THE PROVISIONAL MILITARY GOVERNMENT OF
SOCIALIST ETHIOPIA
NATIONAL REVOLUTIONARY DEVELOPMENT
CAMPAIGN AND CENTRAL PLANNING
SUPREME COUNCIL
CENTRAL STATISTICAL OFFICE

HANDBOOK ON APPLIED ECONOMETRICS

ADDIS ABABA, DECEMBER 1983

THE PROVISIONAL MILITARY GOVERNMENT OF
SOCIALIST ETHIOPIA
NATIONAL REVOLUTIONARY DEVELOPMENT
CAMPAIGN AND CENTRAL PLANNING
SUPREME COUNCIL
CENTRAL STATISTICAL OFFICE

HANDBOOK ON APPLIED ECONOMETRICS

ADDIS ABABA, DECEMBER 1983

CONTENTS

PREFACE.

PART A. BACKGROUND OF ECONOMETRIC MODELLING.

CHAPTER 1. ECONOMETRICS AND ITS OBJECTIVES.

1.1. The Role of Econometric Modelling in the National Economic Planning.

1.2. The Objectives, Structure and Main Stages of Econometric Modelling.

CHAPTER 2. MATHEMATICAL ANALYSIS BACKGROUND OF ECONOMETRICS.

2.1. Sets and Operations on Sets.

2.2. Functions and Differentiation.

CHAPTER 3. LINEAR ALGEBRA BACKGROUND OF ECONOMETRICS.

3.1. Matrices and Operations on Matrices.

3.2. System of Linear Equations and its Solution.

3.3. Quadratic Forms and Definite Matrices.

3.4. Some Special Matrices used in Econometrics

CHAPTER 4. PROBABILITY THEORY BACKGROUNDS OF ECONOMETRICS.

4.1. Random Events and Probability Measure.

4.2. Random Variables and Probability Distributions.

4.3. Parameters of Distribution of the Random Variables.

4.4. Some Important Probability Distributions.

4.5. Regression Functions.

CHAPTER 5. MATHEMATICAL STATISTICS BACKGROUND OF ECONOMETRICS.

5.1. Outline of the Statistical Estimation Theory.

5.2. Outline of the Statistical Hypotheses Testing Theory.

PART B. ECONOMETRIC THEORY AND PRACTICE.

CHAPTER 6. LINEAR MODEL WITH ONE EXPLANATORY VARIABLE.

6.1. Formulation of the model.
6.2. Estimation of the Simplest Linear Model.
6.3. Statistical Inference from the Model.
6.4. Forecasting Based on the Simplest Linear Model.

CHAPTER 7. NON-LINEAR MODELS WITH ONE EXPLANATORY VARIABLE.

7.1. Power-type model, its Reparametrization and Estimation
7.2. Exponential Model, Logarithmic Model.
7.3. Logistic Trend Function Model.

CHAPTER 8. MULTIPLE LINEAR REGRESSION MODEL.

8.1. Estimation by the OLS Method.
8.2. Gauss-Markov Theorem.
8.3. Residual Variance and Estimation of Standard Errors.

CHAPTER 9. STATISTICAL INFERENCE IN THE MULTIPLE NORMAL REGRESSION MODEL.

9.1. Normality Assumption and Its Consequences.
9.2. Interval Estimation and Significance Tests.
9.3. Analysis of Variance and the Coefficient of Determination R^2.
9.4. Forecasting Based on the Multiple Linear Regression Model.
9.5. Use of Computers and Examples.

CHAPTER 10. SOME SPECIAL PROBLEMS.

10.1. Changing the Scale
10.2. Multicollinearity Problem
10.3. Testing the Normality of Residuals

CHAPTER 11. USE OF THE GENERALIZED LEAST SQUARES METHOD FOR MODELS WITH HETEROSCEDASTICITY.

11.1. The Generalized Least Squares Method.
11.2. Heteroscedasticity of Random Disturbances.
11.3. Bartlett's Test for Grouped Heteroscedasticity.

CHAPTER 12. MODELS WITH AUTOCORRELATION OF RANDOM DISTURBANCES.

12.1. The Autoregressive Stochastic Process of Disturbances.
12.2. Application of the GLS Method.
12.3. Use of the Durbin-Watson Test for Autocorrelation.
12.4. Forecasting Based on the Model with Autoregressive Disturbances.

CHAPTER 13. ANALYSIS OF RESIDUALS - A SUMMARIZED PICTURE.

13.1. Flow-Chart of the Proper Sequence of Tests
13.2. Example of Full Analysis.

CHAPTER 14. USE OF PRIOR INFORMATION - NON-CLASSICAL ESTIMATION TECHNIQUES.

14.1. The Restricted Least Squares Method.
14.2. Use of the GLS Method for the Combined Model.
14.3. Bayesian Estimation Technique.

CHAPTER 15. MAIN DOMAINS OF APPLIED ECONOMETRIC MODELS.

15.1. Econometric Demand Functions and Elasticities.
15.2. Econometric Production Functions.

CHAPTER 16. INTRODUCTION TO SIMULTANEOUS EQUATIONS MODELS.

16.1. Classification of Variables.
16.2. Structural Form and Reduced Form of the Model.
16.3. Identification of the Model.

CHAPTER 17. ESTIMATION OF MULTI- EQUATIONAL MODELS.

17.1. Indirect Least Squares Method.
17.2. Two-stage Least Squares Method.

CHAPTER 18. FORECASTING BASED ON MULTI-EQUATIONAL MODELS.

18.1. Use of Reduced Form.
18.2. Ex Post Analysis of Forecasts.

PREFACE

This handbook is designed to be simple book on econometric methods with emphasis on and illustrations from agricultural sector. It emphasises the application of econometric analysis in Ethiopian Agriculture which is the main economic sector of the country.

Econometric analysis plays a vital role as a tool for modelling economic phenomena and its forecasting in socialist countries having a centrally planned economy. It supplies reliable information and scientific methods for analysis and forecasting economic phenomena.

Since 1980 with the support from FAO/UNDP the Project"Integrated System of Food and Agricultural Statistics Programme" was launched. Under this programme, among its various activities, econometric research and training in econometric techniques has been conducted for the staff of the Central Statistics Office (CSO) and other governmental institutions.

This handbook was prepared by Jerzy Gren, FAO, Econometrician, based on the notes given in the lectures on econometric modelling and forecasting. He is a Professor of Econometrics and Mathematical Statistics at the Central School of Planning and Statistics, Warsaw, Poland, now working in the CSO under the above mentioned project.

The handbook consists of two parts. Part A (Chapters 1-5) describes the economic and mathematical background of econometric modelling, while Part B describes the basic econometric theory and practice in estimation, checking and application of econometric models. Even though more emphasis is given to the single equation models, the last few chapters are devoted to the multi-equations econometric models.

We hope this handbook will provide a guide-line to planners and other decision makers in the application of econometric models.

Ato Mitik Beyene
Manager,
Central Statistical Office

ADDIS ABABA, NOVEMBER, 1983.

PART A. BACKGROUND OF ECONOMETRIC MODELLING.

CHAPTER 1. ECONOMETRICS AND ITS OBJECTIVES.

1.1. The Role of Econometric Modelling in the National Economic Planning.

Econometrics as a science is dealing with mathematical and statistical methods for investigation of quantitative regularities within economic phenomena.

Although the econometric methods were developed in the beginning mainly for studying capitalist economies, they can play even more important positive role in socialist countries with a Central Planned Economy. Econometric modelling within the Central Planned Economy is strictly connected with the economic planning and management and its main task is to help and support the planners and managers by providing reliable information and scientific methods for the analysis and forecasting of economic phenomena.

In order to discuss in detail the place and the role of econometrics in the National Economic Planning, let us take the following working definition of the National Economic Plan:

The National Economic Plan is the governmental decision for all sectors of the national economy and their units giving them real tasks and aims in order to reach a fixed numerical proportion in creation and distribution of the Gross National Product during a given period of time.

Although the National Economic Plan is a deterministic decision, its actual realization in real socio-economic life has always a stochastic (random) character. That is because of the human factor involved in an economic system and also because of some random disturbances which always may happen to the economy. As a clear example we can consider the agricultural sector of the National Economy, in which the weather (as a pure stochastic factor) has a great impact on final crop production.

We can illustrate this by the following formula:

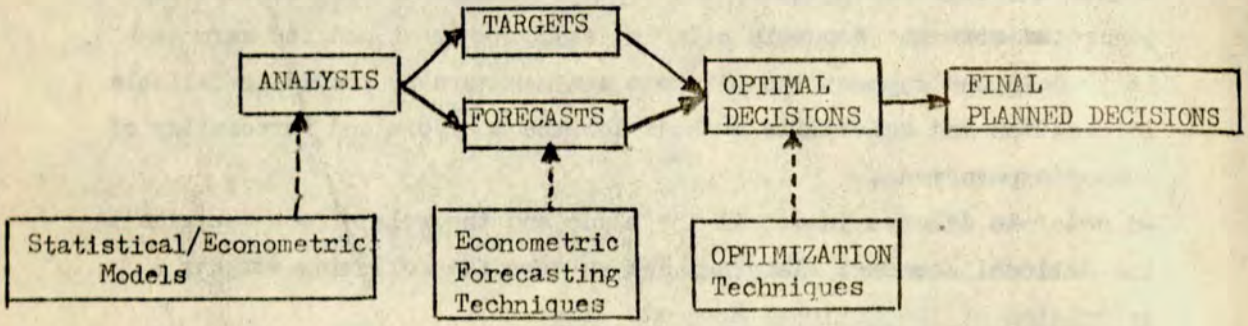
$$R = P + d$$

where P is the planned decision (for example planned wheat production)
R is the actual realization of the plan(final wheat production)
d is the random disturbance(final difference between planned and actual wheat production).

So, the actual realization R of any plan as a function of random variable d, is itself a random variable.

Then it follows that even in the Central Planned Economy any economic variable is in fact a random variable and should be analysed by means of statistical and econometric tools.

In order to distinguish proper statistical/econometric tools which should be used by planners, it is convenient to split the whole planning process into the following main steps (on the scheme below):



From the scheme, which gives a sequence of main steps of planners' work, it follows that the economic planning process should start from the Analysis of all important characters and variables describing the actual status of development in the particular sector of the National Economy. Such an Analysis should give us the true picture of attained growth and should show us all weak points, as well as strong points of the economic development observed during the last period of time. During this first stage of the planning process, mainly statistical and Econometric Models of analysis should be applied in order to obtain comprehensive objective information on the actual situation and trends in economic development, based on past statistical data.

Therefore situated below the Analysis box, the Statistical/Econometric Models box contains the whole set of methods which can be used for analytical purposes (including classical estimation procedures for population means, variances, correlations, regression functions and statistical hypotheses testing procedures, as well as econometric production functions and demand functions with estimated elasticities).

The second step of the planning process consists of the building of Targets and Forecasts independently. These are two different kinds of information about the future. By the Targets we mean values of important economic variables (or proportions) which we would like to obtain in the future for a regular development of the National Economy. The targets can be formulated based on main goals and principles of social and economic policy in the country, and on results of analysis of the actual status. Simplifying, we can say that the targets are our best wishes about future behavior of important economic variables.

On the contrary, the Forecasts are defined as objective (scientific) estimates of what will probably hapen in the future as the realization of important economic variables. More precisely, we can define a forecast as a scientifically based estimate of the future realization of some important economic variable. The term "prediction" is also used sometimes as having the same meaning as "forecasting".

Forecasts are usually calculated based on econometric models and forecasting techniques, with utilization of all historical data on forecasted phenomena. The methods included in the box Econometric Forecasting Techniques (situated below the Forecasts box) will be our special interest in Part B of this Handbook.

The last step of the economic planning process consists of decision making on final planned decisions by finding some kind of compromise between the target figures and the forecasts. To obtain the optimal solution for that compromise, planners should utilize the special Optimization Techniques (out of the rich set of Operations Research methods, especially the Mathematical Programming Techniques for getting the optimal decisions assuming some criterion of optimality connected with the Targets. The box Optimization Techniques situated below the box Optimal Decisions is not under our consideration in this Handbook, since it contains many deterministic (non-stochastic) techniques. Obviously the optimal decisions obtained as the optimal solutions of applied Optimization Techniques are sometimes not applicable by planners for some reasons and have to be changed into other final planned decisions. Anyway in such cases, the planners should know at least how far the final applicable planned decisions are from the optimal decisions obtained from the theoretical model.

From such a review of the economic planning process, the role of Econometrics became clear. The econometric models can be very useful for planners during

their analytical work, as well as during the next stages, since forecasting (providing basic objective information on future) is essential for final planned decisions.

1.2. The Objectives, Structure and Main Stages of Econometric Modelling.

Although sometimes the broad sense of Econometrics is also used in the literature, as the whole set of mathematical method aplicable for a quantitative economic analysis, we shall rather use a strict sense of Econometrics.
We can define Econometrics in the strict sense as the science which studies the statistical relationship between different economic variables based on empirical data.
The basic tool of Econometrics in the strict sense is called the Econometric Model, and it is defined as a stochastic equation(or system of equations) describing the statistical(regressional) relationship between the important economic variables observed in a given economic phenomenon. As we pointed out before, the economic variables have always(even in a Central Planned Economy) a random character. Based on empirical statistical data we are able to investigate the economic variables and the existing relationships between them by means of some special statistical methods known as econometric modelling.

We can distinguish 3 general objectives of econometric modelling:

a) Economic/Numerical Analysis,
b) Forecasting,
c) Optimal Control.

Such a sequence of the main objectives of econometric modelling also reflects the historical development of econometric theory and practice, and it is consistent with the increase in complexity of the problems which appear for these objectives.

a) Historically the first and rather simple purpose of econometric modelling was just an economic/numerical analysis of multiple regression functions between some economic variables. Such exercises started about 1920 for demand analysis and for production functions analysis purposes. The analysis of the numerical values estimated for such important parameters as elasticities could add a lot of new conclusions to the economic analysis.

b) The second and much more complicated purpose of econometric modelling which appeared during and just after World War II, is the forecasting of important economic variables connected with the development of the national economy, like Gross National Product or Private Expenditure. Now-a-days this forecasting objective of econometric modelling has become the most important one.

c) Very recently has appeared the third and rather complicated purpose of econometric modelling, namely the optimal control of some important target variables. The optimal control problem can be defined as the problem of finding such optimal values(for given optimality criterion and for given future periods of time) of some instrumental variables, which give us the needed future values (fixed in advance) of some important target variables. It is in some sense the reverse problem to forecasting, where we are looking for future value of target (dependent) variable for the given values of independent (instrumental) variables.

Depending upon which objective we are especially interested in during econometric modelling, some special forms of considered models should be prepared and some specific steps of work should be carried out. Nevertheless we can describe such a general structure of econometric models and such main stages of the econometric modelling which are common for all objectives.

The general structure of an econometric model can be written as:

$$\mathbb{Y} = f(\mathbb{X}, \Theta, \varepsilon)$$

where $\mathbb{Y}$ is the vector of dependent(explained) variables,

$\mathbb{X}$ is the vector of independent(explanatory) variables,

Θ is the unknown parameters vector,

ε is the random disturbance term,

f is the function of given type.

In order to find out the best general structure of the econometric model reflecting given economic phenomenon, we shall pass through some main stages of econometric modelling.

There are 3 main stages of econometric modelling:

1. Formulation of the model.
2. Estimation of the model.
3. Testing the model.

1. The first stage of econometric modelling consists of several problems:

- choice of the set of dependent variables(at least one) Y
- choice of the set of independent(explanatory) variables X
- choice of the type of function "f"
- choice of the stochastic structure of random disturbance term ε,

i.e. the choice of the set of stochastic assumptions on its parameters (expectation, variances, dependency and distribution type).

Some of these assumptions can be changed when the model is tested.

2. The second stage of an econometric modelling is the statistical estimation of all unknown parameters(mainly θ) of the model.
Based on given statistical data(time series or cross-sectional type) the numerical values of the unknown parameters should be obtained. For this purpose several estimation techniques can be considered (Ordinary Least Squares method, Generalized Least Squares Method, Two Stage Least Squares method, Bayesian method etc) and as appropriate one should be chosen and applied for getting numerical estimates for the unknown parameters.

3. The third stage of econometric modelling (before utilizing the model for instance for forecasting purposes) consists of applying many statistical tests for checking the main structure of the model(i.e. checking the accepted basic assumptions) and its fitting to the empirical data. The testing of the model is a very important stage of econometric modelling and should never be neglected. It needs the full analysis of residuals performed, keeping the proper sequence of applied statistical tests.

The problems enumerated above in the 3 main stages of econometric modelling will be discussed in detail in the second Part B of this handbook as well as the main application of econometric modelling for forecasting purposes. Many numerical examples mainly in an agricultural context will also be included for illustration and better explanation of the presented methods. Because the econometric methods need a rather good knowledge in mathematical analysis, linear algebra, probability theory and statistics, in the next few chapters in this Part A we would like to remind to the reader the most essential topics from that mathematical background of econometric medelling.

CHAPTER 2. MATHEMATICAL ANALYSIS BACKGROUND OF ECONOMETRICS

2.1. Sets and Operations on Sets

A set is a collection of any type of objects or numbers.

Measurements lead to the sets of numbers, which will be essential for us.

The members of a set are called elements of that set.

To describe a set by its elements, two basic notations are used:

(i) listing of elements (usually for finite sets)

(ii) giving a statement of the properties of the elements (usually for infinite sets)

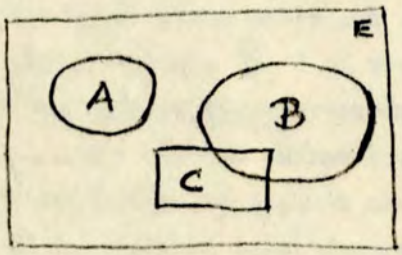
Examples for these two notations are as follows:

$A = \{2,4,6,8,10\}$, $B = \{0,1,2,\ldots, 10\}$, for (i)

$C = \{x : x \geqslant 0\}$, $D = \{x \mid 2 < x < 5\}$, $E = \{(x,y) \mid x \geqslant 1,\ y \geqslant 1\}$ for (ii)

Capital letters A,B,C,... are usually used to denote sets.

Sets can also be represented by circular or similarly enclosed curves, commonly known as Venn diagrams:



Two special sets are: the empty (null) set denoted by ϕ , which contains no elements,

the universal set (space) denoted by E, which contains all possible elements.

The notation $\in$ ("belongs to") stands for a membership of element of sets:

$x \in A$, $y \in Y$, $z \notin X$ (z does not belong to X).

The notation $\subset$ (inclusion) stands for a membership of whole sets (subsets):

$A \subset B$, $C \subset E$

Equal sets have exactly the same elements: $A = B$.

It follows, that if $A \subset B$ and $B \subset A$, then $A = B$.

To obtain new sets from the basic sets, the following operations on sets can be used : complementation, union, intersection, difference.

Complement of a set.

If A is a subset of the universal set E, then the collection of all elements of E not belonging to A is called the complement of A and it is denoted by A'.

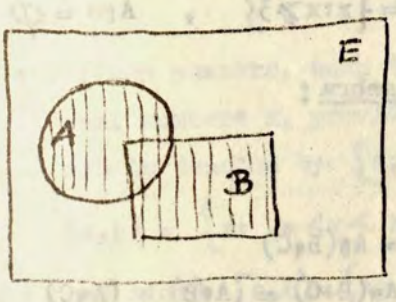
Examples: 1. $E = \{1,2,3,4,5,6\}$, $A = \{1,3,5\}$, $A' = \{2,4,6\}$.

2. E = R(space of all real numbers), $A = \{x : x \geq 2\}$,

$A' = \{x : x < 2\}$.

Union of sets.

The union of two sets A and B, denoted by $A \cup B$ or $A + B$, is the set of all elements belonging to A or B or to both A and B.



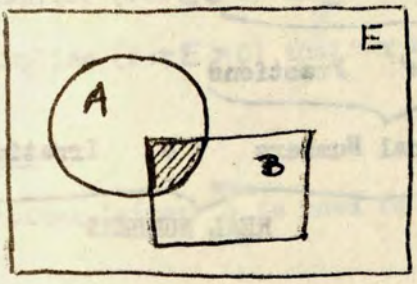
||||| represents $A \cup B$

Examples: 1. $A = \{1,2,3,4\}$, $B = \{2,4,6\}$, $A+B = A \cup B = \{1,2,3,4,6\}$

2. $A = \{x \mid x \geq 2\}$, $B = \{x \mid 1 \leq x \leq 3\}$, $A+B = \{x \mid x \geq 1\}$

Intersection of sets.

The intersection of two sets A and B, denoted by $A \cap B$ or $A \cdot B$, is the set of all elements which belong to both A and B.



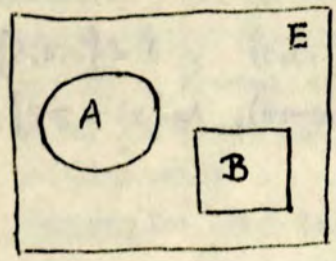
//// represents $A \cap B$

Examples. 1. $A = \{1,2,3,4\}$, $B = \{2,4,6\}$, $A \cap B = \{2,4\}$

2. $A = \{x: x \geq 2\}$, $B = \{x: 1 \leq x \leq 3\}$

$A \cap B = \{x : 2 \leq x \leq 3\}$

If the intersection of two sets does not exist (is an empty set), then these two sets are called disjoint sets.



$A \cap B = \phi$

A and B are disjoint sets

Examples: 1. $A = \{1,2\}$, $B = \{4,5,6\}$, $A \cdot B = \phi$

2. $A = \{x: x \leq 1\}$, $B = \{x: x \geq 3\}$, $A \cdot B = \phi$

There are few basic laws of set algebra :

1. $A+A = A$, $A \cdot A = A$
2. $A+B = B+A$, $A \cdot B = B \cdot A$
3. $(A+B)+C = A+(B+C)$, $(A \cdot B) \cdot C = A \cdot (B \cdot C)$
4. $A+(B \cdot C) = (A+B) \cdot (A+C)$, $A \cdot (B+C) = (A \cdot B) + (A \cdot C)$
5. $A + \phi = A$, $A+E = E$, $A \cdot E = A$, $A \cdot \phi = \phi$
6. $A+A' = E$, $(A')' = A$, $A \cdot A' = \phi$, $E' = \phi$, $\phi' = E$
7. $(A+B)' = A' \cdot B'$, $(A \cdot B)' = A'+B'$ (De Morgan's Laws)

Classification of sets of numbers:

Negative integers, $\{-1, -2, \ldots\}$ Zero , $\{0\}$ Natural numbers $\{1,2,\ldots\}$

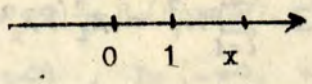
Integers + Fractions

Rational Numbers + Irrational Numbers

REAL NUMBERS

So, any real number is either rational or irrational, and it is related to some corresponding point x on the real number line

$$R = \{x \mid x \text{ is real number}\}$$



Within the set of real numbers we can do all algebraic operations

except: 1. dividing by zero , $\frac{1}{x} \Rightarrow x \neq 0$

2. taking an even root of negative numbers $\sqrt[2n]{x} \Rightarrow x \geq 0$

3. taking the logarithm of negative numbers and of zero $\log x \Rightarrow x > 0$

Intervals.

Let a and b be two given numbers, such that $a < b$.
Then a set of all real numbers x, provided $a \leq x \leq b$ is called a closed interval and is denoted by $[a,b] = \{x: a \leq x \leq b\}$

Open interval $(a,b) = \{x: a < x < b\}$

Half open(right hand side) $[a,b) = \{x: a \leq x < b\}$

Half open(left hand side) $(a,b] = \{x: a < x \leq b\}$

Infinite intervals: $(-\infty, a)$, $(b, +\infty)$, $(-\infty, +\infty) = R$

The modulus (or the absolute value) of real number

$$|x| = \begin{cases} x & \text{if } x \geq 0 \\ -x & \text{if } x < 0 \end{cases}$$

$|x| < \varepsilon$ implies (for $\varepsilon > 0$) that $-\varepsilon < x < \varepsilon$, i.e. $x \in (-\varepsilon, \varepsilon)$

$|x| > \varepsilon$ implies (for $\varepsilon > 0$) that $x < -\varepsilon$, or $x > \varepsilon$.

Sigma notation

The Greek capital letter $\sum$ is used to facilitate the writing of sums of numbers:

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \ldots + x_n$$

Examples: 1. $\sum_{i=1}^{5} i^2 = 1^2+2^2+3^2+4^2+5^2$

2. $\sum_{k=1}^{4} \frac{1}{k} = 1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}$

3. $\sum_{i=1}^{n} x_i^2 = x_1^2+x_2^2+\ldots+x_n^2$

4. $\sum_{k=1}^{n} k = 1+2+\ldots+n = \frac{n(n+1)}{2}$

5. $\sum_{k=1}^{n} k^2 = 1^2+2^2+\ldots+n^2 = \frac{n(n+1)(2n+1)}{6}$

Quantifiers are very useful logical symbols to abreviate of mathematical sentences:

- universal $\bigwedge_x$ (for all x, or for every x)

- existential $\bigvee_x$ (for some x, or there exists an x, such that)

Example. The definition of a limit 'g' of the sequence of numbers $\{a_n\}$ can be written as:

g is the limit of the sequence $\{a_n\}$ if $\bigwedge_{\varepsilon>0} \bigvee_{N \text{ integer}} \bigwedge_{n>N} |a_n-g|<\varepsilon$

2.2. Functions and Differentiation.

Function y =f(x)

Let X be a given set of real numbers, with elements $x \in X$, and let Y be a given another set of real numbers, with elements $y \in Y$. By the function y= f(x) we define a mapping (or transformation) of x into y, such that for each $x \in X$ exists only one number $y \in Y$.

Briefly: $f: X \to Y$; $\bigwedge_{x \in X} \bigvee_{y \in Y} y=f(x)$

only one

The set X is called the domain of the function f(x) , and the set Y is called the range of the function f(x).

