

GLYNN, R., LAIRD, N. M., RUBIN, D. B. (1986). Selection modeling versus mixture modeling with nonignorable nonresponse. In *Drawing Inferences from Self-Selected Samples* (H. Wainer, ed.), 119–146. Springer, New York.

GROVES, R. M., DILLMAN, D. A., ELTINGE, J. L., LITTLE, R. J. A. (2002). *Survey Nonresponse*. Wiley, New York.

HARTLEY, H. O., HOCKING, R. R. (1971). The analysis of incomplete data. *Biometrics* **27** 783–808.

HECKMAN, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement* **5** 475–492.

HORTON, N. J., LIPSITZ, S. R. (2001). Multiple imputation in practice: Comparison of software packages for regression models with missing variables. *American Statistician* **55** 244–254.

LEE, H., RANCOURT, E., SÄRNDAL, C. E. (2002). Variance estimation from survey data under single imputation. In *Survey Nonresponse* (R. M. Groves, D. A. Dillman, J. L. Eltinge, R. J. A. Little, eds.), 315–328. Wiley, New York.

LITTLE, R. J. A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association* **88** 125-134.

LITTLE, R. J. A., RUBIN, D. B. (1987, 2002). *Statistical analysis with missing data.* 1. and 2. ed., Wiley, Hoboken, New Jersey.

MADOW, W. G. , OLKIN, I., RUBIN, D. B. (1983). *Incomplete Data in Sample Surveys.* Academic Press, New York.

MCLACHLAN, G. J., KRISHNAN, T. (1997). *The EM Algorithm and Extensions.* Wiley, New York.

MÜNNICH, R., RÄSSLER, S. (2005). PRIMA: A new multiple imputation procedure for binary variables. *Journal of Official Statistics* (to appear).

OH, J. L., SCHEUREN, F. (1983). Weighting adjustment for unit nonresponse. In *Incomplete Data in Sample Surveys 2* (W. G. Madow, I. Olkin, D. B. Rubin, eds.), 143–184. Academic Press, New York.

RAGHUNATHAN, T. E., RUBIN, D. B. (1998). Roles for Bayesian Techniques in Survey Sampling. *Proceedings of the Silver Jubilee Meeting of the Statistical Society of Canada* 51–55.

RÄSSLER, S., RUBIN, D. B., SCHENKER, N. (2003). Imputation. In *Encyclopedia of Social Science Research Methods* (A. Bryman, M. Lewis-Beck, T. F. Liao, eds.), 477–482. Sage, Thousand Oaks.

RÄSSLER, S., SCHNELL, R. (2004). Multiple imputation for unit nonresponse versus weighting including a comparison with a nonresponse follow-up study. Diskussionspapier der Lehrstühle für Statistik 65/2004, Nürnberg.

RIPHAHN, R. T., SERFLING, O. (2002). Item non-response on income and wealth questions. IZA Discussion Paper No. 573, IZA Bonn.

RIPHAHN, R. T., SERFLING, O. (2005). Item non-response on income and wealth questions. *Empirical Economics* (to appear).

RUBIN, D. B. (1972). A non-iterative algorithm for least squares estimation of missing values in any analysis of variance design. *The Journal of the Royal Statistical Society, Series C* **21** 136–141.

RUBIN, D. B. (1974). Characterizing the estimation of parameters in incomplete-data problems. *Journal of the American Statistical Association* **69** 467–474.

RUBIN, D. B. (1976). Inference and missing data. *Biometrika* **63** 581–592.

RUBIN, D. B. (1978). Multiple imputation in sample surveys - a phenomenological Bayesian approach to nonresponse. *Proceedings of the Survey Research Methods Sections of the American Statistical Association* 20–40.

RUBIN, D. B. (1987, 2004). *Multiple Imputation for Nonresponse in Surveys.* 1. and 2. ed., Wiley, Hoboken, New Jersey.

RUBIN, D. B. (1996). Multiple imputation after 18+ years (with discussion). *Journal of the American Statistical Association* **91** 473–489.

RUBIN, D. B., SCHENKER, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association* **81** 366–374.

SCHAFER, J. L. (1997). *Analysis of Incomplete Multivariate Data.* Chapman and Hall, London.

SCHRÄPLER, J. P. (2004). Respondent behavior in panel studies. A case study for income nonresponse by means of the Germany Socio-Economic Panel (SOEP). *Sociological Methods and Research* **33** 118–156.

SUDMAN, S., BRADBURN, N. M., SCHWARZ, N. (1996). *Thinking about Answers. The Application of Cognitive Processes to Survey Methodology.* Jossey Bass Publishers, San Francisco.