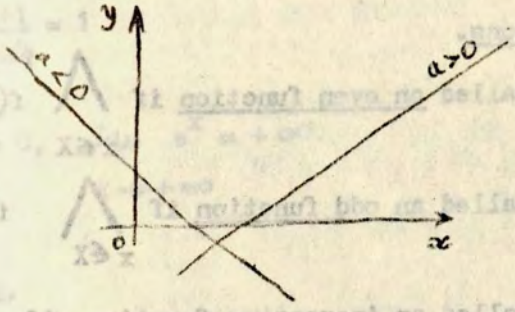
Examples: 1. $y = x^2 + 1$, $X = (-\infty, +\infty) = R$ (domain), $Y = [1, +\infty)$ (range)

2. $y = \log(x-2)$, $X = (2, +\infty)$, $Y = (-\infty, +\infty) = R$

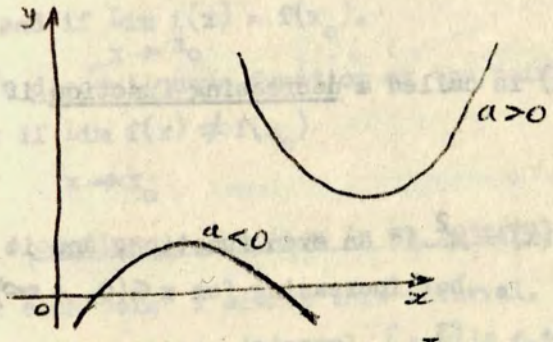
Graphs of function.

The graph of a function $y = f(x)$ consists of all such points in the plane with coordinates $(x, f(x))$.

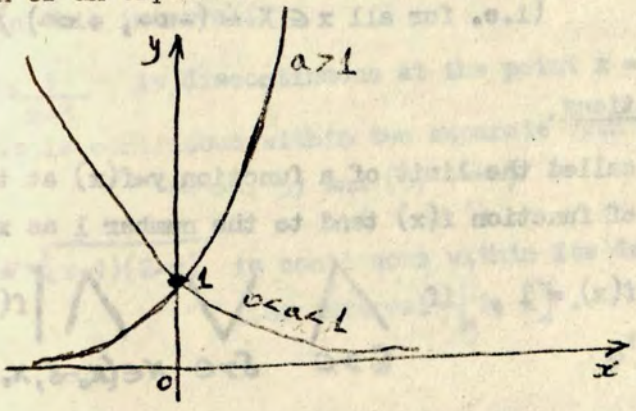
Examples: 1. Graphs of a linear function $y = ax + b$ is a straight line:



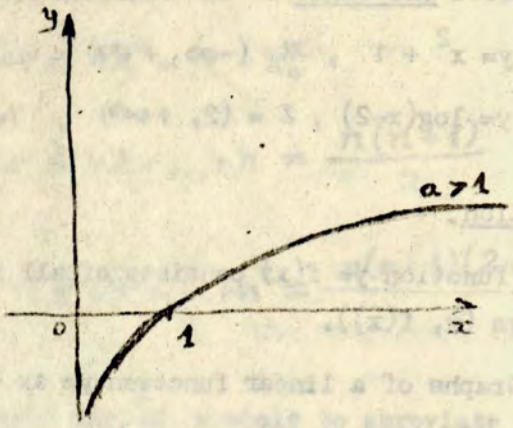
2. Graph of a quadratic function $y = ax^2 + bx + c$ is a parabola:



3. Graph of an exponential function $y = a^x$ is an exponential curve:



4. Graph of a logarithmic function $y = \log_a x$ is a logarithmic curve:



Properties of functions.

A function $f(x)$ is called an even function if $\bigwedge_{x \in X} f(-x) = f(x)$

A function $f(x)$ is called an odd function if $\bigwedge_{x \in X} f(-x) = -f(x)$

A function $f(x)$ is called an increasing function if $\bigwedge_{x_1 < x_2} f(x_1) < f(x_2)$

A function $f(x)$ is called a decreasing function if $\bigwedge_{x_1 < x_2} f(x_1) > f(x_2)$

Examples: 1. $f(x) = x^2$ is an even function, and it is decreasing for $x \in (-\infty,$ but increasing for $x \in (0, +\infty)$

2. $f(x) = x^3$ is an odd function, and it is always increasing (i.e. for all $x \in X = (-\infty, +\infty)$).

Limits of functions.

A number "l" is called the limit of a function $y=f(x)$ at the given point x_0 if the values of function $f(x)$ tend to the number l as x tends to the value x_0.

Strictly $\lim_{x \to x_0} f(x) = l$, if $\bigwedge_{\varepsilon>0} \bigvee_{\delta>0} \bigwedge_{x \in (x_0-\delta, x_0+\delta)} |f(x) - l| < \varepsilon$

$\lim_{x \to \infty} f(x) = 1$ means, that the function $f(x)$ has the limit 1 as x tends to infinity.

$\lim_{x \to \infty} f(x) = \infty$ means, that the function $f(x)$ has not a finite limit as x tends to infinity, but also tends to infinity.

Examples. 1. $\lim_{x \to 3}(x^2 - 4x + 3) = 0$

2. $\lim_{x \to 2} \frac{x^2 - 4}{x-2} = 4$, $\lim_{x \to \infty} \frac{x^2-4}{x-2} = \infty$

3. $\lim_{x \to \infty} \frac{x^2+1}{x^2-1} = 1$

4. $\lim_{x \to -\infty} e^x = 0$, $\lim_{x \to +\infty} e^x = +\infty$.

Continuity of functions,

A function $f(x)$ is a continuous function at the given point x_0, if its value $f(x_0)$ exists and if $\lim_{x \to x_0} f(x) = f(x_0)$.

A function $f(x)$ is called a discontinuous function at the point x_0, if $f(x_0)$ does not exist or if $\lim_{x \to x_0} f(x) \neq f(x_0)$

A function $f(x)$ is called a continuous function in an interval $[a,b]$ if it is continuous at the each point x within this interval.

The graph of a continuous function in an interval $[a,b]$ is a continuous un-broken curve within this interval.

Examples: 1. $f(x) = \frac{1}{x-3}$ is discontinuous at the point $x = 3$ (it is continuous within two separate open intervals $(-\infty, 3)$ and $(3, +\infty)$

2. $f(x) = \sqrt{(x-1)(2-x)}$ is continuous within its domain, i.e. interval $[1, 2]$.

3. $f(x) = e^{-x}$ is always continuous i.e. for each x.

Differentiation.

The derivative of a function $f(x)$ with respect to x is defined as

$$f'(x) = \frac{dy}{dx} = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = \lim_{\Delta x \to 0} \frac{f(x+\Delta x) - f(x)}{\Delta x}$$

The derivative $f'(x)$ is also a function of x, showing the marginal rate of change of the given function $y = f(x)$ at each particular point x. Geometrically, the value of $f(x)$ is equal to the tangent of the angle which the tangent line to the curve $y = f(x)$ at the point x makes with the x-axis. There are some important rules and formulas for differentiating particular types of functions:

$$\left[c\, f(x)\right]' = c\, f'(x)$$

$$\left[f(x) + g(x)\right]' = f'(x) + g'(x)$$

$$\left[f(x) \cdot g(x)\right]' = f'(x) \cdot g(x) + f(x) \cdot g'(x)$$

$$\left[\frac{f(x)}{g(x)}\right]' = \frac{f'(x)\cdot g(x) - f(x)\cdot g'(x)}{g^2(x)}$$

For composite function $y = f\left[g(x)\right] = f(u)$ where $u = g(x)$, we have

$$y' = f'(u) \cdot g'(x)$$

Formulas for some elementary functions: $(c)' = 0$, $(ax + b)' = a$

$$(x^\alpha)' = \alpha x^{\alpha-1} , \quad (\sqrt{x})' = \frac{1}{2\sqrt{x}}$$

$$(e^x)' = e^x , \quad (\sin x)' = \cos x$$

Examples:

1. $y = 5x^3 + 4x^2 - 2x + 1$, $y' = 15x^2 + 8x - 2$.

2. $y = x^2 e^x$, $y' = 2xe^x + x^2 e^x = xe^x(2 + x)$

3. $y = \frac{3x}{x^2+1}$, $y' = \frac{3(x^2+1) - 3x \cdot 2x}{(x^2+1)^2} = \frac{3-3x^2}{(x^2+1)^2}$

4. $y = \sqrt{1+x^2}$, $y' = \frac{1}{2\sqrt{1+x^2}} \cdot 2x = \frac{x}{\sqrt{1+x^2}}$

5. $y = \sin(3x^2 + 2)$, $y' = \cos(3x^2+2) \cdot 6x$

Sign of a derivative.

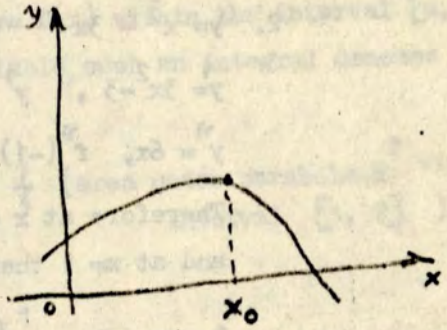
If the derivative $f'(x) > 0$ for $x \in [a,b]$ then the function $y = f(x)$ is an increasing function within this interval.

If the derivative $f'(x) < 0$ for $x \in [a,b]$ then the function $y = f(x)$ is a decreasing function within this interval.

Maxima and minima of functions.

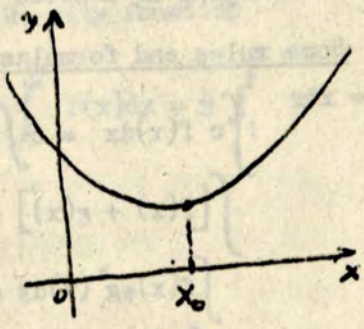
A function $f(x)$ has a maximum value $f(x_0)$ at the point x_0

if $\bigvee_{\varepsilon>0} \bigwedge_{x \in (x_0-\varepsilon, x_0+\varepsilon)} f(x_0) > f(x)$



A function $f(x)$ has a minimum value $f(x_0)$ at the point x_0

if $\bigvee_{\varepsilon>0} \bigwedge_{x \in (x_0-\varepsilon, x_0+\varepsilon)} f(x_0) < f(x)$



The necessary condition for existing a minimum (or a maximum) value $f(x_0)$ at the given point x_0 , is that $f'(x_0) = 0$.

The sufficient condition (additional to the above necessary condition)

- for the minimum value is $f''(x_o) > 0$ (where $f''(x)$ denotes the second derivative)
- for the maximum value is $f''(x_o) < 0$

Examples. 1. $y = e^{-\frac{x^2}{2}}$

$$y' = -x e^{-\frac{x^2}{2}}, \quad y' = 0 \text{ for } x = 0$$

$$y'' = (x^2 - 1) e^{-\frac{x^2}{2}}, \quad f''(0) = -e^0 = -1 < 0$$

Therefore at $x_o = 0$ there is a maximum equal to $f(0) = 1$

2. $y = x^3 - 3x$

$y' = 3x^2 - 3$, $y' = 0$ for $x = -1$ and $x = 1$

$y'' = 6x$, $f''(-1) = -6$, $f''(1) = 6$

Therefore at $x = -1$ there is a maximum equal to $f(-1) = 2$, and at $x = 1$ there is a minimum equal to $f(1) = -2$.

Integration.

An integral of function $f(x)$ with respect to x is defined as a function $F(x)$ such that $F'(x) = f(x)$.

An integral is denoted by $\int f(x)dx = F(x) + C$, where C is any constant

Some rules and formulas for integration:

$$\int c\, f(x)dx = c \int f(x)dx$$

$$\int [f(x) + g(x)]\, dx = \int f(x)dx + \int g(x)dx$$

$$\int f(x) \cdot g'(x)dx = f(x) \cdot g(x) - \int g(x) \cdot f'(x)dx \quad \text{(integration by parts)}$$

$$\int f[g(x)] \cdot g'(x)dx = \int f(u)du, \text{ where } u = g(x) \quad \text{(integration by substitution)}$$

$$\int x^\alpha dx = \frac{x^{\alpha+1}}{\alpha+1} \quad \text{for } \alpha \neq -1$$

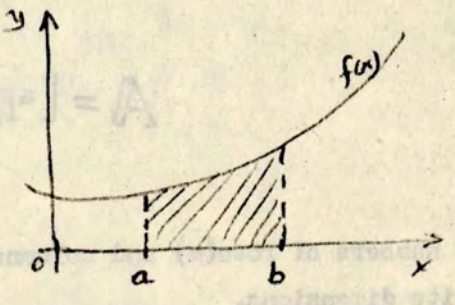
$$\int \frac{1}{x} dx = \ln x$$

$$\int e^x dx = e^x$$

$$\int a^x dx = \frac{a^x}{\ln a}$$

Definite integral

$$\int_a^b f(x)dx = \Big[F(x)\Big]_a^b = F(b) - F(a)$$



Geometrically it is the area under the curve y= f(x) within the interval $[a,b]$
If f(x) is a density function of a random variable, such an integral denotes probability for the interval $[a\ ,\ b]$

Examples. 1. $\int_1^2 x^2 dx = \frac{x^3}{3}\Big]_1^2 = \frac{8}{3} - \frac{1}{3} = \frac{7}{3}$ (area under parabola x^2 within interval $[1,\ 2]$)

2. $\int_0^1 \sqrt{x}\, dx = \int_0^1 x^{\frac{1}{2}} dx = \frac{x^{\frac{3}{2}}}{\frac{3}{2}}\Big]_0^1 = \frac{2}{3} - 0 = \frac{2}{3}$

3. Let $f(x) = \frac{1}{2}x$ for $0 \leq x \leq 2$ is the density function of some random variable X
Then probability $P\{1 \leq X \leq 2\} = \int_1^2 f(x)dx = \frac{1}{2}\int_1^2 x dx =$

$$= \frac{x^2}{4}\Big]_1^2 = 1 - \frac{1}{4} = \frac{3}{4} = 0.75$$

CHAPTER 3. LINEAR ALGEBRA BACKGROUND OF ECONOMETRICS.

3.1. Matrices and operations on matrices.

A matrix $\mathbb{A}$ is a rectangular array of numbers. The elements of a matrix $\mathbb{A}$ are denoted by a_{ij}.

$$\mathbb{A} = [a_{ij}] = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix}$$

The numbers of rows(m) and columns(n) are called the order of a matrix $\mathbb{A}$ or its dimensions.

If m=n, a matrix $\mathbb{A}$ is called the square matrix.

If m= 1, a matrix $\mathbb{A}$ has only one row and it is called a row vetor

$a = (a_1\ a_2 \dots a_n)$.

If n = 1, a matrix $\mathbb{A}$ has only one column called a column vector $a = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{pmatrix}$

If n = m = 1, a matrix $\mathbb{A}$ is a scalar(one number only).

The null matrix denoted by $\mathbb{O}$, is a matrix with all elements being zeros.

The diagonal matrix is a square matrix with all elements zeros except those in the leading diagonal

$$\mathbb{D} = \begin{pmatrix} d_{11} & 0 & \dots & 0 \\ 0 & d_{22} & \dots & 0 \\ \vdots & & & \\ 0 & 0 & \dots & d_{nn} \end{pmatrix}$$

A diagonal matrix with $d_{11} = \dots = d_{nn} = 1$ is called the Identity matrix denoted by $\mathbb{I} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ 0 & 0 & \dots & 1 \end{pmatrix}$

A square matrix with all elements below (or above) the leading diagonal equal to zero, is called the triangular matrix

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ 0 & a_{22} & \cdots & a_{2n} \\ \vdots & & & \vdots \\ 0 & 0 & \cdots & a_{mn} \end{pmatrix}$$

The transpose of a matrix A denoted by A^T or by A' is a matrix with interchanging the rows and columns.

If $A^T = A$ we have a symmetric matrix.

Example

$$A = \begin{pmatrix} 2 & 3 & 4 \\ 1 & 1 & 2 \end{pmatrix} \qquad A^T = \begin{pmatrix} 2 & 1 \\ 3 & 1 \\ 4 & 2 \end{pmatrix}$$

$$A = \begin{pmatrix} 2 & 3 & 4 \\ 3 & 5 & 6 \\ 4 & 6 & 7 \end{pmatrix} = A^T \quad \text{(a symmetric matrix)}$$

For addition $A+B$, both matrices must have the same dimensions m and n. The sum of two matrices of the same dimensions is a matrix $A+B$ with elements which are the sums of the corresponding elements of the given matrices.

$$\begin{pmatrix} 2 & 3 & 1 \\ 4 & 5 & 1 \end{pmatrix} + \begin{pmatrix} 0 & 2 & 3 \\ 1 & 1 & 1 \end{pmatrix} = \begin{pmatrix} 2 & 5 & 4 \\ 5 & 6 & 2 \end{pmatrix}$$

The product of a matrix A by a scalar c, (scalar multiplication) is a matrix cA with elements being c times the elements of given matrix:

$$3\begin{pmatrix} 1 & 2 & 3 \\ 2 & 0 & 3 \end{pmatrix} = \begin{pmatrix} 3 & 6 & 9 \\ 6 & 0 & 9 \end{pmatrix}$$

For multiplication of two matrices A and B, the number of columns in A must be equal to the number of rows in B:

$$A_{m \times n} \cdot B_{n \times k} = C_{m \times k}, \quad \text{where } c_{ij} = \sum_{r=1}^{n} a_{ir} b_{rj}$$

So, the procedure of multiplication of two matrices is such, that each row of A is multiplied by each column of B .

Examples:

$$\begin{pmatrix} 2 & 1 & 0 \\ 1 & 2 & 3 \end{pmatrix} \cdot \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$$ the product does not exist

$$\begin{pmatrix} 2 & 1 & 0 \\ 1 & 2 & 3 \end{pmatrix} \cdot \begin{pmatrix} 3 \\ 4 \\ 2 \end{pmatrix} = \begin{pmatrix} 10 \\ 17 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 2 \\ 2 & 3 \\ 3 & 4 \end{pmatrix} \cdot \begin{pmatrix} 2 & 1 & 0 & 2 & 1 \\ 1 & 1 & 1 & 0 & 2 \end{pmatrix} = \begin{pmatrix} 4 & 3 & 2 & 2 & 5 \\ 7 & 5 & 3 & 4 & 8 \\ 10 & 7 & 4 & 6 & 11 \end{pmatrix}$$

$$(1\ 2\ 3) \cdot \begin{pmatrix} 2 & 1 \\ 3 & 0 \\ 1 & 2 \end{pmatrix} = (11\ 7)$$

For any matrix A, we have $A \cdot I = I \cdot A = A$

In general however $AB \neq BA$, as ilustrated by the following example:

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}, B = \begin{pmatrix} 1 & 5 \\ 2 & 2 \end{pmatrix}, AB = \begin{pmatrix} 5 & 9 \\ 11 & 23 \end{pmatrix}, \text{ but } BA = \begin{pmatrix} 16 & 22 \\ 8 & 12 \end{pmatrix} \neq AB$$

The transpose of a sum $(A+B)^T = A^T + B^T$

The transpose of a product $(A \cdot B)^T = B^T \cdot A^T$

Example:

$$(A \cdot B)^T = \begin{pmatrix} 5 & 9 \\ 11 & 23 \end{pmatrix}^T = \begin{pmatrix} 5 & 11 \\ 9 & 23 \end{pmatrix}$$

$$B^T \cdot A^T = \begin{pmatrix} 1 & 2 \\ 5 & 2 \end{pmatrix} \cdot \begin{pmatrix} 1 & 3 \\ 2 & 4 \end{pmatrix} = \begin{pmatrix} 5 & 11 \\ 9 & 23 \end{pmatrix} = (A \cdot B)^T$$

The trace of a square matrix A is defined as the sum of its diagonal elements:

$$\operatorname{tr} A = \sum_{i=1}^{n} a_{ii}$$

Useful properties of the trace of a matrix are following:

$$\mathrm{tr}\, I_n = n$$

$$\mathrm{tr}(cA) = c\,\mathrm{tr}A, \qquad \mathrm{tr}(A+B) = \mathrm{tr}A + \mathrm{tr}B$$

$$\mathrm{tr}(A\cdot B) = \mathrm{tr}(B\cdot A)$$

Example: $A=\begin{pmatrix}1 & 2\\ 3 & 4\end{pmatrix}$, $B=\begin{pmatrix}1 & 5\\ 2 & 2\end{pmatrix}$, $A+B=\begin{pmatrix}2 & 7\\ 5 & 6\end{pmatrix}$, $A\cdot B=\begin{pmatrix}5 & 9\\ 11 & 23\end{pmatrix}$, $B\cdot A=\begin{pmatrix}16 & 22\\ 8 & 12\end{pmatrix}$

$\mathrm{tr}(A+B) = 2+6=8 = \mathrm{tr}A + \mathrm{tr}B = (1+4) + (1+2)$

$\mathrm{tr}(AB) = 5+23= 28 = \mathrm{tr}(BA) = 16 + 12$

The determinant of a square matrix is a number denoted by $|A| = \det A$ associated with the matrix A in a special way, which is given below for a matrices of 2x2 and 3x3 dimensions:

$$\begin{vmatrix} a & b\\ c & d\end{vmatrix} = ad - bc$$

$$\begin{vmatrix} a & b & c\\ d & e & f\\ g & h & i\end{vmatrix} = aei + bfg + dhc - gec - dbi - hfa$$

Example: $\begin{vmatrix} 2 & 1\\ -1 & 1\end{vmatrix} = 2 -(-1) = 2+1=3$

$$\begin{vmatrix} 1 & 2 & -1\\ 0 & 1 & 3\\ 2 & 1 & 2\end{vmatrix} = 2+12+0+2+0-3=13$$

Determinant of a product of two matrices.

$$|A\cdot B| = \det(A\cdot B) = \det A \cdot \det B$$

Example:

$$A=\begin{pmatrix}1 & 1 & 1\\ 4 & -1 & 1\\ 0 & 1 & -1\end{pmatrix}, \quad B=\begin{pmatrix}2 & 2 & 1\\ 1 & 0 & 2\\ 2 & 1 & 2\end{pmatrix}, \quad A\cdot B=\begin{pmatrix}5 & 3 & 5\\ 9 & 9 & 4\\ -1 & -1 & 0\end{pmatrix}$$

$$\det(A\cdot B)=\begin{vmatrix}5 & 3 & 5\\ 9 & 9 & 4\\ -1 & -1 & 0\end{vmatrix} = 8 = \det A \cdot \det B = \begin{vmatrix}1 & 1 & 1\\ 4 & -1 & 1\\ 0 & 1 & -1\end{vmatrix} \cdot \begin{vmatrix}2 & 2 & 1\\ 1 & 0 & 2\\ 2 & 1 & 2\end{vmatrix} = 8\cdot 1$$

A matrix A is called <u>singular</u> if $\det A = 0$.

If a matrix A is singular, it has <u>linearly dependent</u> rows or columns.

<u>Example</u>:

$$A = \begin{pmatrix} 1 & 1 & 2 \\ 2 & 1 & 3 \\ -1 & 1 & 0 \end{pmatrix}, \quad \det A = 0, \quad A \text{ is singular and its}$$

columns are linearly dependent(third column is the sum of the first two columns).

The <u>minor</u> of an element a_{ij} of a matrix A is a determinant of the matrix obtained by deleting the i-th row and the j-th column.

The <u>cofactor</u> of an element a_{ij} is its minor multiplied by $(-1)^{i+j}$.

The <u>cofactor matrix</u> is the matrix of cofactors.

The <u>adjoint matrix</u> (denoted by A^D) of a matrix A is the transpose of the cofactor matrix.

<u>Example</u>:

$$A = \begin{pmatrix} 1 & 2 \\ -1 & 0 \end{pmatrix}, \quad A^D = \begin{pmatrix} 0 & 2 \\ 1 & 1 \end{pmatrix}$$

The <u>inverse of a matrix</u> A denoted by A^{-1} is defined only for a non-singular matrix, as $A^{-1} = \dfrac{A^D}{\det A}$

<u>Examples</u>:

$$A = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}, \quad \det A = 1 \neq 0$$

$$A^D = \begin{pmatrix} 1 & -2 \\ 0 & 1 \end{pmatrix}, \quad A^{-1} = \frac{A^D}{\det A} = \begin{pmatrix} 1 & -2 \\ 0 & 1 \end{pmatrix}$$

$$B = \begin{pmatrix} 1 & 0 & 1 \\ 2 & 1 & 1 \\ 0 & 1 & 1 \end{pmatrix}, \quad \det B = 2 \neq 0,$$

$$B^D = \begin{pmatrix} 0 & 1 & -1 \\ -2 & 1 & 1 \\ 2 & -1 & 1 \end{pmatrix}, \quad B^{-1} = \begin{pmatrix} 0 & \frac{1}{2} & -\frac{1}{2} \\ -1 & \frac{1}{2} & \frac{1}{2} \\ 1 & -\frac{1}{2} & \frac{1}{2} \end{pmatrix}$$

$$C = \begin{pmatrix} 7 & -11 & 16 \\ -3 & 5 & -7 \\ 1 & -2 & 3 \end{pmatrix}, \quad C^{-1} = \begin{pmatrix} 1 & 1 & -3 \\ 2 & 5 & 1 \\ 1 & 3 & 2 \end{pmatrix}$$

Important properties of the inverse A^{-1} are:

$$A^{-1} \cdot A = A \cdot A^{-1} = I \quad \text{(identity matrix)}$$

$$(A^{-1})^{-1} = A$$

$$(A \cdot B)^{-1} = B^{-1} \cdot A^{-1}$$

$$(cA)^{-1} = \frac{1}{c} A^{-1}, \quad c \neq 0$$

The rank of a matrix A, denoted by $r(A)$ is the dimension of the largest non-zero determinant which can be build from matrix A.

The rank of a matrix A is equal to the maximum number of its linearly independent rows(or columns).

Example:

$$A = \begin{pmatrix} 1 & -1 & 3 \\ 2 & -2 & 6 \\ -1 & 1 & -3 \end{pmatrix}, \quad \det A = 0, \quad \begin{vmatrix} 1 & -1 \\ 2 & -2 \end{vmatrix} = 0$$ as all minors of dimensions 2x2,

$1 = a_{11} \neq 0$ therefore $r(A) = 1$

$$B = \begin{pmatrix} -4 & 0 & 5 \\ 1 & 2 & 3 \end{pmatrix}, \quad \begin{vmatrix} -4 & 0 \\ 1 & 2 \end{vmatrix} = -8 \neq 0, \quad r(B) = 2$$

Important properties of the rank of a matrix.

$$r(A^T) = r(A)$$

$$r(A+B) \leq r(A) + r(B)$$

$$r(A \cdot B) \leq r(A)$$

$$\leq r(B)$$

3.2. System of Linear Equations and its Solution.

Using a matrix notation, any system of n linear equations:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n = b_2 \\ \vdots \quad\quad \vdots \quad\quad \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \dots + a_{nn}x_n = b_n \end{cases}$$

can be briefly written in the form : $\mathbb{A}x = \mathbb{b}$

where: $$\mathbb{A} = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix}, \quad \mathbb{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}, \quad x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

To obtain the solution of the system of linear equations, the Cramer's rule can be used:

$$x_i = \frac{\det \mathbb{A}_i}{\det \mathbb{A}} \qquad \text{for } i = 1,2,\ldots, n$$

where $\det \mathbb{A} \neq 0$ is the determinant of non-singular coefficients matrix $\mathbb{A}$
$\det \mathbb{A}_i$ is the determinant of such a matrix in which the i-th column in matrix $\mathbb{A}$ has been replaced by the vector $\mathbb{b}$.

Examples:

1. $$\begin{cases} 2x_1 + 3x_2 = 9 \\ 3x_1 + 5x_2 = 12 \end{cases}$$

$$\det \mathbb{A} = \begin{vmatrix} 2 & 3 \\ 3 & 5 \end{vmatrix} = 1, \quad \det \mathbb{A}_1 = \begin{vmatrix} 9 & 3 \\ 12 & 5 \end{vmatrix} = 9, \quad \det \mathbb{A}_2 = \begin{vmatrix} 2 & 9 \\ 3 & 12 \end{vmatrix}$$

$$x_1 = \frac{9}{1} = 9, \quad x_2 = \frac{3}{1} = 3$$

2. $$\begin{cases} x_1 - x_3 = 4 \\ 2x_2 - x_3 = 6 \\ x_1 + 2x_2 = 10 \end{cases}$$

$$\det \mathbb{A} = \begin{vmatrix} 1 & 0 & -1 \\ 0 & 2 & -1 \\ 1 & 2 & 0 \end{vmatrix} = 4 \neq 0$$

$$\det \mathbb{A}_1 = \begin{vmatrix} 4 & 0 & -1 \\ 6 & 2 & -1 \\ 10 & 2 & 0 \end{vmatrix} = 16, \quad \det \mathbb{A}_2 = \begin{vmatrix} 1 & 4 & -1 \\ 0 & 6 & -1 \\ 1 & 10 & 0 \end{vmatrix} = 12, \quad \det \mathbb{A}_3 = \begin{vmatrix} 1 & 0 & 4 \\ 0 & 2 & 6 \\ 1 & 2 & 10 \end{vmatrix} = 0$$

$$x_1 = \frac{16}{4} = 4, \quad x_2 = \frac{12}{4} = 3, \quad x_3 = \frac{0}{4} = 0.$$

Alternatively the inverse of a matrix, $\mathbb{A}^{-1}$ can be used to obtain the solution of the system of linear equations, assuming that $\mathbb{A}$ is non singular, so $\mathbb{A}^{-1}$ exists.

Taking into account the matrix notation for a system of linear equations we have :

$$\mathbf{A}\mathbf{x} = \mathbf{b}$$

$$\mathbf{A}^{-1}\cdot\mathbf{A}\mathbf{x} = \mathbf{A}^{-1}\cdot\mathbf{b}$$ (by left-hand side multiplication by $\mathbf{A}^{-1}$)

$$\mathbf{I}\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$$

and finally $$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$$

Examples:

1. $$\begin{cases} 2x_1 + 3x_2 = 9 \\ 3x_1 + 5x_2 = 12 \end{cases}$$

$$\mathbf{A} = \begin{pmatrix} 2 & 3 \\ 3 & 5 \end{pmatrix}, \quad \det \mathbf{A} = 1 \neq 0, \quad \mathbf{A}^{-1} = \begin{pmatrix} 5 & -3 \\ -3 & 2 \end{pmatrix}$$

$$\mathbf{x} = \begin{pmatrix} 5 & -3 \\ -3 & 2 \end{pmatrix}\cdot\begin{pmatrix} 9 \\ 12 \end{pmatrix} = \begin{pmatrix} 9 \\ -3 \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

2. $$\begin{cases} x_1 - x_3 = 4 \\ 2x_2 - x_3 = 6 \\ x_1 + 2x_2 = 10 \end{cases} \qquad \mathbf{A} = \begin{pmatrix} 1 & 0 & -1 \\ 0 & 2 & -1 \\ 1 & 2 & 0 \end{pmatrix}, \quad \det \mathbf{A} = 4 \neq 0, \quad \mathbf{A}^{-1} = \tfrac{1}{4}\begin{pmatrix} 2 & -2 & 2 \\ -1 & 1 & 1 \\ -2 & -2 & 2 \end{pmatrix}$$

$$\mathbf{x} = \tfrac{1}{4}\begin{pmatrix} 2 & -2 & 2 \\ -1 & 1 & 1 \\ -2 & -2 & 2 \end{pmatrix}\cdot\begin{pmatrix} 4 \\ 6 \\ 10 \end{pmatrix} = \tfrac{1}{4}\begin{pmatrix} 16 \\ 12 \\ 0 \end{pmatrix} = \begin{pmatrix} 4 \\ 3 \\ 0 \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

3.3. Quadratic Forms and Definite Matrices.

The scalar $$\mathbf{x}^T\mathbf{A}\mathbf{x} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_i x_j$$ is called a quadratic form

of a vector $$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$, with a symmetric matrix $\mathbf{A} = [a_{ij}]$.

Examples: 1. $$\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 2 & 3 \end{pmatrix}, \quad \mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

The quadratic form $x^T A x = (x_1\ x_2)\begin{pmatrix} 1 & 2 \\ 2 & 3 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = x_1^2 + 4x_1x_2 + 3x_2^2$

2. $A = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 3 & 1 \\ 0 & 1 & 4 \end{pmatrix}$, $x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$.

The quadratic form $x^T A x = (x_1\ x_2\ x_3)\begin{pmatrix} 2 & 0 & 0 \\ 0 & 3 & 1 \\ 0 & 1 & 4 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = 2x_1^2 + 3x_2^2 + 4x_3^2 + 2x_2x_3$

A quadratic form is called <u>positive definite</u> if for all vectors $x \neq 0$, $x^T A x > 0$.
A quadratic form is called <u>positive semi-definite</u> if for all vectors $x \neq 0$, $x^T A x \geq 0$.
A symmetric matrix A is called positive definite, if its quadratic form is positive definite.
A symmetric matrix A is called positive semi-definite if its corresponding quadratic form is positive semi-definite.

<u>Example:</u> $A = \begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix}$ is positive definite, because its quadratic form $x^T A x = 2x_1^2 + 3x_2^2 > 0$ for all vectors $x \neq 0$.

<u>Useful theorems:</u>

1. If A is a nxm matrix with the rank $m < n$, then $A^T A$ is a positive definite matrix.
2. If A is a positive definite matrix, and P is a non-singular matrix, then $P^T A P$ is a positive definite matrix.

<u>Differentiation in the matrix notation</u>.

<u>The derivative of a quadratic form</u> $y = x^T A x$ with respect to the vector x is given by the formula:

$$\frac{\partial y}{\partial x} \equiv \begin{pmatrix} \frac{\partial y}{\partial x_1} \\ \vdots \\ \frac{\partial y}{\partial x_n} \end{pmatrix} = 2Ax.$$

The derivative of a linear form $y = a^T x = a_1x_1 + a_2x_2 + \dots + a_n x_n$

with respect to the vector $\mathbf{x}$ is given by the formula

$$\frac{\partial y}{\partial \mathbf{x}} = \begin{pmatrix} \frac{\partial y}{\partial x_1} \\ \vdots \\ \frac{\partial y}{\partial x_n} \end{pmatrix} = \mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix}$$

Examples: 1. $y = 2x_1^2 + 2x_1x_2 + 3x_2^2 + 4x_2x_3 + 5x_3^2 = \mathbf{x}^T \mathbf{A} \mathbf{x}$.

y is a quadratic form with the matrix $\mathbf{A} = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 3 & 2 \\ 0 & 2 & 5 \end{pmatrix}$

then its derivative

$$\frac{\partial y}{\partial \mathbf{x}} = \begin{pmatrix} \frac{\partial y}{\partial x_1} \\ \frac{\partial y}{\partial x_2} \\ \frac{\partial y}{\partial x_3} \end{pmatrix} = \begin{pmatrix} 4x_1 + 2x_2 \\ 2x_1 + 6x_2 + 4x_3 \\ 4x_2 + 10x_3 \end{pmatrix} = 2\mathbf{A}\mathbf{x}.$$

2. $y = 5x_1 + 6x_2 - 2x_3 = \mathbf{a}^T\mathbf{x}$ is a linear form of the vector $\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$

with the coefficients vector $\mathbf{a}^T = (5\ 6\ -2)$,

then its derivative $\frac{\partial y}{\partial \mathbf{x}} = \begin{pmatrix} \frac{\partial y}{\partial x_1} \\ \frac{\partial y}{\partial x_2} \\ \frac{\partial y}{\partial x_3} \end{pmatrix} = \begin{pmatrix} 5 \\ 6 \\ -2 \end{pmatrix} = \mathbf{a}$

3.4. Some Special Matrices used in Econometrics.

The Gram's matrix is defined for each matrix $\mathbf{A}_{mxn}$

as $\mathbf{G} = \mathbf{A}^T\mathbf{A}$ or $\mathbf{G}_1 = \mathbf{A}\mathbf{A}^T$

$\mathbf{G}$ is a quadratic nxn matrix, while $\mathbf{G}_1$ is quadratic mxm matrix.

There are two useful properties of Gram's matrices:

1. The Gram's matrices G and G_1 are symmetric.
2. The rank $r(G) = r(G_1) = r(A)$

Therefore the smallest Gram's matrix (out of G_1 and G_1) can be used for finding the rank of a given matrix A .

Example

$$A = \begin{pmatrix} 2 & 6 & 8 \\ 3 & 9 & 12 \end{pmatrix}$$

$$G_1 = \begin{pmatrix} 104 & 156 \\ 156 & 234 \end{pmatrix}$$

$\det G_1 = 0$

Therefore $r(G_1) = 1 = r(A)$

The idempotent matrix.

A matrix A is called an idempotent matrix, if it is symmetric i.e. $A^T = A$ and if $A^2 = A \cdot A = A$

The following properties of idempotent matrices are useful in Econometrics:

1. If A is an idempotent matrix, then $B = I - A$ is also an idempotent matrix, and $AB = BA = 0$
2. If A is any matrix, then $A(A^TA)^{-1}A^T$ is an idempotent matrix.
3. If A is an idempotent matrix, then its rank is equal to its trace, i.e. $r(A) = \operatorname{tr} A$.

The last property is specially useful, since it is very easy to find the trace of matrix.

Examples:

1. It can be easily checked, that $A = \begin{pmatrix} \frac{2}{3} & 0 & \frac{\sqrt{2}}{3} \\ 0 & 1 & 0 \\ \frac{\sqrt{2}}{3} & 0 & \frac{1}{3} \end{pmatrix}$ is an idempotent matrix.
2. Let X be nxr matrix, then from above 1-3 properties it follows, that the matrix $Z = I - X(X^TX)^{-1}X^T$ is an idempotent matrix and its rank can be obtained as:

$$r(Z) = \operatorname{tr} Z = \operatorname{tr}\left[I - X(X^TX)^{-1}X^T\right] =$$
$$= \operatorname{tr} I_n - \operatorname{tr}\left[X^T \cdot X(X^TX)^{-1}\right] =$$
$$= n - \operatorname{tr} I_r = n - r.$$

CHAPTER 4. PROBABILITY THEORY BACKGROUND OF ECONOMETRICS.

4.1. Random Events and Probability Measure.

Random experiments.

An experiment is said to be random if we cannot predict the outcome before the experiment is carried out. The set of all possible elementary single outcomes in a given random experiment is called the sample space and it will be denoted by Ω. Any subset of the sample space Ω is called an event (or random event). Events will be denoted by A, B, C, etc, like sets.

Example: If we roll a die, we are performing the random experiment for which the sample space is $\Omega = \{1,2,3,4,5,6,\}$. The following subsets of Ω can be considered as examples of events: $A = \{2,4,6,\}$, $B = \{6\}$, $C = \{1,2,3,\}$, $D = \{1,3,5\}$.

As special events we can consider the certain event denoted also by Ω which is certain to occur in a given random experiment, since it contains all possible outcomes, and the impossible event, denoted by $\emptyset$, which can never occur in a given random experiment, since it does not contain any outcome(it is equal to the empty set of outcomes),

The Borel field of events, denoted by $\mathcal{F}$, is such a set of events, that

1. $\Omega \in \mathcal{F}$
2. If $A_i \in \mathcal{F}$ for i=1,2,..., then the union $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$ and the entersection $\bigcap_{i=1}^{\infty} A_i \in \mathcal{F}$.

The Borel field of events $\mathcal{F}$ is therefore a family of all possible events and it has the important property: all events belonging to $\mathcal{F}$ are measurable sets, i.e. we can define a measure as a set function on them.

Probability measure.

A set function P is called the Probability Measure, or briefly Probability, if it fulfils the following three Kolmogorov axioms:

1. For each event $A \in \mathcal{F}$, $0 \leq P(A) \leq 1$.
2. $P(\Omega) = 1$
3. $P(\bigcup_{i=1} A_i) = \sum_{i=1}^{\infty} P(A_i)$ for the disjoint events $A_1, A_2, \ldots$

So, the probability is a non-negative, normalized and countable additive measure of events and it gives a mathematical measure for the chance of any random event.

It can be proved, that $P(A') = 1-P(A)$ for the complement A' of an event A; and $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ for any two events A and B.

The full mathematical model associated with each random experiment for getting the probability of events is called the probability space, and it is defined as a triple: $(\Omega, \mathcal{F}, P)$,

where Ω is the sample space for a given random experiment,

$\mathcal{F}$ is the Borel field of events,

P is the probability measure, which gives probability P(A) for any event A.

It follows, that in order to obtain the proper probabilities for events in a given random experiment, first of all we have to define the appropriate probability space $(\Omega, \mathcal{F}, P)$ to that trial.

In a such random experiments in which sample space is finite, the simplest so called classical probability measure can be used, which gives the probability for an event A, as $P(A) = \frac{m}{n}$, where m is the number of favourable events (the number of elementary outcomes contained in the event set A) out of a total of n possible elementary outcomes of the experiment.

Examples.

1. A box contains 4 red, 2 white and 6 black balls. When a ball is drawn at random what is the probability that it will be red?
   A red ball is selected. $P(A) = \frac{4}{12} = \frac{1}{3}$.
2. A single card is drawn at random from a deck of 52 bridge cards. What is the probability that it will be a diamond or a queen ?

   Answer: Let D denotes the event that a diamond is selected, and Q denotes the event that a queen is selected,
   then $P(D \cup Q) = P(D) + P(Q) - P(D \cap Q) = \frac{13}{52} + \frac{4}{52} - \frac{1}{52} = \frac{16}{52} = \frac{4}{13}$

4.2. Random Variables and Probability Distributions.

Let $(\Omega, \mathcal{F}, P)$ be a given probability space for some random experiment. Each real-valued function defined on the sample space Ω is called

a random variable. It is a mapping of events into real numbers.

A capital letters X, Y, ... will be used to denote random variables.

$$X : \Omega \rightarrow R .$$

With the real numbers from R which are the values of random variable X, the probabilities are associated, as each event from the domain of the random variable X has the appropriate probability.

Therefore, for each random variable we can define its probability distribution as the set of its values(the range) and probabilities connected with them.

A random variable is said to be discrete if its range is a finite or countable subset of the space of real numbers R.

A random variable is said to be continuous if its range is an interval of real numbers.

The probability distribution of a discrete random variable X can be characterized by so called probability function:

$$P(X = x_i) = p_i \quad , \text{ where } p_i \text{ are the probabilities}$$

$$0 < p_i < 1 \text{ and } \sum_i p_i = 1 .$$

In Econometrics much more important however are continuous random variables, because the numerical results of random experiments (with) measurements of economic variables belong to some intervals of real numbers.

The probability distribution of a continuous random variable X can be characterized by so called probability density function (pdf) denoted by f(x) and satisfying two conditions:

$$1. \quad f(x) \geq 0 \quad .$$

$$2. \quad \int_{-\infty}^{+\infty} f(x)dx = 1$$

To interpret the probability density function f(x) we can imagine that the total unit mass of probability (= 1) has been spread continuously over the real line with non-negative density f(x) at the point x.

It is important to realize, that the pdf values are only the marginal probabilities for each value x of the continuous random variable X.
In order to obtain the proper probability, we have to integrate the pdf.
Therefore <u>for continuous random variable X</u> the probabilities for its values are calculated as follows:

$$P\{a < X < b\} = \int_a^b f(x)dx \quad \text{for any interval } (a,b)$$

$$P\{X = x\} = 0 \quad \text{for any single value } x.$$

A common characteristic for both discrete and continuous random variables is called the <u>distribution function</u> denoted by F(x),

$$F(x) = P\{X < x\} = \begin{cases} \sum_{x_i < x} p_i & \text{for discrete case} \\ \int_{-\infty}^{x} f(t)dt & \text{for continuous case with pdf } f(t) \end{cases}$$

For the continuous random variable we have:

$$\frac{d\,F(x)}{dx} = f(x)$$

<u>Examples:</u>

1. A random trial is defined as a random selection of one household out of the households population. We know, that 90% of households live in rural areas, while only 10% in urban areas.
   If the rural household is selected, the real number "1" is associated with this fact, while the number "0" is associated with selection of an urban household.
   We can define the random variable X, with its probability distribution defined by the following probability function:

$$P(X = 1) = 0.9$$
$$P(X = 0) = 0.1$$

   The distribution function of the random variable X is defined as:

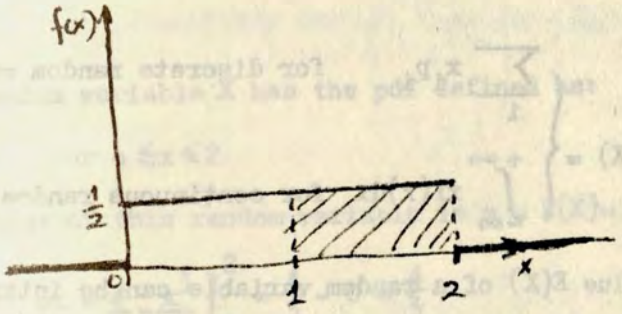
$$F(x) = \begin{cases} 0 & \text{for} \quad x \leq 0 \\ 0.1 & \text{for} \quad 0 < x \leq 1 \\ 1 & \text{for} \quad x > 1 \end{cases}$$

2. The area under wheat in sampled agricultural holding can be considered as the random variable X of continuous type, with the probability density function f(x) defined as:

$$f(x) = \begin{cases} \frac{1}{2} & \text{for} \quad 0 \leq x \leq 2 \quad \text{(hectares)} \\ 0 & \text{elsewhere} \end{cases}$$

So, the probability of obtaining an agricultural holding with the area under wheat within interval $1 \leq x \leq 2$ (see the graph) can be calculated as:

$$P\left(1 \leq X \leq 2\right) = \int_1^2 f(x)dx = \int_1^2 \tfrac{1}{2}dx = \tfrac{1}{2}x\Big]_1^2 = 1 - \tfrac{1}{2} = \tfrac{1}{2} = 0.5$$



Multidimensional random variables.

Let $(\Omega, \mathcal{F}, P)$ be a given probability space for some random experiment and let R^n be the n-dimensional real space containing the vectors

$$\mathbf{x} = (x_1, x_2, \ldots, x_n).$$

Any real-vector function $\mathbf{X} : \Omega \to R^n$ is called a n-dimensional random variable or random vector $\mathbf{X} = (X_1, X_2, \ldots, X_n)$.

The joint probability density function for a continuous random vector $\mathbf{X}$ will be denoted by $f(\mathbf{x}) = f(x_1, x_2, \ldots, x_n)$ and it satisfies two conditions:

1. $f(\mathbf{x}) \geq 0$
2. $\int \ldots \int_{R^n} f(\mathbf{x})d\mathbf{x} = 1$

Two-dimensional random variable (random vector) will be denoted usually by (X,Y) and its joint pdf by $f(x,y)$. The elements of a random vector X can be independent or dependent random variables. They can be also correlated or uncorrelated, which will be explained in the next paragraph.

For example: function $f(x,y)$ defined as

$$f(x,y) = x+y \quad \text{for } 0 \leq x \leq 1, \quad 0 \leq y \leq 1$$

is the pdf of some two-dimensional random variable (X,Y).

4.3. Parameters of Distribution of the Random Variable.

The probability distribution of each random variable can be represented by the set of parameters called the moments.

The main "Location" or "the central tendency" parameter of a random variable is its expected value (or mean value).

The expected value of a random variable X is defined as

$$m = E(X) = \begin{cases} \sum_i x_i p_i & \text{for discrete random variable} \\ \int_{-\infty}^{+\infty} x f(x)\,dx & \text{for continuous random variable} \end{cases}$$

The expected value $E(X)$ of a random variable can be interpreted as the centre of gravity of its probability distribution.

The main properties of the expected value are as follows:

1. $E(c) = c$ for any constant c
2. $E(cX) = cE(X)$
3. $E(X \mp Y) = E(X) \mp E(Y)$
4. $E(X \cdot Y) = E(X) \cdot E(Y)$ for independent random variables X and

The expected value is the special case of moments of the random variable.

The r-th moment of a random variable X is defined as:

$$m_r = E(X^r).$$

The first moment is just the expected value $m = E(X)$.

The second moment $m_2 = E(X^2) = \int_{-\infty}^{+\infty} x^2 f(x)\,dx$ (for continuous case).

The main "scale" or "dispersion" parameter of a random variable is its variance denoted by $\sigma^2 = D^2(X) = \mathrm{Var}(X)$ and defined as $D^2(X) = E\left[X-m\right]^2$ where m is the expected value of the random variable X.

Usually the variance of a random variable is calculated from the formula:

$$\sigma^2 = D^2(X) = E(X^2) - m^2 = m_2 - m^2$$

Very often, the standard deviation of a random variable is used as a dispersion parameter. It is defined as $\sigma = D(X) = \sqrt{D^2(X)}$

The main properties of the variance $D^2(X)$ are:

1. $D^2(c) = 0$ for any constant c.
2. $D^2(cX) = c^2 D^2(X)$.
3. $D^2(X \mp Y) = D^2(X) + D^2(Y)$. for independent random variables X and Y.

Examples:

1. A continuous random variable X has the pdf defined as:

$$f(x) = \tfrac{1}{2}x \quad \text{for } 0 \leq x \leq 2$$

The expected value of this random variable is $m = E(X) = \int_0^2 x \cdot \frac{1}{2} x \, dx =$

$$= \tfrac{1}{2} \int_0^2 x^2 \, dx = \frac{1}{2} \cdot \frac{x^3}{3} \Big]_0^2 = \frac{4}{3} - 0 = \frac{4}{3}.$$

The second moment $m_2 = E(X^2) = \int_0^2 x^2 \cdot \frac{1}{2} x \, dx = \frac{1}{2} \int_0^2 x^3 dx = \frac{1}{2} \cdot \frac{x^4}{4} \Big]_0^2 = 2.$

Then the variance of the random variable X is equal to

$$\sigma^2 = D^2(X) = E(X^2) - m^2 = 2 - \frac{16}{9} = \frac{2}{9},$$

and the standard deviation $\sigma = D(X) = \sqrt{\frac{2}{9}} \approx 0.47$

2. A continuous random variable X has the pdf defined as:

$$f(x) = e^{-x} \quad \text{for } x \geq 0$$ (see the gamma distribution in the next paragraph).

The expected value of this random variable is:

$$m = E(X) = \int_0^{+\infty} x e^{-x} dx = \Gamma(2) = 1! = 1$$

The second moment $m_2 = E(X^2) = \int_0^{+\infty} x^2 e^{-x} dx = \Gamma(3) = 2! = 2.$

Then the variance of the random variable X is:

$$\sigma^2 = D^2(X) = m_2 - m^2 = 2 - 1 = 1,$$

and the standard deviation $\sigma = D(X) = \sqrt{1} = 1.$

For a multidimensional random variable (a random vector $\mathbb{X}$) the mean vector and the covariance matrix are the main parameters.

The mean vector (or the vector of expectations) of the random vector

$$\mathbb{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}$$

is defined as:

$$M = E(\mathbb{X}) = \begin{pmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_n) \end{pmatrix}$$

The covariance matrix of the random vector $\mathbb{X}$ is defined as:

$$V(\mathbb{X}) = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1n} \\ \sigma_{12} & \sigma_2^2 & \dots & \sigma_{2n} \\ \vdots & & \ddots & \vdots \\ \sigma_{1n} & \sigma_{2n} & \dots & \sigma_n^2 \end{pmatrix}$$

where $\sigma_i^2 = D^2(X_i) = \mathrm{Var}(X_i)$ for $i = 1, \dots, n$

$\sigma_{ij} = \mathrm{Cov}(X_i, X_j)$ are the covariances of the random variables X_i and X_j

$$\mathrm{Cov}(X_i, X_j) = E\left[(X_i - m_i)(X_j - m_j)\right], \text{ where } m_i = E(X_i),$$

$$m_j = E(X_j).$$

The covariance matrix $V(\mathbf{X})$ is symmetric, since $\sigma_{ij} = \sigma_{ji}$

The correlation coefficient of the random variables X_i and X_j is defined as:

$$\rho_{ij} = \frac{\mathrm{Cov}(X_i, X_j)}{D(X_i) \cdot D(X_j)} = \frac{\sigma_{ij}}{\sigma_i \cdot \sigma_j} = \frac{Cov(X_i, X_j)}{\sqrt{D^2(X_i) \cdot D^2(X_j)}}$$

If $\rho_{ij} = 0$, random variables X_i and X_j are called uncorrelated

If $\rho_{ij} \neq 0$, random variables X_i and X_j are called correlated

Another concept of relationship between two random variables, namely a stochastic dependency is based on joint and marginal distributions.

Let the random vector (X, Y) be a continuous two-dimensional random variable having a joint probability density function (joint pdf) denoted by $f(x,y)$.

The marginal pdf are defined as:

$$f_2(x) = \int_{-\infty}^{+\infty} f(x,y)\,dy \quad \text{for the random variable X,}$$

and

$$f_2(y) = \int_{-\infty}^{+\infty} f(x,y)\,dx \quad \text{for the random variable Y.}$$

The random variables X and Y are said to be (stochastically) independent if

$$f(x,y) = f_1(x) \cdot f_2(y)$$

If such equational condition for the joint and marginal pdf is not fulfilled, the random variables X and Y are called dependent.

A connection between the correlation and dependency is given by the following two very important theorems:

1. If random variables X and Y are independent; then they are also uncorrelated (i.e. $\rho_{xy} = 0$).
2. If random variables X and Y are correlated (i.e. $\rho_{xy} \neq 0$), then they are also stochastically dependent.

The regression functions(explained in the paragraph 4.5) which reflect the dependency between random variables, are directly used in econometric modelling.

Example.

A two-dimensional continuous random variable (X,Y) has the joint pdf defined as: $f(x,y) = xy$ for $0 \leq x \leq 1$, $0 \leq y \leq 2$.

The marginal pdf are:

$$f_1(x) = \int_0^2 xy\,dy = 2x, \text{ and } f_2(y) = \int_0^1 xy\,dx = \tfrac{1}{2}y$$

Because $f_1(x) \cdot f_2(y) = 2x \cdot \frac{1}{2} y = xy = f(x,y)$

the random variables X and Y are independent, and therefore also uncorrelated, i.e. $Cov(X,Y) = 0$ and $\rho_{xy} = 0$.

$$E(X) = \int_0^1 x \cdot 2x\,dx = \tfrac{2}{3}, \quad D^2(X) = \sigma_x^2 = \frac{1}{18}$$

$$E(Y) = \int_0^2 y \cdot \tfrac{1}{2} y\,dy = \frac{4}{3}, \quad D^2(Y) = \sigma_y^2 = \frac{2}{9}$$

So, the mean vector of the random variable (X,Y) has elements $\frac{2}{3}$ and $\frac{4}{3}$, while the covariance matrix is equal to $\begin{pmatrix} \frac{1}{18} & 0 \\ 0 & \frac{2}{9} \end{pmatrix}$

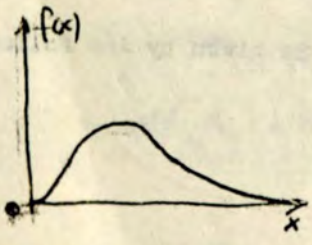
4.4. Some Important Probability Distributions.

The Gamma Distribution.

A continuous random variable X has the gamma distribution if its probability density function(pdf) has the form:

$$f(x) = \frac{b^p}{\Gamma(p)} x^{p-1} e^{-bx} \quad \text{for } x>0, \quad b>0, \quad p>0$$

where $\Gamma(p) = \int_0^{+\infty} x^{p-1} e^{-x}\,dx$ is the gamma function with properties:



$$\Gamma(p+1) = p\,\Gamma(p)$$

$$\Gamma(n+1) = n! \text{ for natural } n$$

$$\Gamma(\tfrac{1}{2}) = \sqrt{\pi}$$

The expected value in the gamma distribution is $E(X) = \frac{p}{b}$, and the variance is $D^2(X) = \frac{p}{b^2}$.

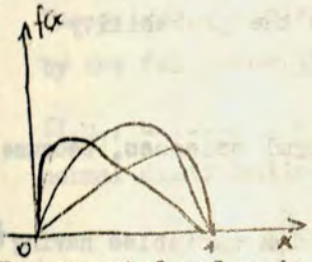
There are two important special cases of the gamma distribution:

(i) for $b = \alpha$, $p = 1$ we have so called exponential distribution with the pdf $f(x) = \alpha e^{-\alpha x}$ for $x > 0$, $\alpha > 0$

(ii) for $b = \frac{1}{2}$, $p = \frac{k}{2}$ we have so called Chi - square (χ^2) distribution with k degrees of freedom and the pdf $f(x) = \frac{1}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})} x^{\frac{k}{2}-1} e^{-\frac{x}{2}}$ for $x > 0$, k- natural number.

The Beta Distribution.

A continuous random variable X has the standard beta distribution if its pdf has the form:



$$f(x) = \frac{1}{B(p,q)} x^{p-1}(1-x)^{q-1} \quad \text{for } 0 \le x \le 1,\ p > 0,\ q > 0$$

$$B(p,q) = \int_0^1 x^{p-1}(1-x)^{q-1}dx = \frac{\Gamma(p)\cdot\Gamma(q)}{\Gamma(p+q)}$$ is the beta function

The expected value in the beta distribution is $E(X) = \frac{p}{p+q}$, while the variance is $D^2(X) = \frac{pq}{(p+q)^2(p+q+1)}$

The special case of the beta distribution is the uniform distribution which is the beta distribution for $p = q = 1$, and the pdf $f(x) = 1$ for $0 \le x \le 1$.

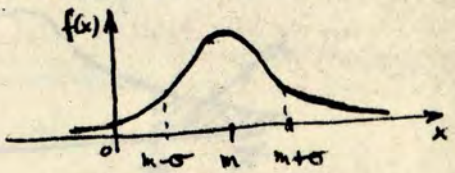
The beta distribution is connected with the gamma distribution by the following theorem: If X_1 and X_2 are two independent random variables having the gamma distribution with parameters b, p_1 and b, p_2 respectively, then the random variable $Z = \frac{X_1}{X_1+X_2}$ has the beta distribution with parameters $p=p_1$ and $q=p_2$.

The Normal Distribution.

A continuous random variable X has a normal distribution if its pdf has the form:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}} \quad \text{for } x(-\infty, +\infty),\ \sigma > 0.$$

The expected value(mean) in the normal distribution is $E(X) = m$, and the variance is $D^2(X) = \sigma^2$.
Very often the symbol $N(m, \sigma^2)$ is used for the normal distribution.



The special case is so called the standardized normal distribution $N(0,1)$ for $m=0$ and $\sigma^2 = 1$, with pdf: $f(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$ which can be easily obtained by standardization defined as

$$u = \frac{X - m}{\sigma}$$

The numerical values of the pdf and the distribution function $F(u)$ of the standardized normal distribution $N(0,1)$ are printed in Statistical Tables and they can be easily used to obtain any probability in the normal distribution.

It is worth to remember that in the normal distribution $N(m, \sigma^2)$ the probability

$$P\{m - 2\sigma < X < m + 2\sigma\} \approx 0.95 \quad \text{, while the probability}$$
$$P\{m - 3\sigma < X < m + 3\sigma\} \approx 0.997$$

The normal distribution is very often used in natural sciences, because of the Central Limit Theorem which says, that:

If $X_1, X_2, \ldots, X_n$ is a sequence of independent random variables having the same or different distributions with a finite means and variances, then the sum $S_n = \sum_{i=1}^{n} X_i$ has in limit (when $n \to \infty$) the normal distribution.

The Central Limit Theorem gives also the background of using the normal distribution for the random disturbance term in econometric models.

The Multivariate Normal Distribution.

A n-dimensional random vector $\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}$ has the multivariate (n-dimensional) normal distribution with the mean vector $\boldsymbol{\mu} = E(\mathbf{X})$ and the covariance matrix $V(\mathbf{X}) = \boldsymbol{\Sigma}$, if the joint probability density function has the form:

$$f(\mathbf{x}) = (2\pi)^{-\frac{n}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left\{-\tfrac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right\}$$

where $|\Sigma|$ is the determinant of the covariance matrix Σ,
Σ^{-1} is the inverse of the covariance matrix Σ

Often the symbol $N(\mu, \Sigma)$ is used for the multivariate normal distribution.

In econometric models it is often assumed, that the random disturbance vector ε has n-dimensional normal distribution $N(0, \sigma^2 I)$ i.e. with the mean vector $\mu = 0$ and the diagonal covariance matrix $\Sigma = \sigma^2 I$. For that case the joint pdf has the simple form:

$$f(\varepsilon) = (2\pi \cdot \sigma^2)^{-\frac{n}{2}} \cdot e^{-\frac{1}{2\sigma^2} \varepsilon^T \varepsilon}$$

The Chi-square (χ^2) Distribution.

The chi-square distribution is connected with the normal distribution by the following theorem:

If $u_1, u_2, \ldots, u_k$ are independent random variables with the standard normal distributions N(0, 1), then the random variable

$\chi^2 = \sum_{i=1}^{k} u_i^2$ has the chi-square distribution with k degrees of freedom (df = k).

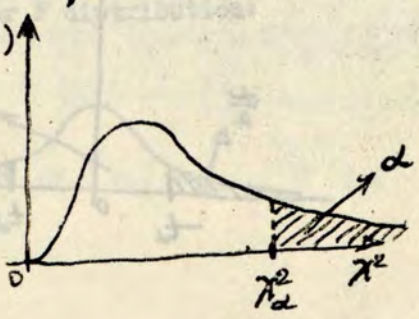
The pdf of the χ^2 distribution being a special case of the gamma distribution is in the form:

$$f(\chi^2) = \frac{1}{2^{\frac{k}{2}} \cdot \Gamma(\frac{k}{2})} (\chi^2)^{\frac{k}{2}-1} e^{-\frac{\chi^2}{2}} \text{ for } \chi^2 > 0, \text{ and k natural number.}$$

The only parameter in the chi-square distribution is k (the number of degrees of freedom), which is also the mean (expected value) : $E(\chi^2) = k$.
The variance of the χ^2 distribution is $D^2(\chi^2) = 2k$.
For $\chi^2 = k-2 > 0$ exists the maximum value of the pdf.

The statistical tables usually contain a table of the χ^2 distribution in such a form, that for given probability α and for given number of degrees of freedom k= 1,2,..., such a value χ^2 is printed, that

$$P\{\chi^2 \geq \chi^2_\alpha\} = \alpha,$$



There are two important theorems for the chi-square distribution:

1. If $\mathbf{U}$ is a random vector having n-dimensional standard normal distribution $N(\mathbf{0}, \mathbf{I})$, then the quadratic form $\mathbf{u}^T\mathbf{A}\mathbf{u}$ has the χ^2 distribution with the number of degrees of freedom equal to the rank of matrix $\mathbf{A}$ (df= $r(\mathbf{A})$).

2. If df = $k \to \infty$, then χ^2 distribution tends to the normal distribution

$$\text{i.e. } \sqrt{2\chi^2} \xrightarrow[k\to\infty]{} N\left(\sqrt{2k-1},\ 1\right).$$

The t-Student's Distribution.

The t-Student's distribution is connected with the normal and with the chi-square distributions by the following theorem:

If U and V are two independent random variables, and U has the normal distribution N(0,1), while the random variable V has the χ^2 distribution with df = k, then the random variable t defined as

$$t = \frac{U}{\sqrt{\frac{V}{k}}}$$ has the Student's t distribution with df=k

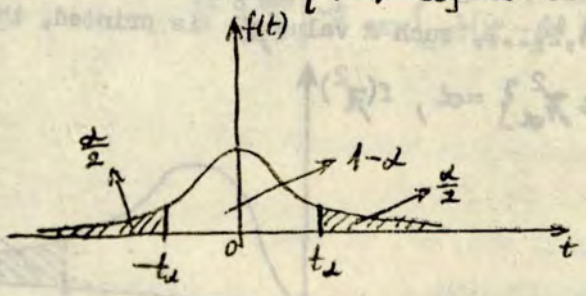
The probability density function of the t Student's distribution has the form:

$$f(t) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{k\pi}\,\Gamma\left(\frac{k}{2}\right)} \left(1 + \frac{t^2}{k}\right)^{-\frac{k+1}{2}} \quad \text{for } t \in (-\infty, +\infty),$$ and k-natural number.

The expected value in the t-Student's distribution is $E(t) = 0$, while the variance $D^2(t) = \frac{k}{k-2}$ for $k > 2$

The Student's t distribution tends to the standard normal distribution N(0,1) as df= $k \to \infty$

The statistical tables usually contain a table of the Student's distribution in such a form, that for given probability α and for given df=k=1, 2..., such a value t is printed, that $P\{|t| \geq t_\alpha\} = \alpha$



The F-Fisher-Snedecor Distribution

The F distribution is connected with the chi-square distribution by the following theorem:

If V_1 and V_2 are two independent random variables, both having χ^2 distribution with $df=k_1$ and $df=k_2$ respectively, then the random variable F defined as

$$F = \frac{V_1/k_1}{V_2/k_2} = \frac{V_1}{V_2} \cdot \frac{k_2}{k_1}$$ has the F-Fisher-Snedecor distribution with k_1 and k_2 degrees of freedom.

The pdf of the F distribution is in the form:

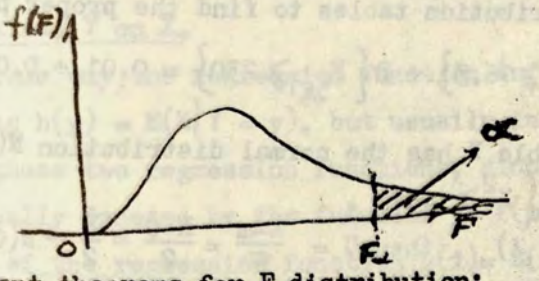
$$f(F) = \frac{\Gamma\left(\frac{k_1 + k_2}{2}\right)}{\Gamma\left(\frac{k_1}{2}\right)\cdot\Gamma\left(\frac{k_2}{2}\right)} \left(\frac{k_1}{k_2}\right)^{\frac{k_1}{2}} \cdot \frac{F^{\frac{k_1}{2}-1}}{\left(1 + \frac{k_1}{k_2}F\right)^{\frac{k_1+k_2}{2}}} \quad \text{for } F \geq 0$$

and k_1 , k_2 natural numbers.

The expected value in the F distribution is $E(F) = \frac{k_2}{k_2 - 2}$ for $k_2 > 2$

and the variance is $D^2(F) = \frac{2k_2^2(k_1 + k_2 - 2)}{k_1(k_2 - 4)(k_2-2)^2}$ for $k_2 > 4$

The Statistical Tables usually contain a table of F distribution in such a form, that for given probability α and for given pair k_1 and k_2 of degrees of freedom, such a value F_α is printed, that $P\{F \geq F_\alpha\} = \alpha$



There are three important theorems for F distribution:

1. If t is a random variable having the Student distribution with df=k then $F=t^2$ has the F Snedecor distribution with $k_1 = 1$ and $k_2 = k$ degrees of freedom.

2. If the random variable F has F Snedecor distribution with k_1 and k_2 degrees of freedom, than the random variable $Z = \frac{1}{F}$ has the F distribution with respectively k_2 and k_1 degrees of freedom.

3. If F is a random variable having F Snedecor distribution with k_1 and K_2 degrees of freedom, then the random variable defined as

$X = \frac{1}{1 + \frac{k_1}{k_2} F}$ has the beta distribution with parameters :

$$p = \frac{k_2}{2}, \quad q = \frac{k_1}{2}$$

Examples:

1. Find the expected value and the variance of the continuous random variable X having the pdf : $f(x) = 4\,x \cdot e^{-2x}$ for $x > 0$.
To solve the problem we can check, that X has infact the gamma distribution with parameters $p = b = 2$.
So, the mean value $E(X) = \frac{p}{b} = 1$, and the variance $D^2(X) = \frac{p}{b} = \frac{1}{2}$.

2. A random variable X has the t Student distribution with df= 5. Find the probability: $P\{X^2 \geq 16.3\} + P\{\frac{1}{X^2} \geq 230\}$.
Recalling that X^2 has the F distribution with df $k_1 = 1$ and $k_2 = 5$, and $\frac{1}{X^2}$ has the F distribution with df $k_1 = 5$ and $k_2 = 1$ we shall use the F distribution tables to find the proper probabilities.

$$P\{F_{1\times5} \geq 16.3\} + P\{F_{5\times1} \geq 230\} = 0.01 + 0.05 = 0.06$$

3. A random variable X has the normal distribution N(0,4). Find the probability $P\{X^2 > 1.82\}$.
Since $X \sim N(0, 4)$, then $U = \frac{X-m}{\sigma} = \frac{X-0}{2} = \frac{X}{2} \sim N(0, 1)$,
and then $U^2 = \frac{X^2}{4}$ has the χ^2 distribution with df = 1.
Therefore we have $P\{X^2 > 1.82\} = P\{\frac{X^2}{4} > \frac{1.82}{4}\} =$

$= P\{\chi_1^2 > 0.455\} \approx 0.50$ (from the table of χ^2 distribution).

4.5. Regression Functions.

For stochastically dependent random variables the regression functions can be very useful for studying (and for prediction) of realizations of some random variable provided given values for other random variables. In definition of the regression function the conditional probability distribution is involved.

Let (X, Y) be a two-dimensional continuous random variable having the probability distribution described by the joint probability density function $f(x,y)$.

The one-dimensional marginal distribution of the random variable X is described by the marginal pdf $f_1(x) = \int_{-\infty}^{+\infty} f(x,y)dy$.

The one-dimensional conditional distribution of the random variable Y can be defined now by means of the conditional pdf $f(y|x) = \frac{f(x,y)}{f_1(x)}$

Let us consider now the main parameter in that conditional distribution, i.e. the conditional expected value of the random variable Y:

$$E(Y|X = x) = \int_{-\infty}^{+\infty} y \cdot f(y|x)dy$$

For any given x(value of the random variable X) the conditional expected value $E(Y|X=x)$ is a number, but if x varies within the range of the random variable X, we can define the function:

$$g(x) = E(Y|X = x)$$

Such a function $g(x)$ giving the conditional means of the random variable Y for each value of the second random variable X is called the regression function (of the I-st kind) Y on X.

Symmetrically, in the same way, the regression function (of the I-st kind) X on Y can be defined as $h(y) = E(X|Y = y)$, but usually in practice, we use only the first one $g(x)$ out of these two regression functions, since a causal relationship (if exists) is usually denoted by the function $y = f(x)$.

The important property of the regression function $g(x) = E(Y|X = x)$ is, that for this particular function, $E[Y - g(x)]^2$ reaches the minimum value compared to those for any other function $g(x)$.

Since the regression function Y on X, i.e. $g(x) = E(Y|X = x)$ transforms each realization of the random variable X into a conditional expected value of the random variable Y, it can be used as the tool for the prediction (forecasting) of y values, provided given values x.

There is however one practical limitation here: in order to calculate (by integration) the regression function of the I-st kind, we have to know fully the joint probability distribution i.e. the joint pdf $f(x,y)$.

Because usually the joint pdf $f(x,y)$ is not known, we have to estimate the regression function $g(x)$ (based on statistical sample from the two-dimensional distribution of (X,Y)).

The estimated (by the Ordinary Least Squares method) regression function is called the II-nd kind of regression and it is commonly used in a statistical practice, also in Econometrics.

Only sometimes, knowing the joint pdf $f(x,y)$ we are able to calculate (and not to estimate) directly the regression function of the I-st kind.

The regression function can be similarly defined for any k+1-dimensional random vector $(Y,\bar{X}) = (Y,X_1,X_2,\ldots,X_k)$ assuming that the joint probability distribution is known.

Such a multiple regression function is defined as

$$g(x_1,x_2,\ldots,x_k) = E(Y|X_1 = x_1, X_2 = x_2,\ldots,X_k = x_k)$$

The type of the regression function depends on the type of joint probability distribution for which the regression function is calculated. Very often the regression functions are non-linear. Only in the multivariate normal distribution the regression functions are linear.

Examples:

1. A two-dimensional continuous random variable (X,Y) has the joint pdf defined as $f(x,y) = \begin{cases} xy & \text{for } 0 \leq x \leq 1, \quad 0 \leq y \leq 2 \\ 0 & \text{elsewhere} \end{cases}$

   So, the marginal pdf for the random variable X can be calculated as

$$f_1(x) = \int_{-\infty}^{+\infty} f(x,y)dy = \int_0^2 xydy = x\int_0^2 ydy = x\cdot\frac{y^2}{2}\Big]_0^2 = 2x \text{ for } 0 \leq x \leq 1.$$

Now, the conditional pdf for the random variable Y can be obtained as $f(y|x) = \frac{f(x,y)}{f_1(x)} = \frac{xy}{2x} = \frac{1}{2}y$ for $0 \leq y \leq 2$

The regression function(of the I-st kind) Y on X can be calculated now as

$$g(x) = E\left[Y \mid X = x\right] = \int_{-\infty}^{+\infty} y \cdot f(y|x)dy = \int_0^2 y \cdot \frac{1}{2}y dy = \frac{1}{2}\int_0^2 y^2 dy =$$

$$= \frac{1}{2} \cdot \frac{y^3}{3}\Big]_0^2 = \frac{4}{3}.$$

So, the regression function Y on X is a constant function $g(x) = \frac{4}{3}$ (it does not depent on x), which is clear if we realize, that X and Y are stochastically independent variables,

as $f_2(y) = \int_0^1 xydx = y\int_0^1 xdx = y \cdot \frac{x^2}{2}\Big]_0^1 = \frac{1}{2}y$ for $0 \leq y \leq 2$,

and $f_1(x) \cdot f_2(y) = 2x \cdot \frac{1}{2}y = xy = f(x,y)$.

Because of that independency of X and Y, the constant regression function $g(x) = \frac{4}{3}$ is not so useful for prediction of values y, as in other cases, when X and Y are dependent random variables.

2. A two-dimensional continuous random variable (X,Y) has the joint pdf defined as: $f(x,y) = x+y$ for $0 \leq x \leq 1$, $0 \leq y \leq 1$.
So, the marginal pdf for X can be calculated as:

$$f_1(x) = \int_0^1 (x+y)dy = x\int_0^1 dy + \int_0^1 ydy = x + \frac{1}{2} \quad \text{for } 0 \leq x \leq 1.$$

Because of symmetry, the marginal pdf for Y is $f_2(y) = y + \frac{1}{2}$ for $0 \leq y \leq 1$.

Now, we may conclude, that the random variables X and Y are dependent, because $f_1(x) \cdot f_2(y) = (x + \frac{1}{2})(y + \frac{1}{2}) \neq f(x,y)$

One can check, that X and Y are also correlated($\rho_{xy} = -\frac{1}{11} \neq 0$).

The conditional pdf for Y is as follows:

$$f(y|x) = \frac{f(x,y)}{f_1(x)} = \frac{x+y}{x+\frac{1}{2}} \quad \text{for } 0 \leq y \leq 1$$

The regression function Y on X can be calculated now as

$$g(x) = E(Y|X=x) = \int_0^1 y\,\frac{x+y}{x+\frac{1}{2}}\,dy = \frac{1}{x+\frac{1}{2}}\int_0^1 (xy + y^2)dy =$$

$$= \frac{1}{x+\frac{1}{2}}\left(x\cdot\frac{y^2}{2}\Big]_0^1 + \frac{y^3}{3}\Big]_0^1\right) = \frac{\frac{1}{2}x + \frac{1}{3}}{x+\frac{1}{2}} = \frac{3x+2}{6x+3} \quad \text{for } 0 \le x \le 1.$$

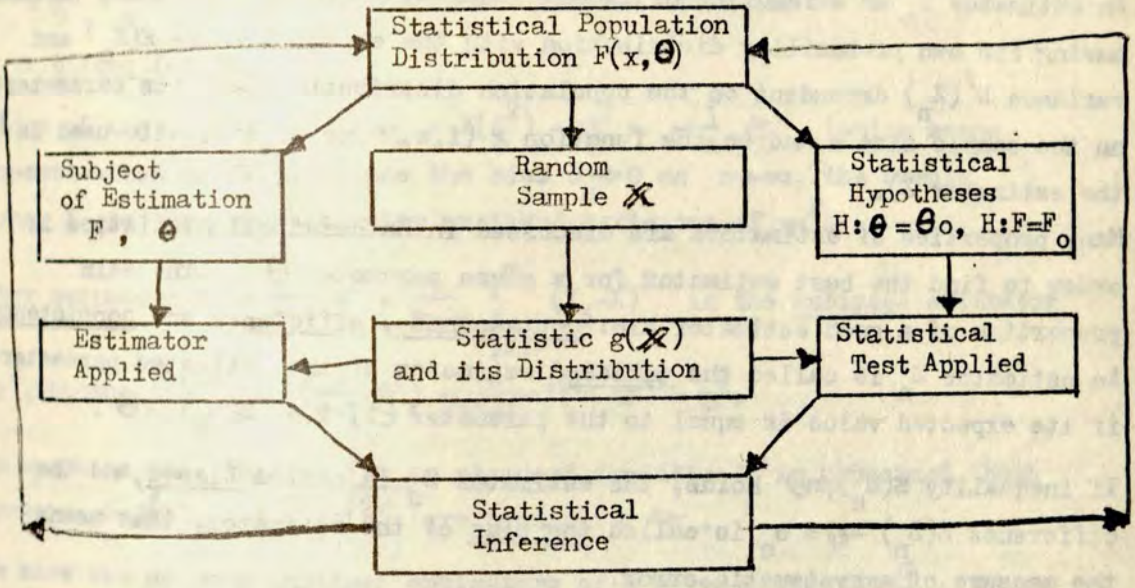
So, the regression function $g(x) = \frac{3x+2}{6x+3}$ is non-linear (hyperbolic). The linear function $h(x) = -\frac{1}{9}x + \frac{2}{3}$ can be taken as a good approximation of the $g(x)$. For example for $x = \frac{1}{2}$ we can predict the value y, based on the true regression function as $g(\frac{1}{2}) = \frac{7}{12} \approx 0.58$, while the linear approximation by the function $h(x)$, gives us $h(\frac{1}{2}) = \frac{11}{18} \approx 0.61$

CHAPTER 5. MATHEMATICAL STATISTICS BACKGROUND OF ECONOMETRICS.

5.1. Outline of the Statistical Estimation Theory.

A scientific concept of making inference about statistical populations based on random samples is the main subject of Mathematical Statistics. Mathematical Statistics deals with a rules, tools and principles of a statistical inference based on random samples and the probability theory. There are two basic kinds of the statistical inference: the estimation of distributions and the statistical hypotheses testing. The basic tools are estimators and statistical tests respectively and the statistic is the common idea used in both tools.

A general scheme of the statistical inference procedure is as follows:



It is assumed that for a given Statistical Population in which the variable X is observed its probability distribution described by the distribution function $F(x,\theta)$ is under investigation.

The type of distribution F or its parameter θ is the subject of an estimation or a statistical hypothesis. A sample $\mathcal{X}=(x_1, x_2, \ldots, x_n)$ is randomly selected from the population and its function $g(\mathcal{X})$ called the statistic have to be

properly chosen for the appropriate estimator or test application. Finally some new information about the statistical population come out, and the statistical inference can be repeated if necessary for an another aspect of the population distribution.

In any econometric modelling both types of a statistical inference (the estimation and the hypotheses testing) are in use.

The estimation procedures applied in an econometric modelling mainly concern the parametric estimation i.e. estimation of unknown parameters θ.

The basic tool in the parametric estimation is called the estimator and it is defined as follows:

Any statistic $Z_n = g(X)$ having the probability distribution which depends on the estimated parameter θ is called an estimator of that parameter θ.

A numerical value (realization) of the Estimator for a given sample is called the estimate of θ.

An estimator Z_n as a function of random sample X is in fact a random variable having its own probability distribution with the expected value $E(Z_n)$ and variance $D^2(Z_n)$ dependent on the population distribution (and its parameter θ), on the sample size n, and on the function g (i.e. type of statistic used as the estimator Z).

Many properties of estimators are discussed in Mathematical Statistics in order to find the best estimator for a given parameter θ. The main properties of a good estimator are: unbiasedness, efficiency and consistency.

An estimator Z_n is called the unbiased estimator of the estimated parameter θ, if its expected value is equal to the parameter θ, i.e. $E(Z_n) = \theta$.

If inequality $E(Z_n) \neq \theta$ holds, the estimator Z_n is called biased, and the difference $E(Z_n) - \theta = b_n$ is called the bias of the estimator, that means the measure of a systematic error.

If an estimator is biased, but its bias tends to zero $b_n \to 0$ as $n \to \infty$ the estimator is called asymptotically unbiased.

Examples:

1. A statistical population has the distribution with unknown mean value $E(X) = m$. As an estimator of m, the sample mean $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ is usually taken.

It is easy to check, that $\bar{X}$ is unbiased for m:

$$E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} m = \frac{nm}{n} = m.$$

2. A statistical population has the normal distribution $N(m, \sigma^2)$ with unknown variance σ^2. As an estimator of σ^2 the sample variance S^2 is taken:

$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$, where $\bar{X}$ is the sample average.

To study if S^2 is an unbiased estimator for the population variance σ^2, we can utilize chi-square distribution. It can be proved, that $\frac{n S^2}{\sigma^2} = \chi^2$ with df = n-1. Recalling that $E(\chi^2_k) = k$, we have:

$$E\left(\frac{n S^2}{\sigma^2}\right) = E(\chi^2_{n-1}) = n-1.$$

So, $\frac{n}{\sigma^2} E(S^2) = n-1$, and finally: $E(S^2) = \frac{n-1}{n}\sigma^2 =$

$= \sigma^2 - \frac{1}{n}\sigma^2 \neq \sigma^2$, which means that the estimator S^2 is <u>biased</u> for σ^2.

The bias of that estimator is $b_n = E(S^2) - \sigma^2 = -\frac{1}{n}\sigma^2$ (which means under-estimation of σ^2). Since the bias $b_n \to 0$ as $n \to \infty$, the sample variance S^2 is the <u>asymptotically</u> unbiased estimator of σ^2.

Another estimator $\hat{S}^2 = \frac{n}{n-1} S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is the <u>unbiased</u> estimator of σ^2, because $E(\hat{S}^2) = E\left(\frac{n}{n-1} S^2\right) = \frac{n}{n-1} \cdot \frac{n-1}{n}\sigma^2 = \sigma^2$.

It is worth to add, that as far as standard deviation σ is concerned, both estimators $S = \sqrt{S^2}$ and $\hat{S} = \sqrt{\hat{S}^2}$ are biased for σ.

If we have two or more unbiased estimators of a given parameter θ, we shall choose the estimator having the smallest variance. Such an estimator which is unbiased and has the <u>minimum variance</u> among the all unbiased estimators of the parameter θ is called <u>the most efficient estimator</u> Z_n^* for θ.
If an estimator Z_n is unbiased for θ, its standard deviation $D(Z_n)$ is called the <u>standard error</u> of the estimator Z_n, and it is denoted by $SE(Z_n)$

It follows, that to obtain the smallest standard error (SE) of an estimate for given parameter Θ we shall use the most efficient estimator (if it exists) of Θ.

To compare the efficiency of two unbiased estimators of the same parameter the relative efficiency is defined as:

$$e = \frac{D^2(Z_n^*)}{D^2(Z_n)}$$, where $D^2(Z_n^*)$ is the variance of the most efficient estimator and $D^2(Z_n)$ is the variance of any estimator.

$0 < e \leq 1$. $e = 1$ if the estimator Z_n is the most efficient,

Example:

It can be proved, that the sample average $\bar{X}$ is the most efficient estimator of the population mean value m in the normal distribution $N(m, \sigma^2)$ provided simple random sampling.

The variance of that estimator is given by $D^2(\bar{X}) = \frac{\sigma^2}{n}$ and its standard error $SE(\bar{X}) = D(\bar{X}) = \frac{\sigma}{\sqrt{n}}$.

Another unbiased estimator of population mean m in the normal distribution is sample median Me (middle value) having the variance $D^2(Me) = \frac{\pi}{2} \cdot \frac{\sigma^2}{n}$.

So, the relative efficiency of the sample median is

$$e = \frac{\frac{\sigma^2}{n}}{\frac{\pi}{2} \cdot \frac{\sigma^2}{n}} = \frac{2}{\pi} \approx 0.64 .$$

It means that the sample average $\bar{X}$ has the variance of 36% less than the variance of the sample median Me.

An estimator Z_n is said to be a consistent estimator of the parameter Θ if the sequence of estimates produced by the estimator is stochastically convergent to the parameter Θ when the sample size $n \to \infty$, i.e.

$$\lim_{n \to \infty} P\{|Z_n - \Theta| < \varepsilon\} = 1$$

or briefly: $\operatorname{p\,lim}_{n \to \infty} Z_n = \Theta$

In other words, there is full equivalence of the definition of consistency with the law of large numbers for the statistic Z_n.

From the practical viewpoint when using a consistent estimator we can approach the estimated parameter Θ by increasing the sample size. To check if some estimator is consistent or not we can utilize the following theorems:

1. If the estimator Z_n is an unbiased estimator of the parameter Θ (or it is asymptotically unbiased), and if its variance $D^2(Z_n) \to 0$, as $n \to \infty$, then the estimator Z_n is a consistent estimator of Θ.
2. (Slutsky theorem). If $\operatorname{plim}_{n\to\infty} Z_n = \Theta$, and $g(x)$ is a continuous or rational function, then $\operatorname{plim}_{n\to\infty} g(Z_n) = g(\Theta)$.

Examples:

1. Sample average $\bar{X}$ is an unbiased estimator of the population mean m, and its variance (provided simple random sampling) is $D^2(\bar{X}) = \frac{\sigma^2}{n} \to 0$ as $n \to \infty$. Then $\bar{X}$ is a consistent estimator of m. Next it follows, that $g(\bar{X}) = \bar{X}^2$ is consistent estimator of m^2, since the quadratic function g is continuous.
2. Sample variance $\hat{S}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$ is an unbiased estimator for the population variance σ^2, and its variance is $D^2(\hat{S}^2) = \frac{2\sigma^4}{n-1} \to 0$ as $n \to \infty$. Then $\hat{S}^2$ is consistent estimator of σ^2, and $\hat{S}$ is consistent estimator of the standard deviation σ.

There are two basic techniques of estimation of the unknown parameter Θ: the point estimation and the interval estimation.

The point estimation is simple: assuming that Z_n is a good estimator of (i.e. unbiased, efficient and consistent), its value (realization) z_n is calculated for the given sample $\mathcal{X}$. Such a numerical value (number, or point on real axis) is taken as an estimate of the unknown parameter Θ and is called the point estimate for Θ. The standard error(SE) of the point estimate is usually used as a measure of its precision.

The interval estimation gives a set of estimates for the unknown parameter Θ,

provided given high probability that this set contains the true value θ. Such set of estimates is called the confidence interval for the estimated parameter θ, and is denoted by :

$$P\left\{a(Z_n) < \theta < b(Z_n)\right\} = 1 - \alpha$$

where the lower limit $a(Z_n)$ and the upper limit $b(Z_n)$ of the confidence interval are random variables (as functions of an estimator Z_n), and $1 - \alpha$ is the given probability (near to 1) called the confidence coefficient.

An interval estimation through confidence intervals is much more safe technique than a point estimation, but it can be applied only when the type of probability distribution of the estimator Z_n is known. Otherwise the confidence interval(a,b) can not be obtained.

Example:

Sample average $\bar{X}$ is a good estimator for the population mean m in the normal distribution $N(m, \sigma^2)$. Its numerial value $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ calculated for the given sample results $(x_1, x_2, \ldots, x_n)$ is the point estimate for m. In addition it is well known, that $\bar{X}$ has the normal distribution $N(m, \frac{\sigma^2}{n})$. Provided that the population variance σ^2 is known, the confidence interval for the population mean m is given by the formula:

$$P\left\{\bar{X} - u_\alpha \frac{\sigma}{\sqrt{n}} < m < \bar{X} + u_\alpha \frac{\sigma}{\sqrt{n}}\right\} = 1 - \alpha$$

where u_α is the appropriate number taken from the tables of the standard normal distribution N(0,1) for the fixed confidence coefficient $1 - \alpha$, in such a way, that $P\left\{-u_\alpha < U < u_\alpha\right\} = 1 - \alpha$.

Mathematical Statistics(as well as Econometrics) has elaborated several general methods for getting estimators with good properties for a given estimated parameter θ : the Moments Method, the Maximum Likelihood Method, the Least Squares Method, etc. The last two methods (and their further extensions) are especially important and very often used in econometric modelling.

The Maximum Likelihood Method is based on the likelihood function defined as: $L(\mathcal{X}, \Theta) = f(x_1, \Theta) \cdot f(x_2, \Theta) \cdot \ldots \cdot f(x_n, \Theta)$

where $\mathcal{X} = (x_1, x_2, \ldots, x_n)$ are given sample results.

The maximum value of the $L(\mathcal{X}, \Theta)$ with respect to Θ can be obtained for some estimate $\hat{\Theta}$, which is the function g of sample results $\mathcal{X}$. Such an estimator $\hat{\Theta} = g(\mathcal{X})$ is called the Maximum Likelihood estimator (ML). It can be proved, that the ML estimators are asymptotically unbiased, consistent and asymptotically most efficient estimators for the estimated paramater Θ. The ML estimators have asymptotically(i.e. for large samples) normal distribution, that facilitates a confidence interval building.

The Least Squares Method is based on the minimization of the sum of squares of differences between the empirical and the estimated values of given function.

Let $y = g(\mathcal{X}, \Theta)$ be a given function for the random véctor $(Y, \mathcal{X})$ with Θ as unknown parameters vector. The sample results$(y_i, \mathcal{X}_i)$ for $i=1,2,\ldots,n$ are given joint observations on the random vector $(Y, \mathcal{X})$.

The Least Squares (LS) estimator of vector Θ is defined as such a vector $\hat{\Theta}$, which minimizes the criterion function S:

$$(\min) \quad S = \sum_{i=1}^{n} \left[y_i - g(\mathcal{X}_i, \Theta) \right]^2$$

The necessary condition for the minimum needs solution of the system of equations: $\frac{\partial S}{\partial \Theta} = 0$.

For non-linear functions "g" the solution can not be obtain sometimes analytically, therefore the usual way is to transform the function g into linear.

Examples:

1. A statistical population has the normal distribution $N(m, \sigma^2)$ with unknown mean value m, and known variance σ^2. We would like to find the Maximum Likelihood(ML) estimator of m. The likelihood function

for the normal distribution is defined as:

$$L = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - m)^2\right\}$$

Its logarithm is: $\ln L = -\frac{n}{2}\cdot\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - m)^2$

and the derivative $\frac{\partial \ln L}{\partial m} = \frac{1}{\sigma^2}\cdot\sum_{i=1}^{n}(x_i - m) = 0$

$$\text{for } \hat{m} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$$

Since $\frac{\partial^2 \ln L}{\partial m^2} = -\frac{n}{\sigma^2} < 0$, the likelihood function has the minimum for $\hat{m} = \bar{x}$ which is the ML estimator of m.

2. A statistical population (XY) has the two-dimensional distribution in which the regression function Y on X is linear $g(x) = \beta x + \alpha$. For given n sample observations (x_i, y_i) the unknown parameters β and α of the regression function can be estimated by the Least Squares (LS) method, i.e. by finding the minimum of the function $S = \sum_{i=1}^{n}(y_i - \beta x_i - \alpha)^2$. The well known solution of that problem is:

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

5.2. Outline of the Statistical Hypotheses Testing Theory.

A testing of the statistical hypotheses is the second kind of statistical inference based on random samples. It is widely used during a checking (verification) of estimated econometric models.

A statistical hypothesis is defined as any supposition about the probability distribution of a random variable, and it is usually denoted by $H: F(x) \in \Omega$, where $F(x)$ is the distribution function of a random variable X, and Ω is a given set of distribution functions, called the set of admissible hypotheses.

A statistical hypothesis is called a parametric hypothesis, if the set of admissible hypotheses contains such hypotheses which differ from themselves only in values of the parameter θ (within the same type of probability distribution), otherwise the hypothesis is called a non-parametric.

A statistical hypothesis is called a simple hypothesis, if the set of admissible hypotheses contains only one element, i.e. if it describes the probability distribution in the unique way. Otherwise the hypothesis is called a composite.

Examples:

1. $H: F(x) \in \Omega_N$, (where Ω_N is the set of all normal distribution functions) is a parametric and composite hypothesis (because of two parameters m and σ^2).
2. $H: \sigma^2 = 2$, (where σ^2 is the variance in a normal distribution with the known mean m) is a parametric simple hypothesis.

A statistical hypothesis can be tested by its confrontation with results of a sample drawn from the probability distribution, i.e. by checking if the sample result supports the hypothesis or denies it.

The proper tool for a verification of the statistical hypothesis is known in Mathematical Statistics as a statistical test.

The statistical test for a given statistical hypothesis H is defined as the decision rule in which one, out of the two decisions: to accept the hypothesis H, or to reject it, is connected with each random sample.

As a result of performing of a statistical test the final decision about the hypothesis H may be correct, or may be erroneous, as indicated by the following table:

Hypothesis / Decision	True	False
Accept	correct decision	II Kind error
Reject	I Kind error	correct decision

So, during the statistical hypothesis testing we risk that one out of the two errors will be made.

When we reject the hypothesis H, it might be a correct decision or the I Kind error is made (rejection the hypothesis H, while it is true).

When we accept the hypothesis H, it might be a correct decision, or the II kind error is made (acceptation the hypothesis H, while it is false).

A risk of making of those errors can be measured by the probabilities:

$$P\{\text{I kind error}\} = \alpha$$
$$P\{\text{II kind error}\} = \beta$$

For a fixed sample size it is not possible to construct such a statistical test, that both probabilities α and β are as small as possible.

Therefore some compromising solutions usually are obtained.

One type of solution is to construct so called the most powerful test. It is defined as such a test in which the probability α (of the I kind error) is fixed at the sufficiently low level, while the probability β (of the II kind error) is minimized.

The second type of solution (which is much more often used in statistical practice) is to construct the significance test.

A significance test is defined as such a test, which allows only to reject the hypothesis H (with fixed small probability α of the I kind error), or even such decision (rejection) can not be made, but never the decision "accept the hypothesis H" is made (since the probability β of the II kind error is not considered at all in that test).

The general principles for construction of a significance test for the given statistical hypothesis H_o are as follows:

1. For the statistical hypothesis H_o which is under testing we shall find an appropriate statistic Z_n such that its probability distribution is known when the hypothesis H_o is true.

2. For fixed small probability α (which is called also the significance level) such a critical set (region) Q of values of the statistic Z_n should be defined, that $P\{Z_n \in Q \mid H_o\} = \alpha$.
It means there is given small probability that the statistic Z_n will take its value within the set Q, provided that the hypothesis H_o is true.

3. Assuming given sample results $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ the realization z of the statistic Z_n is calculated for given sample and it is compared to the critical set Q.
If $z \in Q$, the hypothesis H_o is rejected with the probability of the I kind error equal to the small number α (significance level fixed in advance).
If $z \notin Q$, the hypothesis H_o can not be rejected (since the probability $P\{Z_n \notin Q \mid H_o\} = 1 - \alpha$ is rather high). It does not mean that we may to accept the H_o, since the probability β of the II kind error is not considered at all in that test.

So, in a significance test we may only reject the hypothesis H_o, or we may say, that we can not reject it (if $z \notin Q$).
Some useful statistical significance tests used in Econometrics for the verification of econometric models purposes will be described in the Part B of this handbook.

Examples:

1. A random variable X has the normal distribution $N(m, \sigma^2)$ with known variance σ^2. An inference has to be made about the mean value m, based on given n observations of the random variable X. We would like to test the hypothesis $H_o : m = m_o$, against the $H_1 : m \neq m_o$ (where m_o is a given number) by means of a significance test.
The statistic $U = \frac{\bar{X} - m_o}{\sigma}\sqrt{n}$ having the normal distribution $N(0,1)$, provided that H_o is true, can be used for construction of the significance test. The critical set for a given significance level α, can be defined as: $Q = \{u: |u| \geq u_\alpha\}$, where $P\{|u| \geq u_\alpha\} = \alpha$

So, if $u \in Q$, we have to reject the hypothesis H_o.

For example, let $n=60$, $\sigma = 20$, $\bar{x} = 72$ and $H_o : m = 55$, $H_1 : m \neq 55$.

For $\alpha = 0.01$ we have $u_\alpha = 2.576$ (from the normal distribution table).

Therefore $u = \frac{72-55}{20}\sqrt{60} = 6.75 \in Q = \{u : |u| \geq 2.576\}$, and the hypothesis $H_o : m = 55$ is rejected (at the significance level 0.01).

2. A statistical population has a normal distribution $N(m, \sigma^2)$ with unknown parameters. Based on the random sample of n elements we would like to test the hypothesis $H_o : \sigma^2 = \sigma_o^2$, against $H_1 : \sigma^2 > \sigma_o^2$ (where σ_o^2 is a given number), by using of a significance test.

The statistic $\chi^2 = \frac{n S^2}{\sigma_o^2}$ having the Chi-square distribution with $df = n - 1$, provided that H_o is true, can be used for construction of the significance test. The critical set can be defined as:

$Q = \{\chi^2 : \chi^2 \geq \chi_\alpha^2\}$, where $P\{\chi^2 \geq \chi_\alpha^2\} = \alpha$ (significance level)

So, if $\chi^2 \in Q$, the hypothesis H_o should be rejected.

Let's assume, that $n = 11$, $S^2 = 0.9$ (sample variance), and the hypothesis to be tested is $H_o : \sigma^2 = 0.6$

For $\alpha = 0.05$ (significance level), $\chi_\alpha^2 = 18.3$ (from the χ^2 distribution table for $df = 10$), and the critical set is defined as:

$Q = \{\chi^2 : \chi^2 \geq 18.3\}$.

Since $\chi^2 = \frac{11 \cdot 0.9}{0.6} = 16.5 \notin Q$, we can not reject the H_o.

PART B. ECONOMETRIC THEORY AND PRACTICE

CHAPTER 6. LINEAR MODEL WITH ONE EXPLANATORY VARIABLE.

6.1. Formulation of the Model.

In the strict sense Econometrics is the science which deals with methods for studying the statistical relationship between different economic variables. As we pointed out in Chapter 1, the economic variables have always a random (stochastic) character. Based on empirical statistical data one can investigate the economic variables and the existing relationships between them by means of some special statistical methods called an econometric modelling.
For example we may be interesting in the numerical relationship between the area under particular cereal crop and the crop yield in Ethiopian agricultural holdings. For this purpose a special econometric model can be formulated, estimated and analysed.
The basic tool of Econometrics i.e. the Econometric Model can be generally defined as a stochastic equation (or system of equations) describing the statistical relationship between the important variables observed in a given economic phenomenon.
It is clear that the idea of regression function (see paragraph 4.5) must be the direct background of any econometric model, since it deals with dependency between random variables.
The process of econometric modelling can be split into 3 main stages:

1. Formulation of the model,
2. Estimation of the model,
3. Testing the model.

Each separate stage needs an application of its own statistical techniques and has its own specific problems which have to be solved.
In the next few chapters of this handbook we shall concentrate on one-equational econometric models. The general structure of these models can be written in the form:

$$Y = f(\mathcal{X}, \Theta, \varepsilon)$$

where Y is the dependent (explained) variable,

$\mathcal{X}$ is the vector of independent (explanatory) variables,

Θ is the vector of unknown parameters,

ε is the random disturbance term,

f is the function of given type.

A formulation of the model contains such problems like:

(i) choice of the set of explanatory variables ($\mathcal{X}$) adequate for a given explained variable Y,

(ii) choice of the type of function "f" (linear, exponential, power, etc)

(iii) choice of the stochastic structure of the random disturbance ε .

Let us take now under detailed consideration the simplest econometric model i.e. one-equational linear model with one explanatory variable only. One well known special case of such a model is the linear trend function with the time variable as only explanatory variable. Another example of such a simplest model may be the linear demand function for some agricultural food commodity with the income per caput as a single independent variable.

Let Y stands for the given dependent economic variable, which is essential for the observed economic phenomenon. It is often called the explained variable, because we would like to explain the variation of its values by means of other (idependent) variable.

Let X stands for the independent economic variable which was chosen as an adequate explanatory variable which has to explain the variation of the dependent variable Y. Often the causal relationship, $X \Rightarrow Y$ is in mind while choosing the X, but not necessarly.

Let us assume the linear relationship between variables X and Y with some random disturbance denoted by ε .

So, the structure of our simplest model is: $Y = \beta X + \alpha + \varepsilon$, where the constants β and α are unknown structural parameters (to be estimated).

Let us assume, that we have numerical results of n joint observations (realizations) of the variables X and Y.

Let us denote those joint observations by (x_i, y_i) for i=1,2,..., n.

The detailed structure of the simplest linear model is completely described by the set of following assumptions:

Assumption 1. For each joint observation (x_i, y_i), i=1,2,..., n the equation $y_i = \beta x_i + \alpha + \varepsilon_i$ holds.

Assumption 2. Observations x_i are given non-random numbers, only y_i are random variables.

Assumption 3. ε_i for i = 1,2,...,n are independent random variables (and therefore also uncorrelated, cov $(\varepsilon_i, \varepsilon_j) = 0$) with the common mean value $E(\varepsilon_i) = 0$ and the variance $D^2(\varepsilon_i) = \sigma^2$.

From above assumptions it follows that:

(i) $E(y_i \mid x_i) = E(\beta x_i + \alpha + \varepsilon_i) = \beta x_i + \alpha$ (analogy to the regression function)

(ii) $D^2(y_i \mid x_i) = D^2(\varepsilon_i) = \sigma^2$

(iii) y_i are independent random variables and their average $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ has the mean value $E(\bar{y}) = \frac{1}{n}\sum_{i=1}^{n} E(y_i) = \frac{1}{n}\sum_{i=1}^{n} (\beta x_i + \alpha) =$

$$= \beta \bar{x} + \alpha$$

and the variance $D^2(\bar{y}) = \frac{1}{n}\sum_{i=1}^{n} D^2(y_i) = \frac{\sigma^2}{n}$.

It is worth to notice that in addition to the structural parameters β (the regression coefficient) and α (intercept), the unknown variance σ^2 of the random disturbance ε has to be estimated as well.
Although (i) gives a full analogy to the linear regression function of the I-st kind, nevertheless its parameters β and α are unknown. Therefore by estimation of these parameters we will be dealing in fact with the regression function of the II-nd kind.

6.2. Estimation of the Simplest Linear Model.

The unknown structural parameters β and α of the model can be estimated based on observations (x_i, y_i) by the Ordinary Least Squares (OLS) Method, i.e. by minimizing of the criterion function:

$$\min_{\beta,\alpha} S = \sum_{i=1}^{n} \left[y_i - E(y_i \mid x_i) \right]^2 = \sum_{i=1}^{n} \left[y_i - \beta x_i - \alpha \right]^2$$

The differentiation gives the following system of linear equations to be solved:

$$\begin{cases} \frac{\partial S}{\partial \beta} \equiv 2\sum_{i=1}^{n} (y_i - \beta x_i - \alpha)(-x_i) = 0 \\ \frac{\partial S}{\partial \alpha} \equiv 2\sum_{i=1}^{n} (y_i - \beta x_i - \alpha)(-1) = 0 \end{cases} \quad \text{or} \quad \begin{cases} \beta\sum_{i=1}^{n} x_i^2 + \alpha\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i \\ \beta\sum_{i=1}^{n} x_i + \alpha \cdot n = \sum_{i=1}^{n} y_i \end{cases}$$

Since a solution of that system of equations will depend on random variables y_i, then itself it will be a random variable. It means, that by solving of the above system of equation we will obtain the estimators denoted by "b" and "a" of unknown parameters β and α respectively. Using the Cramer's rule(see paragraph 3.2) we can obtain solution in the form:

$$b = \frac{\begin{vmatrix} \sum_{i=1}^{n} x_i y_i & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} y_i & n \end{vmatrix}}{\begin{vmatrix} \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & n \end{vmatrix}} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2} =$$

$$= \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$a = \bar{y} - b\bar{x} \quad , \qquad \text{where } \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \; , \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

It is worth noticing, that the OLS estimators b and a are linear (since they are linear functions of y_i).

For the variance of the estimator b we have:

$$D^2(b) = D^2\left[\frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right] =$$

$$= \frac{1}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^2} \sum_{i=1}^{n}\left[(x_i - \bar{x})^2 \cdot D^2(y_i)\right] = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Since b is an unbiased estimator and its variance $D^2(b) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \longrightarrow 0$, as $n \to \infty$, it follows that it is a consistent estimator of the parameter β, i.e. $\operatorname{plim}_{n\to\infty} b = \beta$

Similarly we can find that the variance of the estimator "a" is:

$$D^2(a) = \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2} \longrightarrow 0, \text{ as } n \to \infty$$

It follows, that the estimator "a" is a consistent estimator of α, i.e. $\operatorname{plim}_{n\to\infty} a = \alpha$.

One can also prove(as it is shown using matrix notation in Chapter 8), that the OLS estimators b and a are the best (i.e. the most efficient) estimators within the class of linear estimators.

Standard errors of these estimators are given by the formulas:

$$SE(b) = D(b) = \sigma_b = \frac{\sigma}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

$$SE(a) = D(a) = \sigma_a = \frac{\sigma\sqrt{\sum_{i=1}^{n} x_i^2}}{\sqrt{n\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

In fact, we can not calculate of those Standard Errors, since the variance σ^2 of the random disturbance ε is unknown and also has to be estimated by means of so called residual variance.

Residuals are defined as the differences:

$$\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - bx_i - a \quad \text{for } i=1, 2,\ldots,n.$$

As residuals can be treaten as the estimates for non-observable random disturbances ε_i, we can take the variance of residuals as an estimate of the unknown variance σ^2 of the random disturbances.
The residual variance is defined as:

$$S^2 = \frac{\sum_{i=1}^{n} \hat{\varepsilon}_i^2}{n-2} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}$$

One can also prove (see chapter 8 in matrix notation), that
S^2 is an unbiased estimator for σ^2.
Now, we can estimate the standard errors of the OLS estimates "b" and "a", as:

$$\text{Est}\left(SE(b)\right) = S_b = \frac{S}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}, \text{ where } S = \sqrt{S^2}$$

$$\text{Est}\left(SE(a)\right) = S_a = \frac{S\sqrt{\sum_{i=1}^{n} x_i^2}}{\sqrt{n\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Example:

Let us estimate the linear model $y_i = \beta x_i + \alpha + \varepsilon_i$ (with assumptions 1-3) based on the following empirical observations (x_i, y_i), i=1,2,3,4.

x_i	1	3	4	8
y_i	2	3	3	4

It will be convenient to carry on the calculations in the working table:

x_i	y_i	x_i^2	$x_i y_i$	$\hat{y}_i$	$\hat{\varepsilon}_i^2$	$(x_i - \bar{x})^2$
1	2	1	2	2.19	0.0361	9
3	3	9	9	2.73	0.0729	1
4	3	16	12	3.00	0	0
8	4	64	32	4.08	0.0064	16
16	12	90	55		0.1154	26

$\bar{x} = 4$, $\bar{y} = 3$, $n = 4$

So, we obtain the OLS estimates as:

$$b = \frac{55 - \frac{1}{4} \cdot 16 \cdot 12}{90 - \frac{1}{4} \cdot 16^2} = \frac{7}{26} \approx 0.27$$

$$a = 3 - 0.27 \cdot 4 \approx 1.92$$

$\hat{y}_i = 0.27x_i + 1.92$ for $i = 1,2,3,4$.

The residual variance $S^2 = \frac{0.1154}{2} = 0.0577$, and $S = \sqrt{S^2} \approx 0.24$

The standard errors of the OLS estimates are estimated as:

$$S_b = \frac{0.24}{\sqrt{26}} \approx 0.05$$

$$S_a = \frac{0.24\sqrt{90}}{\sqrt{4 \cdot 26}} \approx 0.22$$

6.3. Statistical Inference from the Model.

Often we need not only the point estimate but also a confidence interval for the main structural parameter β i.e. regression coefficient.

For that purpose (as well as for testing of the hypothesis about its significance) we have to add assumption about normality of disturbances distribution:

Assumption 4.

Random disturbances ε_i have an identical normal distribution $N(0,\sigma^2)$ with the pdf as $f(\varepsilon_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{\varepsilon_i^2}{2\sigma^2}}$.

Now, for our simplest linear model with assumptions 1-4 one can prove (see Chapter 9 in matrix notation) that the statistic defined as

$t = \frac{b - \beta}{S_b}$ has the t - Student's distribution with df = n-2.

Based on the above t statistic we can construct the confidence interval for the regression coefficient β as:

$$P\left\{b - t_\gamma \cdot S_b < \beta < b + t_\gamma \cdot S_b\right\} = 1 - \gamma$$

where t_γ is such a value from t - Student's distribution with df = n - 2, that for given probability $1 - \gamma$ (confidence coefficient) $P\left\{-t_\gamma < t < t_\gamma\right\} = 1-\gamma$

The significance test for the hypothesis $H_o : \beta = 0$ against $H_i : \beta \neq 0$ is based also on the t-Student's statistic.

If the hypothesis $H_o : \beta = 0$ is true, the statistic $t = \frac{b}{S_b}$

has the Student's distribution with df = n - 2.

So, the critical region is given by: $Q = \left\{t: |t| \geq t_\gamma\right\}, P\left\{|t| \geq t_\gamma\right\} = \gamma$

where t_γ is the critical value of the statistic t for the fixed significance level γ.

If the empirical value $t \in Q$, the hypothesis H_o is rejected, which means that the regression coefficient significantly differs from zero.

If $t \notin Q$, we can not reject the hypothesis H_o (the regression coefficient is statistically not significant).

Example:

Taking into account the numerical results obtained for the former example in the paragraph 6.2, we can construct the confidence interval for the

regression parameter β with the given confidence coefficient $1 - \gamma = 0.95$ (which gives $t_\gamma = 4.3$ for df = 2)

$$0.27 - 4.3 \cdot 0.05 < \beta < 0.27 + 4.3 \cdot 0.05$$
$$0.27 - 0.22 < \beta < 0.27 + 0.22$$
$$0.05 < \beta < 0.49$$

Alternatively, we can test the hypothesis $H_o: \beta = 0$ against $H_1: \beta \neq 0$. The empirical value of statistic t is equal to $t = \frac{0.27}{0.05} = 5.4$, while the critical value for the significance level $\gamma = 0.05$ and df = 2, is again $t = 4.3$, so the critical region is defined by

$$Q = \{t: |t| \geq 4.3\} .$$

Since the empirical value $t = 5.3 \in Q$, the hypothesis $H_o: \beta = 0$ should be rejected, which means that the estimated regression coefficient (b = 0.27) is statistically significant.

6.4. Forecasting Based on the Simplest Linear Model.

A forecasting can be defined as the estimation of the future realization of a random variable.

Let us assume, that joint observation (x_t, y_t) for t = 1,2, ..., n are the time series, and we are interested to predict (forecast) the future realization y_T of the random variable Y, based on given future value x_T of the explanatory variable in our simplest linear model.

For a forecasted value y_T we have the equation:

$$y_T = \beta x_T + \alpha + \varepsilon_T$$

with assumptions $E(\varepsilon_T) = 0$

$$D^2(\varepsilon_T) = \sigma^2$$

$$\mathrm{cov}(\varepsilon_T, \varepsilon_t) = 0 \quad \text{for all } t=1,\ldots,n.$$

The expected value of the forecasted value is now $E(y_T) = \beta x_T + \alpha$, and $y_T = E(y_T) + \varepsilon_T$.

So, the forecasting of the future single value y_T can be split into two problems: the prediction of the expected value $E(y_T)$ and the forecasting

of the future value of random disturbance ε_T.

Assuming given OLS estimates "b" and "a" for the parameters β and α, and given future value x_T, we can predict the expected value $E(y_T)$ as $\hat{y}_T = bx_T + a$.

The variance of that predictor can be obtained as:

$$D^2(\hat{y}_T) = D^2(bx_T + a) = D^2\left[bx_T + (\bar{y} - b\bar{x})\right] = D^2\left[\bar{y} + b(x_T - \bar{x})\right] =$$

$$= D^2(\bar{y}) + (x_T - \bar{x})^2 \cdot D^2(b) = \frac{\sigma^2}{n} + (x_T - \bar{x})^2 \cdot \frac{\sigma^2}{\sum_{i=1}^{n} (x_t - \bar{x})^2} =$$

$$= \sigma^2\left[\frac{1}{n} + \frac{(x_T - \bar{x})^2}{\sum_{i=1}^{n} (x_t - \bar{x})^2}\right].$$

Now, as the best forecast for the future value ε_T of the random disturbance is its expected value $E(\varepsilon_T) = 0$.

So, the final formula for the forecast of the future single value y_T is just equal to $\hat{y}_T = bx_T + a$, and its variance is given as:

$$D^2(\hat{y}_T) + \sigma^2 = \sigma^2\left[1 + \frac{1}{n} + \frac{(x_T - \bar{x})^2}{\sum_{t=1}^{n} (x_t - \bar{x})^2}\right]$$

The unknown variance σ^2 of the random disturbance can be estimated as

$$S^2 = \frac{\sum_{t=1}^{n} (y_t - \hat{y}_t)^2}{n-2}$$, which allows to estimate the standard error of our forecast as:

$$SE(\hat{y}_T) = S\sqrt{1 + \frac{1}{n} + \frac{(x_T - \bar{x})^2}{\sum_{t=1}^{n} (x_t - \bar{x})^2}}$$

Examples:

1. Taking into account the numerical estimation results obtained for the example given in the paragraph 6.2. we can calculate the forecast of a future value y_T, assuming the given linear model and the future value $x_T = 10$.
The estimated model was $\hat{y}_t = 0.27\, x_t + 1.92$, $s \approx 0.24$,
so for the future value $x_T = 10$, we obtain the forecast of y_T as:

$$\hat{y}_T = 0.27 \cdot 10 + 1.92 = 2.7 + 1.92 = 4.62$$

with the standard error

$$SE(\hat{y}_T) = 0.24\sqrt{1 + \tfrac{1}{4} + \frac{(10-4)^2}{26}} = 0.24\sqrt{1.25 + 1.38} \approx 0.39$$

2. The time series of the teff production y_t in the Ethiopian private peasants holdings for a given 7 years is as follows:(in thousand tons)

year	t	1	2	3	4	5	6	7
production	y_t	826	973	981	1001	1060	1286	1277

Based on the linear trend function model $y_t = \beta t + \alpha + \varepsilon_t$
we would like to calculate the forecast $\hat{y}_8$ of the next year teff production y_8.
For this purpose we shall adapt the formulas given in this paragraph with only change of the notation for explanatory variable : $x_t = t$ (time variable) , and $x_T = T$.
By means of the OLS method we obtain the estimate of the linear trend function as: $\hat{y}_t = 73.5t + 764$, $s \approx 61.4$.

So, for the comming year $T = 8$ we obtain the forecast of the teff production as $\hat{y}_8 = 73.5 \cdot 8 + 764 = 1352$ (of thousand tons)
with the standard error:

$$SE(\hat{y}_8) = 61.4\sqrt{1 + \frac{1}{7} + \frac{(8-4)^2}{28}} = 61.4\sqrt{1.71} \approx 80.3$$

The relative standard error of that forecast is equal to $\frac{80.3}{1352} = 6\%$.

If we are also interesting in getting the forecast of the teff production for two years forward, i.e. for $T = 9$, we can apply the same formulas for forecasting based on the estimated linear trend function model.

For $T = 9$ we obtain the forecast of the teff production as:

$$\hat{y}_9 = 73.5 \cdot 9 + 764 = 1425 \text{(of thousand ton)}$$

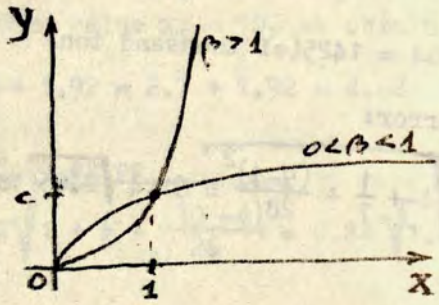
with the standard error:

$$SE(\hat{y}_9) = 61.4\sqrt{1 + \frac{1}{7} + \frac{(9-4)^2}{28}} = 61.4\sqrt{2.03} \approx 91.4 \ .$$

CHAPTER 7. NONLINEAR MODELS WITH ONE EXPLANATORY VARIABLE.

7.1. Power-type Model, its Reparametrization and Estimation.

One out of the non-linear models most often used in econometric practice is the power-type function model $y = c\,x^{\beta}$.



A direct application of the Ordinary Least Squares (OLS) method to that model is very complicated, due to the system of non-linear equations (which has not analytical solution) obtained after differentiation with respect to the parameters c and β.

Instead, the logarithmic transformation of the power-type model into the linear model is commonly in use, since it gives an easy way for estimation of the model. Such logarithmic transformation needs however a multiplicative disturbance error in the original power-type model, which should be written in the form:

$$y_t = c x_t^{\beta} \cdot \varepsilon_t \quad \text{for} \quad t = 1,2,\ldots, n \;, \quad c>0 \;, \; \varepsilon_t > 0.$$

Now, we can apply the logarithmic transformation, which gives:

$$\log y_t = \log c + \beta \log x_t + \log \varepsilon_t \quad , \quad \text{or}$$

$$y_t^* = \beta x_t^* + \alpha + v_t \quad , \quad \text{where} \begin{cases} y_t^* = \log y_t \\ x_t^* = \log x_t \\ \alpha = \log c \\ v_t = \log \varepsilon_t \end{cases}$$

Since the last model obtained after logarithmic transformation is linear, its parameters β and α can be easily estimated by the OLS method and the formulas given in the former

Chapter 6 (see paragraph 6.2) can be directly used with one small change only, that instead of original empirical data (x_t, y_t) their logarithms (x_t^*, y_t^*) should be used.

It is important to realize that only parameter β remains unchanged during that logarithmic transformation, while parameter c was transformed into the new parameter α and can be estimated indirectly from the estimate of α. That indirect estimation of c needs an unbiasedness in the estimation of α. For this purpose a special reparametrization of the obtained linear model should be introduced, based on the existing connection between the distributions of the additive disturbance error v_t in the linear model obtained after transformation, and the multiplicative disturbance error ε_t in the original power-type model.

For the obtained (after logarithmic transformation) linear model

$$y_t^* = \beta x_t^* + \alpha + v_t$$

we would like to assume that its disturbance error v_t has the normal distribution $N(0, \sigma^2)$. Because of the relationship $v_t = \log \varepsilon_t$ it follows, that the multiplicative disturbance error in the original power-type model will have the log-normal distribution with the mean value $E(\varepsilon_t) = e^{\frac{1}{2}\sigma^2}$, according to the following theorem:

If a random variable V has the normal distribution $N(m, \sigma^2)$, and $V = \log \varepsilon$, then the random variable ε has so called log-normal distribution, with mean (expected value) $E(\varepsilon) = e^{m+\frac{1}{2}\sigma^2}$.

The consequence of that fact is, that for our original model we obtain:

$$E(y_t) = cx_t^{\beta} \cdot E(\varepsilon_t) = cx_t^{\beta} \cdot e^{-\frac{1}{2}\sigma^2}$$

But we would like to have the power-type regression function rather in the form $E(y_t) = cx_t^{\beta}$, that means we would like to have $E(\varepsilon_t) = 1$, which is equivalent to:

$$e^{m + \frac{1}{2}\sigma^2} = 1$$

$$e^{m + \frac{1}{2}\sigma^2} = e^0$$

$$m + \tfrac{1}{2}\sigma^2 = 0$$

$$m = -\tfrac{1}{2}\sigma^2 = E(v_t)$$

The last result means, that in order to obtain $E(\varepsilon_t) = 1$, we have to assume, that v_t has the normal distribution $N(-\frac{1}{2}\sigma^2, \sigma^2)$ i.e. with mean value $E(v_t) = -\frac{1}{2}\sigma^2$ which is not convenient for the estimation of the linear model. Therefore a special reparametrization should be introduced in the linear model.

Let us denote $w_t = v_t - E(v_t) = v_t + \frac{1}{2}\sigma^2$, with $E(w_t) = 0$,

Now $v_t = w_t - \frac{1}{2}\sigma^2$, and we can rewrite our model in the form:

$$y_t^* = \beta x_t^* + \alpha + v_t$$

$$= \beta x_t^* + \alpha + w_t - \tfrac{1}{2}\sigma^2$$

$$= \beta x_t^* + (\alpha - \tfrac{1}{2}\sigma^2) + w_t$$

$$= \beta x_t^* + \alpha' + w_t \quad , \text{ where } \alpha' = \alpha - \tfrac{1}{2}\sigma^2$$

So, after such a reparametrization we have α' as constant term in our linear model, and a new disturbance error with the properties: $E(w_t) = 0$, $D^2(w_t) = \sigma^2$, which are good for estimation of the model by the OLS method.

Let $\hat{\beta}$ and $\hat{\alpha}'$ be the OLS estimates of β and α' respectively. $\hat{\beta}$ can be taken directly as good (unbiased) estimate of β. To obtain the estimate $\hat{\alpha}$ of α from the estimate $\hat{\alpha}'$ of α', we shall apply the formula:

$$\hat{\alpha} = \hat{\alpha}' + \tfrac{1}{2}S^2,$$

where S^2 is the residual variance as an estimate of unknown σ^2 (see Chapter 6).

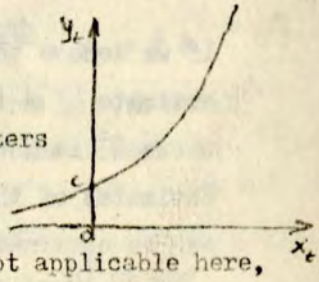
Of course, if the correction $\frac{1}{2}S^2$ is very small, it can be neglected, otherwise we shall remember to add $\frac{1}{2}S^2$ to the OLS estimate $\hat{\alpha}'$ of the constant term α in the linear model obtained from the original power-type model after logarithmic transformation.

The estimate of the original parameter c can be written as:

$$\hat{c} = \text{Anti-log}(\hat{\alpha}) = \text{Anti-log}(\hat{\alpha}' + \tfrac{1}{2}S^2).$$

In any case, the forecasting is based on the estimated linear model obtained after transformation, i.e. on $\hat{y}_t^* = \hat{\beta} x_t^* + \hat{\alpha}$.

7.2. Exponential Model, Logarithmic Model.



The exponential model can be defined as:

$y_t = c \cdot a^{x_t} \cdot \varepsilon_t$, where $c>0$, $a>0$ are parameters and $\varepsilon_t>0$ is the multiplicative disturbance error.

Like for the power-type model, a direct OLS method is not applicable here, and the logarithmic transformation of the exponential model is necessary for an indirect estimation of the unknown parameters c and a.
After logarithmic transformation of the original exponential model, we obtain again a linear model in the form:

$$\log y_t = \log c + x_t \cdot \log a + \log \varepsilon_t$$

or $y_t^* = \beta x_t + \alpha + v_t$, where $\begin{cases} y_t^* = \log y_t \\ \beta = \log a \\ \alpha = \log c \\ v_t = \log \varepsilon_t \end{cases}$

For such a linear model we can apply the OLS method for estimation of the unknown parameters β and α.
The formulas given in the paragraph 6.2 can be directly applied with one small change only, namely instead of empirical data (x_t, y_t), a partly transformed data (x_t, y_t^*) should be used.
Like for power-type model, the same reparametrization problem appears in the transformed exponential model, due to the condition $E(\varepsilon_t) = 1$ for the multiplicative disturbance error ε_t in the original model and its connection with the additive disturbance error v_t in the linear model.

Finally, the estimated linear model is in the form

$$y_t^* = \beta x_t + \alpha' + w_t \quad , \quad \text{where } \alpha' = \alpha - \tfrac{1}{2}\sigma^2$$

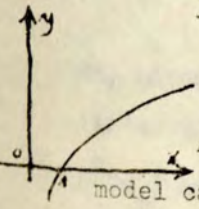
$$w_t = v_t + \tfrac{1}{2}\sigma^2 \quad , \text{ and } E(w_t) = 0,$$

$$D^2(w_t) = \sigma^2.$$

If we denote the OLS estimates of β and α' by $\hat{\beta}$ and $\hat{\alpha}'$ respectively, the estimate of α is given as $\hat{\alpha} = \hat{\alpha}' + \frac{1}{2} S^2$,
where S^2 is the residual variance.
Estimates of the original parameters "c" and "a" in the exponential model can be obtained as: $\hat{c} = \text{Anti-log}(\hat{\alpha}) = \text{Anti} - \log(\hat{\alpha}' + \frac{1}{2} S^2)$

$\hat{a} = \text{Anti-log}(\hat{\beta})$.

Ofcourse for forecasting purposes the estimated linear model $y_t^* = \beta x_t + \alpha$ is much more convenient instead of original exponential model.



The Logarithmic Model is defined as:

$y_t = \beta \log x_t + \alpha + \varepsilon_t$, where ε_t is the additive disturbance error

After changing of the variable $\log x_t = x_t^*$, the simple linear model can be obtained:

$$y_t = \beta x_t^* + \alpha + \varepsilon_t$$

and it can be directly estimated by the OLS method, based on the partly transformed empirical data (x_t^*, y_t).
There is no need for the reparametrization here, since the original (additive) disturbance error is not under logarithmic transformation. So, the formulas given in the paragraph 6.2 for the OLS estimates of the unknown parameters β and α can be used directly (with only change $x_t^* = \log x_t$ instead of x_t).

7.3. Logistic Trend Function Model.

In econometric modelling a special so called logistic trend function model is used sometimes. It is defined as:

$$y_t = \frac{k}{1 + be^{-at}} \quad \text{for } a > 0, \quad k > 0$$

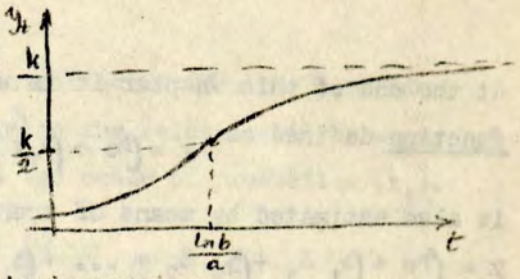
It is easy to check, that:

(i) $0 < y_t < k$, and $\lim_{t \to \infty} y_t = k$

k is called a level of satiation

(ii) $y'_t = \frac{dy_t}{dt} = \frac{a}{k} \cdot y_t(k - y_t) > 0$

which means, that the function y_t is increasing.



It has one point of flexion, with coordinates $t = \frac{\ln b}{a}$, $y_t = \frac{1}{2} k$

Generally speaking, the logistic trend function is a special growth model (Robertson's law) with the growth rate y'_t proportional to the actual level y_t and to the difference $k - y_t$.

Simultanuous estimation of all (3) unknown parameters (k, a, b) of the logistic trend function is very difficult. Some kind of two-stage estimation procedure known as Hotelling' method is usually used in that case which needs a special computer programme.
Often the level of satiation i.e. parameter "k" is known in advance, and only two parameters "a" and "b" are under estimation.
In that case, the logistic trend function can be transform into linear function for the estimation purposes as follows:

$$\frac{k}{y_t} = 1 + be^{-at}$$

$$\frac{k}{y_t} - 1 = be^{-at}$$

$$\log\left(\frac{k}{y_t} - 1\right) = \log b - at$$

$$y^*_t = -at + b^* \quad , \text{ where } \begin{cases} y^*_t = \log\left(\frac{k}{y_t} - 1\right) \\ b^* = \log b \end{cases}$$

The obtained model is linear and its parameters a and b^* can be estimated by the OLS method based on the empirical data y_t transformed into y^*_t for each t = 1,2,...n.

At the end of this Chapter it is worth mentioning, that the polynomial function defined as $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_k x^k$

is also estimated by means of transformation into linear function: $y = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \ldots + \beta_k Z_k$, where $Z_1 = x$, $Z_2 = x^2$, ..., $Z_k = x^k$, which can be considered as the linear multiple regression described in next Chapter 8.

It is worth also to add, that a commonly used criterion for choosing the model which has the best fitting to the empirical data is the coefficient of determination r^2 calculated for transformed(linear) model as:

$$r^2 = 1 - \frac{\sum_{t=1}^{n} (y_t - \hat{y}_t)^2}{\sum_{t=1}^{n} (y_t - \bar{y})^2}, \quad 0 \leq r^2 \leq 1.$$

A value of r^2 near to 1 means good fitting of the model to the empirical data.

The detailed explanation of r^2 is given in Chapter 9.

The list of non-linear models(with one explanatory variable) discussed in this Chapter and their appropriate transformations into linear models for the estimation purposes can be summarized by the following table:

Type of Model	Transformation of data	Remarks
Power-type	$(\log x_t , \log y_t)$	Reparametrization needed
Exponential	$(x_t , \log y_t)$	Reparametrization needed
Logarithmic	$(\log x_t , y_t)$	No needs for reparametrization
Polynomial	$(Z_1=x, Z_2=x^2, \ldots, Z_k=x^k)$	Multiple regression model
Logistic Trend Function	$\left(t, \log\left(\frac{k}{y_t} - 1 \right) \right)$	Parameter k is known

Examples:

1. The following empirical data have been collected from the sampled enterprises (manufacturing some article) on the relationship between the total cost of production (y_t) and the scale of production (x_t).

x_t	1	2	3	4	5	6	7
y_t	5	14	21	24	26	28	30

Based on the data, four types of cost function $y_t = f(x_t)$ were estimated: linear, power-type, exponential and logarithmic. Out of these functions, the estimated logarithmic model $\hat{y}_t = 12.86 \log x_t + 5.48$ has a best fitting to the data ($r^2 = 0.99$).

2. The time series of the teff production y_t in the Ethiopian private peasant holdings for given 7 years is as follows(in thousand tons):

year t	1	2	3	4	5	6	7
production y_t	826	973	981	1001	1060	1286	1277

In the example 2 (on page 74) we estimated the linear trend function model. Now, let us estimate the power-type trend function $y_t = c \cdot t^{\beta} \cdot \varepsilon_t$, or after logarithmic transformation $y_t^* = \beta \log t + \alpha + v_t$, where $y_t^* = \log y_t$, $\alpha = \log c$, $v_t = \log \varepsilon_t$.
The numerical results are:

$$\hat{y}_t^* = \underset{(0.041)}{0.211} \log t + \underset{(0.057)}{8.999}, \text{ with } r^2 = 0.84, \text{ and the residual variance}$$

$s^2 = 0.005$ which can be neglected as correction in the reparametrization.

The forecast $\hat{y}_8$ for the year T = 8 obtained based on the estimated power-type model is equal to $\hat{y}_8 = 1255$ (of thousand tons), which is less than forecast obtained based on the linear trend function model.

3. In some research experiment performed in Ethiopia for establishing the relationship between the crop yield Y and the Water-Balance

Index (X), four types of function $y = f(x, \varepsilon)$, namely: linear, power-type, exponential and logarithmic were estimated. Out of these functions estimated for maize in the Wollo region, the exponential function: $\log \hat{y}_t = 0.031\ x_t - 1.576$
(0.008)
had a best fitting to the empirical data ($r^2 = 0.725$).

CHAPTER 8. MULTIPLE LINEAR REGRESSION MODEL.

8.1. Estimation by the OLS Method.

Models with only one explanatory variable X are often not satisfactory for good explanation of variation of the given dependent variable Y. Econometric models with many independent variables are based on the multiple regression functions. Because they involve a very time-consuming computation for the estimation of a large number of parameters, multiple regression models were not often used before 1960. Nowadays, quick and powerful computers enable to estimate even a very big models with hundreds explanatory variables.

Let's start from basic assumptions.

We shall discuss the linear model for a given dependent(explained) variable Y with k independent (explanatory) variables $X_1, X_2, \ldots, X_k$.

On all these variables n joint observations have been collected as:

$$(y_t, x_{t1}, x_{t2}, \ldots, x_{tk}) \quad \text{for } t = 1,2,\ldots,n.$$

These empirical statistical data can be either time-series or cross-sectional data, so the index "t" relates either to the consecutive period of time (usually-year) or to the consecutive cross-sectional item (e.g. agricultural holding).

A detailed structure of the multiple regression model is completely described by the set of following assumptions:

Assumption 1. $y_t = \beta_1 x_{t1} + \beta_2 x_{t2} + \ldots + \beta_k x_{tk} + \beta_{k+1} + \varepsilon_t$ for $t=1,2,\ldots n$.

Assumption 2. Observations $x_{t1}, x_{t2}, \ldots, x_{tk}$ on explanatory variables are given (non-random) numbers for each t.

Assumption 3. Random disturbances ε_t are uncorrelated between themselves (i.e. $\operatorname{cov}(\varepsilon_t, \varepsilon_{t'}) = 0$) and they have common expected value $\bigwedge_t \; E(\varepsilon_t) = 0$

and common variance $\bigwedge_t \; D^2(\varepsilon_t) = \sigma^2$

From these assumptions if follows, that:

$$E(y_t \mid x_{t1}, x_{t2}, \ldots, x_{tk}) = \beta_1 x_{t1} + \beta_2 x_{t2} + \ldots + \beta_k x_{tk} + \beta_{k+1}$$

It means a linear multiple regression function with k + 1 unknown parameters.

A matrix notation is very useful in estimation of such a models.

Let's introduce the following symbols:

$$\mathcal{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$ is n x 1 column vector of all observations on dependent variable Y.

$$\mathcal{X} = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1k} & 1 \\ x_{21} & x_{22} & \ldots & x_{2k} & 1 \\ \vdots & \vdots & & \vdots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{nk} & 1 \end{pmatrix}$$ is n x(k+1) matrix of all observations on k explanatory variables and additional variable $X_{k+1} \equiv 1$

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{k+1} \end{pmatrix}$$ is (k+1) x 1 column vector of unknown (structural) parameters

$$\mathcal{E} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$ is nx1 column vector of random disturbances

Using these symbols we can rewrite briefly the assumptions of our model as:

Assumption 1. $\mathcal{Y} = \mathcal{X}\beta + \mathcal{E}$

Assumption 2. Matrix $\mathcal{X}$ has given(non-random) elements and it has full rank: $r(\mathcal{X}) = k + 1 \leq n$,
i.e. all its columns are linearly independent.

Assumption 3. The n-dimensional random (disturbance) vector $\mathcal{E}$ has the mean vector $E(\mathcal{E}) = \mathcal{O}$, and the covariance matrix $V(\mathcal{E}) = E(\mathcal{E} \cdot \mathcal{E}^T) = \sigma^2 I =$

$$= \begin{pmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{pmatrix}.$$

It follows, that $E(\mathcal{Y} | \mathcal{X}) = \mathcal{X}\beta$, or $\mathcal{Y} = E(\mathcal{Y}|\mathcal{X}) + \mathcal{E}$, or $\mathcal{E} = \mathcal{Y} - E(\mathcal{Y}|\mathcal{X})$

So, each element ε_t of the random disturbance vector $\mathcal{E}$ can be interpreted as the difference between the empirical value y_t of the explained variable and the corresponding value of the regression function. Minimization of the sum of squares of such differences is infact the basic criterion in the Ordinary Least Squares (OLS) Method for getting estimates of the unknown regression parameters (vector β):

$$(\min_{\beta}) \; S = \sum_{t=1}^{n} \varepsilon_t^2 = \mathcal{E}^T \mathcal{E} = (\mathcal{Y} - \mathcal{X}\beta)^T (\mathcal{Y} - \mathcal{X}\beta)$$

Let's split the function S into linear and quadratic form of the vector β for differentiation purposes (since minimization of S needs a derivation with respect to the vector β):

$$S = (\mathcal{Y} - \mathcal{X}\beta)^T (\mathcal{Y} - \mathcal{X}\beta) = \mathcal{Y}^T\mathcal{Y} - \beta^T\mathcal{X}^T\mathcal{Y} - \mathcal{Y}^T\mathcal{X}\beta + \beta^T\mathcal{X}^T\mathcal{X}\beta =$$

$$= \underbrace{\mathcal{Y}^T\mathcal{Y}}_{\text{constant}} - 2\underbrace{\mathcal{Y}^T\mathcal{X}\beta}_{\text{linear}} + \underbrace{\beta^T\mathcal{X}^T\mathcal{X}\beta}_{\text{quadratic}}$$

Recalling the differentiation of linear and quadratic forms in matrix notation (see paragraph 3.3): $\frac{\partial(a^T x)}{\partial x} = a$, $\frac{\partial(x^T A x)}{\partial x} = 2Ax$,

we have $\frac{\partial S}{\partial \beta} = -2\mathcal{X}^T\mathcal{Y} + 2\mathcal{X}^T\mathcal{X}\beta$

From the necessary condition for the minimum $\frac{\partial S}{\partial \beta} = \mathcal{O}$, we obtain the following system of linear equations:

$$\mathcal{X}^T\mathcal{X}\beta = \mathcal{X}^T\mathcal{Y}$$

Now multiplying (left-hand side) by the inverse matrix $(X^TX)^{-1}$ (which exists because the Gram matrix X^TX is not singular, since it has rank $r(X^TX) = r(X) = k + 1$ as was assumed in 2), and denoting the solution of that system of equation by b, we obtain finally the OLS estimator of β as: $b = (X^TX)^{-1}X^Ty$. Since the matrix of second derivatives $\frac{\partial^2 S}{\partial \beta^2} = 2X^TX$ is positive definite (see paragraph 3.3), the criterion function $S = \sum_{t=1}^{n} \varepsilon_t^2$ has really the minimum value for such an estimate of β which is equal to $b = (X^TX)^{-1}X^Ty$.

It is easy to realize, that the OLS estimator b of the parameters vector β is a <u>linear estimator</u>, since it is a linear function of the random vector y (and consequently also is a linear function of the random vector ε).

Of course, the formula for the OLS estimator b of the vector β is general, and it comprises as the special case those formulas given in paragraph 6.2 for the simplest linear model with one explanatory variable. This is clear, if the system of linear equations for a general case $X^TX\beta = X^Ty$ is written for that special case, i.e. for the simplest model $y_t = \beta x_t + \alpha + \varepsilon_t$ for which we have:

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix}$$

$$X^TX = \begin{pmatrix} x_1 & x_2 \cdots & x_n \\ 1 & 1 \cdots & 1 \end{pmatrix} \cdot \begin{pmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix} = \begin{pmatrix} \sum_{t=1}^{n} x_t^2 & \sum_{t=1}^{n} x_t \\ \sum_{t=1}^{n} x_t & n \end{pmatrix}$$

$$X^Ty = \begin{pmatrix} x_1 & x_2 & \cdots & x_n \\ 1 & 1 & \cdots & 1 \end{pmatrix} \cdot \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \sum_{t=1}^{n} x_t y_t \\ \sum_{t=1}^{n} y_t \end{pmatrix}$$

So, for that case the system of equations $X^T X \beta = X^T y$ can be written in the form:

$$\begin{pmatrix} \sum_{t=1}^{n} x_t^2 & \sum_{t=1}^{n} x_t \\ \sum_{t=1}^{n} x_t & n \end{pmatrix} \begin{pmatrix} \beta \\ \alpha \end{pmatrix} = \begin{pmatrix} \sum_{t=1}^{n} x_t y_t \\ \sum_{t=1}^{n} y_t \end{pmatrix}$$

or
$$\begin{cases} \beta \sum_{t=1}^{n} x_t^2 + \alpha \sum_{t=1}^{n} x_t = \sum_{t=1}^{n} x_t y_t \\ \beta \sum_{t=1}^{n} x_t + \alpha \cdot n = \sum_{t=1}^{n} y_t \end{cases}$$

as we obtained in paragraph 6.2.

8.2. Gauss-Markov Theorem.

The statistical properties of the OLS estimator b of the parameters vector β are considered by the essential Gauss-Markov Theorem.

Theorem.

For the multiple linear regression model with assumptions 1-3, the OLS estimator $b = (X^T X)^{-1} X^T y$ is the Best Linear Unbiased Estimator (BLUE) of the parameters vector β.

Let's consider now in detail each part of Gauss-Markov theorem.

First of all, the unbiasedness of the OLS estimator b can be checked.

Recalling, that by the Assumption 2, the matrix X is constant(non-random), the expected value of the linear estimator b is:

$$E(b) = E[(X^T X)^{-1} X^T y] = (X^T X)^{-1} X^T \cdot E(y) =$$
$$= (X^T X)^{-1} X^T \cdot X\beta = I \cdot \beta = \beta.$$

$E(b) = \beta$ means, that b is an unbiased linear estimator of the parameters vector β. Now let's find the covariance matrix for the unbiased estimator b.

For this purpose let us first consider the estimation error, i.e. the difference $\mathbb{b} - \beta = (X^TX)^{-1}X^Ty - \beta = (X^TX)^{-1}X^T(X\beta+\varepsilon) - \beta =$

$= (X^TX)^{-1}X^TX\beta + (X^TX)^{-1}X^T\varepsilon - \beta = I\beta + (X^TX)^{-1}X^T\varepsilon - \beta =$

$= (X^TX)^{-1}X^T\varepsilon.$

The <u>covariance matrix</u> of the OLS estimator $\mathbb{b}$ can be now obtained (recalling, that by Assumption 3, $V(\varepsilon) = E(\varepsilon \cdot \varepsilon^T) = \sigma^2 I$ and that $(A \cdot B)^T = B^T \cdot A^T$)

$$V(\mathbb{b}) = E\left[(\mathbb{b}-\beta)(\mathbb{b}-\beta)^T\right] = E\left[(X^TX)^{-1}X^T\varepsilon \cdot \varepsilon^T X (X^TX)^{-1}\right] =$$

$$= (X^TX)^{-1}X^T \cdot E(\varepsilon\varepsilon^T) \cdot X(X^TX)^{-1} = \sigma^2 (X^TX)^{-1} \cdot I =$$

$$= \sigma^2 (X^TX)^{-1}.$$

The principal diagonal of the covariance matrix $V(\mathbb{b})$ consists of the variances $D^2(b_i)$ of respective elements b_i $(i = 1,2,\ldots, k+1)$ of the vector $\mathbb{b}$.

So, $D^2(b_i) = \sigma^2_{bi} = \sigma^2 d_{ii}$, where d_{ii} is i-th diagonal element of the inverse matrix $(X^TX)^{-1}$.

For example, for the simplest linear model $y_t = \beta x_t + \alpha + \varepsilon_t$, the covariance matrix of the OLS estimator $\mathbb{b} = \begin{pmatrix} b \\ a \end{pmatrix}$ of the parameters vector $\beta = \begin{pmatrix} \beta \\ \alpha \end{pmatrix}$ is in the form:

$$V(\mathbb{b}) = V\begin{pmatrix} b \\ a \end{pmatrix} = \sigma^2 (X^TX)^{-1} = \sigma^2 \begin{pmatrix} \dfrac{1}{\sum_{t=1}^{n}(x_t - \bar{x})^2} & \dfrac{-\bar{x}}{\sum_{t=1}^{n}(x_t - \bar{x})^2} \\ \dfrac{-\bar{x}}{\sum_{t=1}^{n}(x_t - \bar{x})^2} & \dfrac{\sum_{t=1}^{n} x_t^2}{n\sum_{t=1}^{n}(x_t - \bar{x})^2} \end{pmatrix}$$

So, the variance of the estimator b of the regression coefficient β in the simplest linear model is equal to the first (upper) element of this covariance matrix, i.e. $D^2(b) = \dfrac{\sigma^2}{\sum_{t=1}^{n}(x_t - \bar{x})^2}$

Now we can prove that the OLS estimator b with its covariance matrix $V(b) = \sigma^2(X^TX)^{-1}$ is the best (i.e. most efficient) linear unbiased estimator of β.

The OLS linear estimator $b = (X^TX)^{-1}X^Ty$ can be written as $b = Ay$ where the matrix $A = (X^TX)^{-1}X^T$.

Let's consider for the vector β any other linear unbiased estimator $\hat{\beta} = Cy$ (linearity) with a constant matrix C, and $E(\hat{\beta}) = \beta$ (unbiasedness).

This estimator $\hat{\beta}$ can be written as $\hat{\beta} = C(X\beta + \varepsilon) = CX\beta + C\varepsilon$.

Since $E(\varepsilon) = 0$, from the unbiasedness of the estimator $\hat{\beta}$ it follows, that $CX = I$

Since the constant matrix C can be split as $C = A + D$, we have:

$$(A+D)X = I$$

$$AX + DX = I$$

$$I + DX = I, \quad \text{because } AX = (X^TX)^{-1}X^TX = I$$

or $DX = 0$ (null matrix)

Now, the covariance matrix of that linear unbiased estimator $\hat{\beta}$ is obtained as:

$$V(\hat{\beta}) = E[(\hat{\beta}-\beta)(\hat{\beta}-\beta)^T] =$$

$$= E[C\varepsilon \cdot \varepsilon^T C^T], \text{ since } \hat{\beta}-\beta = Cy-\beta = C(X\beta+\varepsilon)-\beta = C\varepsilon.$$

Next, recalling that $V(\varepsilon) = E(\varepsilon\varepsilon^T) = \sigma^2 I$, we obtain:

$$V(\hat{\beta}) = C \cdot E(\varepsilon\varepsilon^T) \cdot C^T = C\sigma^2 I C^T = \sigma^2 CC^T =$$

$$= \sigma^2 (A+D)(A+D)^T = \sigma^2 [AA^T + DA^T + AD^T + DD^T] =$$

$$= \sigma^2 [(X^TX)^{-1}X^T \cdot X(X^TX)^{-1} + DX(X^TX)^{-1} + (X^TX)^{-1}X^TD^T + DD^T] =$$

$$= \sigma^2 [(X^TX)^{-1} + DD^T] = \sigma^2 (X^TX)^{-1} + \sigma^2 DD^T =$$

$$= V(b) + \sigma^2 DD^T, \text{ where } DD^T \text{ is positive definite matrix.}$$

It follows, that for each element on the principal diagonal i.e. for respective variances we have inequality $D^2(\hat{\beta}_i) > D^2(b_i)$, which means that any other linear unbiased estimator of β_i has always a greater variance compared to the variance of the OLS estimator b_i.
In other words it means, that the OLS estimator $\mathbf{b}$ is the most efficient estimator within the class of linear unbiased estimators of $\boldsymbol{\beta}$. It completes the proof of the Gauss-Markov theorem, that the OLS estimator $\mathbf{b}$ is the BLUE (Best Linear Unbiased Estimator) for the parameters vector $\boldsymbol{\beta}$ in the multiple regression model with so called classic assumptions 1 - 3.
We have to add here, that OLS method(derived first by K. Gauss and investigated later by A. Markov) is no longer the best method for estimation of $\boldsymbol{\beta}$ if we put other (more general) assumptions than those 1-3 for the estimated model(see Chapter 11).

8.3. Residual Variance and Estimation of Standard Errors.

Since the variance σ^2 of the random disturbance ε_t is unknown, we have to estimate it also for getting variance $D^2(b_i) = \sigma^2 d_{ii}$ for i=1,2,..., k+1 as diagonal elements of the covariance matrix $V(\mathbf{b}) = \sigma^2 (\mathbf{X}^T\mathbf{X})^{-1}$, and in consequence for getting the standard errors of estimates $SE(b_i) = D(b_i) = \sigma\sqrt{d_{ii}}$.
It is easy to guess, that so called residuals i.e. differences $\hat{\varepsilon}_t = y_t - \hat{y}_t$ where $\hat{y}_t = b_1 x_{t1} + b_2 x_{t2} + \ldots + b_k x_{tk} + b_{k+1}$ can be taken as estimates of unobservable random disturbances ε_t. Consequently the residual variance can be taken as an estimator for the unknown variance σ^2.
The following related Theorem can be formulated:

The residual variance defined as $S^2 = \dfrac{\sum_{t=i}^{n} \hat{\varepsilon}_t^2}{n-k-1} = \dfrac{\sum_{t=i}^{n} (y_t - \hat{y}_t)^2}{n-k-1}$ is an unbiased estimator of the unknown parameter σ^2 i.e. variance of random disturbances in the multiple linear regression model with assumptions 1 - 3 .

The short proof of this theorem is very instructive since it fully utilises the matrix notation. Let's denote the vector of residuals as $\hat{\varepsilon} = y - \hat{y}$ where $\hat{y} = Xb$.

It can be transform into a linear function of the disturbance vector ε as follows:

$$\hat{\varepsilon} = y - Xb = X\beta + \varepsilon - Xb = \varepsilon - X(b - \beta) =$$
$$= \varepsilon - X(X^TX)^{-1}\varepsilon = [I - X(X^TX)^{-1}X^T]\varepsilon =$$
$$= M\varepsilon,$$

where $M = I - X(X^TX)^{-1}X^T$ is an idempotent matrix with property that $r(M) = tr(M)$ (see paragraph 3.4)

Now, let's take under consideration the numerator of the residual variance S^2.

$$\sum_{t=i}^{n} \hat{\varepsilon}_t^2 = \hat{\varepsilon}^T\hat{\varepsilon} = (M\varepsilon)^T(M\varepsilon) = \varepsilon^T M^T M \varepsilon = \varepsilon^T M \varepsilon$$

(since M is an idempotent matrix).

The expected value of that numerator is equal to:

$$E\left(\sum_{t=i}^{n} \hat{\varepsilon}_t^2\right) = E(\varepsilon^T M \varepsilon) = E\,tr(\varepsilon^T M \varepsilon) \quad \text{(since } \varepsilon^T M \varepsilon \text{ is a scalar)}$$
$$= E\,tr(M\varepsilon\varepsilon^T) = tr\,E(M\varepsilon\varepsilon^T) \quad \text{(since tr and E are linear operators)}$$
$$= tr[M \cdot E(\varepsilon\varepsilon^T)] = tr[M \cdot \sigma^2 I]$$
$$= \sigma^2 \cdot tr\,M$$

Trace of the matrix M can be easily obtained as: $tr\,M = tr[I_n - X(X^TX)^{-1}X^T]$
$$= n - tr[X^TX(X^TX)^{-1}],$$

$$tr\,M = n - tr\,I_{k+1} = n - (k+1) = n - k - 1$$

So, $$E(S^2) = E\left(\frac{\sum_{t=i}^{n} \hat{\varepsilon}_t^2}{n-k-1}\right) = \frac{1}{n-k-1} E\left(\sum_{t=i}^{n} \hat{\varepsilon}_t^2\right)$$

$$= \frac{1}{n-k-1}\sigma^2 tr\,M = \frac{1}{n-k-1}\sigma^2(n-k-1) = \sigma^2.$$

$E(S^2) = \sigma^2$ means that the residual variance $S^2 = \frac{\sum_{t=i}^{n} \hat{\varepsilon}_t^2}{n-k-1}$ is an unbiased estimator of the variance σ^2 of random disturbances.

For the numerical calculations purposes we can transform the numerator of the residual variance S^2 into simpler form:

$$\sum_{t=1}^{n} \hat{\varepsilon}_t^2 = \hat{\varepsilon}^T \hat{\varepsilon} = (y - Xb)^T (y - Xb) = y^T y - b^T X^T y - y^T X b + b^T X^T X b = y^T y - 2 b^T X^T y + b^T X^T X \cdot (X^T X)^{-1} X^T y = y^T y - b^T X^T y$$

So, $$S^2 = \frac{\sum_{t=1}^{n} \hat{\varepsilon}_t^2}{n-k-1} = \frac{y^T y - b^T X^T y}{n-k-1}$$

Now, we can estimate the full covariance matrix $V(b) = \sigma^2 (X^T X)^{-1}$ as $\hat{V}(b) = s^2 (X^T X)^{-1}$, and consequently its diagonal elements i.e. variances $D^2(b_i) = \sigma^2 \cdot d_{ii} = \sigma_{b_i}^2$ can be estimated as $S_{b_i}^2 = S^2 d_{ii}$ where d_{ii} is the i-th diagonal element of the inverse matrix $(X^T X)^{-1}$.

Finally, the standard errors of the OLS estimators i.e. $SE(b_i) = D(b_i) = \sigma \sqrt{d_{ii}} = \sigma_{b_i}$ can be estimated as $S_{b_i} = S \sqrt{d_{ii}}$, where $S = \sqrt{S^2}$ is a consistent estimator of $\sigma = \sqrt{\sigma^2}$.

Let us summarize the basic results discussed in this chapter in the following table:

NAME	SYMBOLS	ESTIMATES AND PROPERTIES
Structural Parameters	β	$b = (X^T X)^{-1} X^T y$, BLUE
Variance of Disturbances	σ^2	$s^2 = \frac{\hat{\varepsilon}^T \hat{\varepsilon}}{n-k-1} = \frac{y^T y - b^T X^T y}{n-k-1}$ Unbiased
Covariance Matrix	$V(b) = \sigma^2 (X^T X)^{-1}$	$\hat{V}(b) = s^2 (X^T X)^{-1}$ Unbiased
Variances of the OLS Estimates	$D^2(b_i) = \sigma^2 d_{ii} = \sigma_{b_i}^2$	$S_{b_i}^2 = S^2 d_{ii}$ Unbiased
Standard Errors of the OLS Estimates	$SE(b_i) = D(b_i) = \sigma \sqrt{d_{ii}} = \sigma_{b_i}$	$S_{b_i} = S \sqrt{d_{ii}}$ Consistent

Examples:

1. We suppose, that the monthly expenditure for non-food commodities Y in a household depends on the household's income (X_1) and on the expenditure on food commodities (X_2). For n = 5 households the empirical data (in hundred Birr) are as follows:

t	x_{t1}	x_{t2}	y
1	2	1	0.7
2	3	2	0.9
3	4	2	1.3
4	5	3	1.5
5	6	3	2.0

Based on these data we want to estimate the linear model:

$$y_t = \beta_1 x_{t1} + \beta_2 x_{t2} + \beta_3 + \varepsilon_t \qquad \text{for } t = 1,2,\dots 5,$$

or in matrix notation $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

where $\mathbf{y} = \begin{pmatrix} 0.7 \\ 0.9 \\ 1.3 \\ 1.5 \\ 2.0 \end{pmatrix}$, $\mathbf{X} = \begin{pmatrix} 2 & 1 & 1 \\ 3 & 2 & 1 \\ 4 & 2 & 1 \\ 5 & 3 & 1 \\ 6 & 3 & 1 \end{pmatrix}$, $\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}$, $\boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \end{pmatrix}$

with $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $V(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}$.

The OLS estimator of $\boldsymbol{\beta}$ is $\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$

The successive calculations are:

$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} 90 & 49 & 20 \\ 49 & 27 & 11 \\ 20 & 11 & 5 \end{pmatrix}, \quad \det(\mathbf{X}^T\mathbf{X}) = 15$$

$$(\mathbf{X}^T\mathbf{X})^{-1} = \frac{1}{15}\begin{pmatrix} 14 & -25 & -1 \\ -25 & 50 & -10 \\ -1 & -10 & 29 \end{pmatrix}, \quad \mathbf{X}^T\mathbf{y} = \begin{pmatrix} 28.8 \\ 15.6 \\ 6.4 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 0.453 \\ -0.267 \\ 0.053 \end{pmatrix}$$

$\mathbf{y}^T\mathbf{y} = 9.24$, $\mathbf{b}^T\mathbf{X}^T\mathbf{y} = 9.2356$, $S^2 = \frac{9.24-9.2356}{2} = 0.0022$, $S \approx 0.05$

So, the OLS estimate of the model is $\hat{y}_t = 0.453\, x_{t1} - 0.267\, x_{t2} + 0.053$, and the covariance matrix of the estimator $\mathbf{b}$ is estimated as:

$$\hat{V}(\mathbf{b}) = 0.0022 \begin{pmatrix} \frac{14}{15} & \frac{-25}{15} & \frac{-1}{15} \\ \frac{-25}{15} & \frac{50}{15} & \frac{-10}{15} \\ \frac{-1}{15} & \frac{-10}{15} & \frac{29}{15} \end{pmatrix}$$

The standard errors of the OLS estimates $SE(b_1)$ and $SE(b_2)$ are estimated respectively as $S_{b_1} = 0.05\sqrt{\frac{14}{15}} \approx 0.0485$,

$S_{b_2} = 0.05\sqrt{\frac{50}{15}} \approx 0.0910$.

Numerical results of the estimation $\hat{y}_t = \underset{(0.048)}{0.453\, x_{t1}} - \underset{(0.091)}{0.267\, x_{t2}} + 0.053$

can be interpreted as follows: the increase of the monthly household's income by 100 Birr will increase the monthly expenditure for non-food commodities by 45 Birr on average, while the increase of the monthly expenditure on food commodities by 100 Birr will decrease the expenditure for non-food commodities by 27 Birr (on average).

2. Numerical results of the estimation (by the OLS method) of the linear model for production of wheat in private peasant holdings in Arsi region of Ethiopia (based on cross-sectional data) are as follows:

$$\hat{y} = \underset{(1.14)}{11.70\, x_1} + \underset{(1.30)}{2.82\, x_2} + \underset{(0.94)}{1.62\, x_3} - \underset{(0.97)}{3.39\, x_4} + 2.72$$

where y is the production of wheat (in quintals) in a private peasant holding

x_1 is the area under wheat (in hectares) in a private peasant holding

x_2 is the use of improved seed (0 - 1 variable) in a peasant holding

x_3 is the fertilizer applied (0-1 variable) in a peasant holding

x_4 is the crop damage observed in a peasant holding

From the estimated model we can conclude, that the marginal wheat yield is 11.7 q/ha, the wheat production increases on average by 2.82 q per holding in which an improved seed was used, the wheat production increases on average by 1.62 q per holding in which a fertilizer was applied, and it decreases on average by 3.39 q per holding due to crop damage which occured in a peasant holding.

CHAPTER 9. STATISTICAL INFERENCE IN THE MULTIPLE NORMAL REGRESSION MODEL.

9.1. Normality Assumption and Its Consequences.

We are still discussing the multiple regression model with classic assumptions:

1. $\mathcal{Y} = \mathcal{X}\beta + \mathcal{E}$
2. Matrix $\mathcal{X}$ is non-random, with full rank $r(\mathcal{X}) = k+1 \leq n$.
3. Random disturbance vector $\mathcal{E}$ has the mean value $E(\mathcal{E}) = \mathcal{O}$, and the covariance matrix $V(\mathcal{E}) = \sigma^2 I$.

The point estimation technique by means of the Ordinary Least Squares (OLS) method for getting an estimate of the unknown parameters vector β does not utilize any assumption about distribution type of the random disturbance vector $\mathcal{E}$.

However, for the interval estimation as well for the statistical hypotheses testing purposes the additional assumption on normality of the probability distribution of disturbances has to be included.

So, let's put the following additional assumption:

4. Random disturbance vector $\mathcal{E}$ has n-dimensional normal distribution $N(\mathcal{O}, \sigma^2 I)$ with the joint probability density function(pdf) defined as:

$$f(\mathcal{E}) = (2\pi\sigma^2)^{-\frac{n}{2}} \cdot e^{-\frac{1}{2\sigma^2}\mathcal{E}^T\mathcal{E}}$$

There are two main consequences ofthat additional normality assumption 4:

(i) It allows to apply the Maximum Likelihood (ML) method for getting an estimator of the parameters vector β of the model.

(ii) It allows to perform the statistical inference (by means of confidence intervals and statistical tests) concerning the β .

For(i) let's remind, that the likelihood function for a continous probability distribution is defined as the joint density of a given sample.

In our case, for the normal distribution of the vector $\mathcal{E} = \mathcal{Y} - \mathcal{X}\beta$ (and in consequence of the vector $\mathcal{Y}$ of sample results), the likelihood function is given by:

$$L = f(y) = f(\varepsilon) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}(y-X\beta)^T(y-X\beta)\right\}$$

Its natural logarithm is:

$$\ln L = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y-X\beta)^T(y-X\beta)$$

Now, the necessary condition for maximum, i.e. the equation:

$$\frac{\partial \ln L}{\partial \beta} \equiv -\frac{1}{2\sigma^2}\cdot\frac{\partial[(y-X\beta)^T(y-X\beta)]}{\partial \beta} = 0$$

has the same solution as the equation $\frac{\partial S}{\partial \beta}=0$ in the OLS method.

So, the Maximum Likelihood(ML) estimator of β (under additional mormality assumption) is exactly the same as the OLS estimator $b=(X^TX)^{-1}X^Ty$ which is BLUE of the parameters vector β.

(ii) The statistical inference concerning elements of the vector β, is based on the basic theorem on the t-Student's statistic.

Because of assumed normal distribution of the sample vector y, its linear function $b=(X^TX)^{-1}X^Ty$ i.e. the OLS estimator has also the normal distribution (k+1- dimensional) with the mean vector $E(b)=\beta$ and with the covariance matrix $V(b)=\sigma^2(X^TX)^{-1}$.

It follows, that each element b_i $(i=1,2,\ldots, k+1)$ of the vector b has one-dimensional(marginal) normal distribution $N(\beta_i, \sigma^2_{b_i})$, i.e. with the expected value $E(b_i)=\beta_i$ and with the variance $D^2(b_i)=\sigma^2_{b_i}=\sigma^2\cdot d_{ii}$, where d_{ii} is the i-th diagonal element of the inverse matrix $(X^TX)^{-1}$.

So, by standardization we have, that the random variable

$$U=\frac{b_i-\beta_i}{\sigma_{b_i}} \sim N(0,1) \quad \text{where} \quad \sigma_{b_i}=\sigma\sqrt{d_{ii}}$$

Unfortunately, we can not utilize that normal distribution for statistical inference purposes, because the variance σ^2 of the random disturbances is unknown.

Instead, estimating σ^2 by the residual variance $S^2 = \frac{\hat{\varepsilon}^T\hat{\varepsilon}}{n-k-1}$,

and consequently σ_{b_i} by $S_{b_i} = S\sqrt{d_{ii}}$, we shall utilize the following basic theorem:

Theorem.

If the linear multiple regression model with the assumptions 1-4 has been estimated by the OLS method, then the statistic t defined for each element b_i of the OLS estimator-vector b as

$$t = \frac{b_i - \beta_i}{s_{b_i}} = \frac{b_i - \beta_i}{S\sqrt{d_{ii}}}$$ has the Student's t distribution with df = n-k-1

Let us give a full proof of this theorem, since it is instructive and short. The proof is based on the definition of the Student's t distribution, i.e.

$$t = \frac{u}{\sqrt{\chi^2_k/k}}$$ with df = k, if $U \sim N(0,1)$,

$\chi^2_k \sim$ Chi-square distribution with df = k,

U and χ^2_k independent random variables.

For our purposes, let $U = \frac{b_i - \beta_i}{\sigma_{b_i}} \sim N(0,1)$ and let's take under consideration the following random variable:

$$t = \frac{b_i - \beta_i}{S_{b_i}} = \frac{\frac{b_i - \beta_i}{\sigma_{b_i}}}{\frac{S_{b_i}}{\sigma_{b_i}}} = \frac{U}{\sqrt{\frac{S^2_{b_i}}{\sigma^2_{b_i}}}} = \frac{U}{\sqrt{\frac{S^2 \cdot d_{ii}}{\sigma^2 \cdot d_{ii}}}} = \frac{U}{\sqrt{\frac{S^2}{\sigma^2}}} =$$

$$= \frac{U}{\sqrt{\frac{\hat{\varepsilon}^T\hat{\varepsilon}}{\frac{n-k-1}{\sigma^2}}}} = \frac{U}{\sqrt{\frac{\varepsilon^T M \varepsilon}{\frac{\sigma^2}{n-k-1}}}}$$

Now, because the random vector $\varepsilon \sim N(0, \sigma^2 I)$,
then $\frac{\varepsilon}{\sigma} \sim N(0, I)$, and consequently (see paragraph 4.4)

$\frac{\varepsilon^T M \varepsilon}{\sigma^2} = \left(\frac{\varepsilon}{\sigma}\right)^T M \left(\frac{\varepsilon}{\sigma}\right) = \chi^2_{n-k-1}$ has the Chi-square distribution with the number of degrees of freedom $df = r(M) = tr M = n-k-1$, since the matrix M is an idempotent matrix.

So, we obtain $t = \frac{U}{\sqrt{\frac{\chi^2_{n-k-1}}{n-k-1}}}$, and this statistic is the Student's t statistic,

if the random variable U and χ^2_{n-k-1} are stochastically independent.
To show this, let's remind that $b_i - \beta_i$ in the random variable U is an element of the random vector $b - \beta = A\varepsilon$, where $A = (X^T X)^{-1} X^T$.
The quadratic form $\varepsilon^T M \varepsilon$ and the linear form $A\varepsilon$ are stochastically independent, because the product of their matrices is a null matrix, i.e.

$$AM = (X^T X)^{-1} X^T [I - X(X^T X)^{-1} X^T] =$$

$$= (X^T X)^{-1} X^T - \underbrace{(X^T X)^{-1} X^T \cdot X (X^T X)^{-1} X^T}_{I} = (X^T X)^{-1} X^T - (X^T X)^{-1} X^T = 0$$

That completes the proof of the theorem, that the statistic

$$t = \frac{b_i - \beta_i}{S_{b_i}} \sim \text{Student's-t distribution with } df = n-k-1$$

In the next paragraph we shall see its utilization.

9.2. Interval Estimation and Significance Tests.

Based on the Student's-t statistic from the last theorem, the confidence interval for each estimated parameter β_i of the multiple regression model can be constructed.
For the given confidence coefficient $1 - \alpha$, we can obtain (from Statistical Tables of the Student's t distribution for $df = n-k-1$) such a value t_α, that probability $P\{-t_\alpha < t < t_\alpha\} = 1 - \alpha$.

If we take now $t = \frac{b_i - \beta_i}{S_{b_i}}$ where b_i is the OLS estimate of the β_i and S_{b_i} is its estimated standard error $S_{b_i} = S\sqrt{d_{ii}}$, we have $P\left\{-t_\alpha < \frac{b_i - \beta_i}{S_{b_i}} < t_\alpha\right\} = 1 - \alpha$.

After elementary transformation of the inequality, we obtain the following formula for the Confidence Interval for the parameter β_i ($i = 1,2, \ldots, k+1$):

$$P\left\{b_i - t_\alpha \cdot S_{b_i} < \beta_i < b_i + t_\alpha \cdot S_{b_i}\right\} = 1 - \alpha$$

Testing of statistical significance of the estimated parameter β_i can be performed based on the following significance test:
Let $H_0 : \beta_i = 0$ against the alternative $H_1 : \beta_i \neq 0$ be the statistical hypothesis, which we would like to verify based on the sample from which we obtained the OLS estimate of β_i as b_i with standard error of that estimate as S_{b_i}.

If the hypothesis Ho is true, the statistic defined as:

$t = \frac{b_i - 0}{S_{b_i}} = \frac{b_i}{S_{b_i}}$ has the Student's t distribution with $df = n-k-1$.

Based on this statistic we can build the critical region Q as:
$Q = \left\{t : |t| \geq t_\alpha\right\}$, where t_α is the critical value obtained from Statistical Tables of t Student's distribution with $df = n-k-1$ in such a way, that for given small significance level α, $P\left\{|t| \geq t_\alpha\right\} = \alpha$.
So, the significance test is as follows:
After estimation of the model, the empirical value of the t Student's statistic should be computed as $t = \frac{b_i}{S_{b_i}}$ and then compared with the critical value t_α, i.e. with the critical region Q constructed for the given in advance small significance level α.

If $t \in Q$, the hypothesis H_o is rejected, which means that the estimate of the parameter β_i is statistically significant at the significance level α . (i.e. it significantly differs from zero).
If $t \notin Q$, we can not reject the hypothesis $H_o : \beta_i = 0$, which means that we have no statistical evidence, that the explanatory variable X_i really has any impact on the dependent variable Y, since the estimate of its parameter β_i is statistically non-significant.
So, only the rejection of the hypothesis H_o gives us a justification for keeping the explanatory variable X_i in the model.

An utilization of that significance test during the stage of checking of estimated econometric model is essential, because only models having all explanatory variables with statistically significant parameters can be used for forecasting purposes.

The additional normality assumption 4 allows also to build the significance test for the full parameters vector β simultanuously based on F-Snedecor statistic.
Sometimes we would like to test the vector hypothesis:
$H_o : \beta = \beta_o$ against alternative $H_i : \beta \neq \beta_o$ where β is a given vector of numbers.

Let's consider the statistic defined as $B = (b-\beta_o)^T X^T X (b-\beta_o)$.
If the null hypothesis H_o is true $b-\beta_o = b-\beta = (X^TX)^{-1}X^T\varepsilon$.
In that case, the statistic B can be transformed into such a quadratic form of the disturbance vector ε :

$$B = [(X^TX)^{-1}X^T\varepsilon]^T \cdot (X^TX) \cdot [(X^TX)^{-1}X^T\varepsilon] =$$
$$= \varepsilon^T X (X^TX)^{-1} \underbrace{(X^TX)(X^TX)^{-1}}_{I} X^T\varepsilon = \varepsilon^T X (X^TX)^{-1} X^T\varepsilon =$$
$$= \varepsilon^T [I - M] \varepsilon$$

, where $M = I - X(X^TX)^{-1}X^T$ is an idempotent matrix, and $\varepsilon \sim N(0, \sigma^2 I)$

So $\frac{B}{\sigma^2} = \frac{\varepsilon^T[I-M]\varepsilon}{\sigma^2}$ has the χ^2-distribution with

$df = r(I-M) = tr(I-M) =$
$= tr\, I_n - tr M = n - (n - k - 1) = k+1$

Previously we had, that $\frac{\hat{\varepsilon}^T\hat{\varepsilon}}{\sigma^2} = \frac{\varepsilon^T M \varepsilon}{\sigma^2}$ has the χ^2 distribution with df = n-k-1

So, we have two quadratic forms of the same random vector ε :

$$\frac{B}{\sigma^2} = \chi^2_{k+1} \quad \text{and} \quad \frac{\hat{\varepsilon}^T\hat{\varepsilon}}{\sigma^2} = \chi^2_{n-k-1}.$$

These quadratic forms are stochastically independent because the product of their matrices is a null matrix : $(I-M)M = M - M^2 = M - M = 0$.

Therefore, according to the definition of the F-Snedecor distribution (see paragraph 4.4) the quotient statistic defined as:

$$F = \frac{\dfrac{\chi^2_{k+1}}{k+1}}{\dfrac{\chi^2_{n-k-1}}{n-k-1}} = \frac{\dfrac{B}{\sigma^2(k+1)}}{\dfrac{\hat{\varepsilon}^T\cdot\hat{\varepsilon}}{\sigma^2(n-k-1)}} = \frac{\dfrac{B}{k+1}}{S^2} = \frac{B}{(k+1)S^2}$$

has the F-Snedecor distribution with the numbers of degrees of freedom respectively k+1 and n-k-1.

For a given significance level α and for k+1 and n-k-1 degrees of freedom we can find in the Statistical Tables of F-distribution such a critical value F_α , that probability $P\{F \geqslant F_\alpha\} = \alpha$.

Now, the critical region for the significance test can be defined as $Q = \{F : F \geqslant F_\alpha\}$.

Finally, the significance test for the vector hypothesis $H_o : \beta = \beta_o$ is as follows:

After estimation of the multiple regression model by the OLS method, the empirical value of the statistic F should be computed as $F = \frac{B}{(k+1)S^2}$ and then compared with the critical region Q constructed for the given significance level .

If $F \in Q$, the vector hypothesis H_o is rejected at the significance level α.

If $F \notin Q$, we can not reject the hypothesis $H_o : \beta = \beta_o$.

For the special case $\beta_o = 0$, we have $B = b^T X^T X b = \hat{y}^T\hat{y}$.

9.3. Analysis of Variance and the Coefficient of Determination R^2.

To continue to discuss the multiple regression model with 1-4 assumptions (including normality of random disturbances) we shall deal now with a decomposition of sum of squares of observations on the explained variable Y. It will lead us to an interpretation of the determination coefficient R^2 used for measurement of goodness of fit of the model to the empirical data.

The starting point is the equation $\hat{\varepsilon}^T\hat{\varepsilon} = y^Ty - b^TX^Ty$ obtained in paragraph 8.3.

Recalling that $\hat{y} = Xb$, and, $b = (X^TX)^{-1}X^Ty$ we have also:

$$\hat{y}^T y = (Xb)^T \cdot (Xb) = b^TX^TXb = b^T\underbrace{X^TX \cdot (X^TX)^{-1}}_{I}X^Ty = b^TX^Ty$$

So, we obtain $\hat{\varepsilon}^T\hat{\varepsilon} = y^Ty - \hat{y}^T\hat{y}$

or $y^Ty = \hat{\varepsilon}^T\hat{\varepsilon} + \hat{y}^T\hat{y}$

or in conventional notation

we obtain the following primarly decomposition:

$$\sum_{t=1}^{n} y_t^2 = \sum_{t=1}^{n} \hat{\varepsilon}_t^2 + \sum_{t=1}^{n} \hat{y}_t^2$$

Now we shall prove the following useful Theorem:

If the model with assumptions 1-3 was estimated by the OLS method then the average of residuals is zero, i.e. $\bar{\hat{\varepsilon}} = \frac{1}{n}\sum_{t=1}^{n} \hat{\varepsilon}_t = 0.$

To prove this Theorem, let's consider the OLS estimator b of the parameters vector β : $b = (X^TX)^{-1}X^Ty$

$$(X^TX)b = \underbrace{(X^TX)\cdot(X^TX)^{-1}}_{I}X^Ty$$ (by the left hand side multiplication)

$$X^TXb = X^Ty$$

So, for the last (k+1) -th equation we have :

$$1^TXb = 1^Ty, \quad \text{where } 1^T = (1\ 1\ \dots\ 1)$$

or $1^T\hat{y} = 1^Ty$

$$\text{or} \quad \sum_{t=1}^{n} \hat{y}_t = \sum_{t=1}^{n} y_t , \qquad \text{or} \quad \sum_{t=1}^{n} (y_t - \hat{y}_t) = 0, \qquad \text{or} \quad \sum_{t=1}^{n} \hat{\varepsilon}_t = 0 .$$

From the last equation we obtain finally, that:

$$\bar{\hat{\varepsilon}} = 0 \qquad \text{and} \qquad \bar{y} = \bar{\hat{y}}.$$

Now, let us come back to the primary decomposition of sum of squares

$$y^T y = \hat{\varepsilon}^T \hat{\varepsilon} + \hat{y}^T \hat{y}$$

and let's subtract from both sides the term:

$$n\bar{y}^2 = n\left(\frac{1}{n}\sum_{t=1}^{n} y_t\right)^2 = \frac{1}{n}\left(1^T y\right)^2 = n\bar{\hat{y}}^2 = \frac{1}{n}\left(1^T \hat{y}\right)^2$$

So, we have now $y^T y - \frac{1}{n}\left(1^T y\right)^2 = \hat{\varepsilon}^T \hat{\varepsilon} + \hat{y}^T \hat{y} - \frac{1}{n}\left(1^T \hat{y}\right)^2$

or in conventional notation:

$$\underbrace{\sum_{t=1}^{n} (y_t - \bar{y})^2}_{\text{SST}} = \underbrace{\sum_{t=1}^{n} \hat{\varepsilon}_t^2}_{\text{SSE}} + \underbrace{\sum_{t=1}^{n} (y_t - \bar{\hat{y}})^2}_{\text{SSR}}$$

$$\text{SST} = \text{SSE} + \text{SSR}$$

So, the final decomposition of the empirical total sum of squares (SST) is as follows: the total empirical sum of squares(SST) has been split into two separate terms-the sum of squares of residuals(errors) SSE and the sum of squares of estimated(regression) values SSR.
Now, taking into account normality assumption 4, we have:

$$\frac{\text{SSE}}{\sigma^2} = \chi^2_{n-k-1} , \qquad \frac{\text{SSR}}{\sigma^2} = \chi^2_k$$

and

$$\frac{\chi^2_k / k}{\chi^2_{n-k-1} / n-k-1} = \frac{\text{SSR}/k}{\text{SSE}/n-k-1} = F \sim F$$ - Snedecor distribution with K and n-k-1 degrees of freedom

Summarizing we have the following Analysis of Variance for our linear multiple regression model with assumptions 1-4:

Analysis of Variance Table

Source of Dispersion	Sum of Squares	df	Mean Square	Test F
Regression	SSR = $= b^T X^T y - \frac{1}{n}(1^T y)^2$	k	$S_1^2 = \frac{SSR}{k}$	$F = \frac{S_1^2}{S^2} = \frac{b^T X^T y - \frac{1}{n}(1^T y)^2}{k S^2}$
Residual (Random Error)	SSE = $= \hat{\varepsilon}^T \hat{\varepsilon} = y^T y - b^T X^T y$	n-k-1	$S^2 = \frac{SSE}{n-k-1}$	
Total	SST = $= y^T y - \frac{1}{n}(1^T y)^2$	n-1	×	×

Based on the above decomposition of the empirical total sum of squares SST we can define the Coefficient of Determination R^2 as follows:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = 1 - \frac{y^T y - b^T X^T y}{y^T y - \frac{1}{n}(1^T y)^2}$$

It follows, that $0 \leq R^2 \leq 1$,

$R^2 = 0 \Leftrightarrow SSR = 0 \Rightarrow SSE = SST$, i.e. $\bigwedge_t \hat{y}_t = \bar{y}$

$R^2 = 1 \Leftrightarrow SSE = 0 \Rightarrow SSR = SST$, i.e. $\bigwedge_t \hat{y}_t = y_t$

Therefore the Determination Coefficient R^2 is a measure of goodness of fit of the linear model to the empirical data.

The numerical value of R^2 means how much (in percentages) of the total empirical dispersion of the dependent variable Y can be explained by the linear regression mo

e.g. $R^2 = 0.80$ means, that 80% of the empirical dispersion of the dependent variable is explained by the regression on explanatory variables $X_1, X_2, \ldots, X_k$, and remaining 20% of the empirical dispersion is due to the random disturbances.

It can be proved, that if $X_1, X_2, \ldots, X_k$ are random variables, $R = \sqrt{R^2}$ is the multiple correlation coefficient.

However it is important to remember, that for our model with assumption 2 (about non-random matrix $\mathbb{X}$), we shall rather use R^2 only (and not R).

Taking into account the normality assumption 4, we can construct the significance test for the Determination Coefficient R^2, based on the following F-Snedecor statistic:

$$F = \frac{R^2}{1-R^2} \cdot \frac{n-k-1}{k} = \frac{SSR}{SST} \cdot \frac{SST}{SSE} \cdot \frac{n-k-1}{k} = \frac{SSR}{SSE} \cdot \frac{n-k-1}{k} = \frac{SSR}{k\,S^2} =$$

$$= \frac{b^T X^T y - \frac{1}{n}(1^T y)^2}{k\,S^2} \sim$$

F- Snedecor distribution with k and n-k-1 degrees of freedom.

It is worth to realize, that the significance test F for the Determination coefficient R^2 is equivalent to the test F for the vector hypothesis $H_0: \beta = 0$ (except the constant term β_{k+1}) described in former paragraph 9.2. The critical region is defined as $Q = \{ F: F \geq F_\alpha \}$, where F_α is the critical value from Tables of F-distribution for the given significance level α and for k and n-k-1 degrees of freedom.

If the empirical value of the statistic $\frac{R^2}{1-R^2} \cdot \frac{n-k-1}{k} = F \in Q$, the Determination coefficient R^2 is statistically significant.

If $F \notin Q$, R^2 is not significant (multiple regression is statistically non-significant).

When different linear multiple regression models with a different number of variables have to be compared, so called Adjusted R^2 is utilized.

The Adjusted R^2 is denoted usually by $\bar{R}^2$, and it is defined as

$$\bar{R}^2 = R^2 - \frac{k}{n-k-1}(1-R^2)$$

9.4. Forecasting Based on the Multiple Linear Regression Model.

Let t=T be future (forecasted) period of time and let's assume, that the classical assumptions 1-3 concerning the multiple regression model are valid also for the future period of time T.
It means, that

$$y_T = \beta_1 X_{T1} + \beta_2 X_{T2} + \dots + \beta_k X_{Tk} + \beta_{k+1} + \varepsilon_T$$

where y_T is the forecasted value of the dependent variable Y,

$X_{T1}, X_{T2}, \dots, X_{Tk}$ are given (assumed) future realizations of explanatory variables,

ε_T is the future random disturbance with $E(\varepsilon_T) = 0$, $D^2(\varepsilon_T) = \sigma^2$, and $Cov(\varepsilon_T, \varepsilon_t) = 0$ for $t=1,2,\dots,n$.

If we denote by $\mathbf{X}_{(T)}$ the vector of given future realizations of explanatory variables $\mathbf{X}^T_{(T)} = (x_{T1}, x_{T2}, \dots, x_{Tk}, 1)$ the forecasted value of the explained variable can be written as:

$$y_T = \mathbf{X}^T_{(T)}\boldsymbol{\beta} + \varepsilon_T =$$

$$= E(y_T) + \varepsilon_T, \quad \text{where } E(y_T) = \mathbf{X}^T_{(T)}\boldsymbol{\beta}$$

Any forecast of y_T should therefore consist of an estimate of the mean value $E(y_T) = \mathbf{X}^T_{(T)}\boldsymbol{\beta}$, and of a forecast of ε_T.
The best estimate of the first term $E(y_T)$ is of course equal to $\mathbf{X}^T_{(T)}\mathbf{b}$ (where $\mathbf{b}$ is the OLS estimate of the parameters vector $\boldsymbol{\beta}$), while the best forecast of ε_T is its mean value, i.e. zero. So, the forecast $\hat{y}_T$ of the forecasted value y_T is given by:

$$\hat{y}_T = \mathbf{X}^T_{(T)}\mathbf{b} + 0 = \mathbf{X}^T_{(T)}\mathbf{b}$$

The variance of that forecast will be equal to

$$\mathbb{D}^2(\hat{y}_T) = \mathbb{D}^2(X_{(T)}^T \cdot b) + \mathbb{D}^2(\varepsilon_T) =$$

$$= \mathbb{D}^2(X_{(T)}^T \cdot b) + \sigma^2$$

The first term, i.e. the variance of the estimate $X_{(T)}^T b$ can be obtain as:

$$\mathbb{D}^2(X_{(T)}^T b) = E[X^T b - X^T \beta]^2 = E[X_{(T)}^T (b-\beta)]^2 =$$

$$= E\{[X_{(T)}^T (b-\beta)] \cdot [X_{(T)}^T (b-\beta)]^T\} =$$

$$= X_{(T)}^T \cdot E[(b-\beta)(b-\beta)^T] \cdot X_{(T)} = X_{(T)}^T \cdot V(b) \cdot X_{(T)} =$$

$$= X_{(T)}^T \cdot \sigma^2 (X^T X)^{-1} \cdot X_{(T)} = \sigma^2 X_{(T)}^T (X^T X)^{-1} X_{(T)}$$

So, the variance of the forecast $\hat{y}_T$ of the future value y_T is finally given by the formula:

$$\mathbb{D}^2(\hat{y}_T) = \sigma^2 X_{(T)}^T (X^T X)^{-1} X_{(T)} + \sigma^2 =$$

$$= \sigma^2 [1 + X_{(T)}^T (X^T X)^{-1} X_{(T)}]$$

The standard error $SE(\hat{y}_T)$ of the forecast $\hat{y}_T$ is therefore given by:

$$SE(\hat{y}_T) = \sigma_{\hat{y}_T} = \mathbb{D}(\hat{y}_T) = \sigma \sqrt{1 + X_{(T)}^T (X^T X)^{-1} X_{(T)}}$$

and its estimate is defined as:

$$s_{\hat{y}_T} = s \sqrt{1 + X_{(T)}^T (X^T X)^{-1} X_{(T)}} \quad , \text{ where } s = \sqrt{s^2} = \sqrt{\frac{e^T e}{n-k-1}}$$

By adoption of the normality assumption 4, it is easy to prove, that the statistic defined as:

$$t = \frac{\hat{y}_T - y_T}{S_{\hat{y}T}}$$ has t-Student's distribution with df = n-k-1.

Based on that Student's -t distribution one can build an interval forecast i.e. the Confidence Interval for the forecasted value y_T as:

$$P\left\{\hat{y}_T - t_{\alpha,\, n-k-1} \cdot S_{\hat{y}T} < y_T < \hat{y}_T + t_{\alpha,\, n-k-1} \cdot S_{\hat{y}T}\right\} = 1-\alpha$$

where $t_{\alpha,\, n-k-1}$ is the value of t-Student's statistic taken from the Statistical Tables for the given confidence coefficient $1-\alpha$ and for df=n-k-1.

9.5. Use of Computers and Examples.

During the last twenty five years collections of computer programs (Packages) have been developed for carrying out all calculations for the estimation and checking of the multiple regression models. Examples of such packeges are: SPSS, BMDP , GLIM and ISEA, which has been particularly designed for econometric modelling. Versions of these packages exist for many different types of computers like IBM, HP, ICL, etc. One of the most widely distributed of those packeges is SPPS(Statistical Package for the Social Sciences) containing computer programs for regression analysis. Use of these programs is described in the manual SPSS and SPSS Update.

The only work for the user of these packages is to prepare his data in a form acceptable to the package and to prepare instructions to run the appropriate program.

An example of SPSS regression printout from the HP-3000 computer installed at CSO Addis Ababa is given on the next page.

It contains regression analysis of the linear model for production of wheat in private peasant holdings in Shoa region of Ethiopia. The model included 6 explanatory variables listed at the beginning of the printout.

Then we have results on Determination Coefficient R^2 (R SQUARE), $\bar{R}^2$(ADJUSTED R SQUARE), Analysis of Variance with F statistic, results of the OLS estimation(estimates b_i for each i-th variable are printed as "B") with standard errors (SEB) and t-Student's statistics (T).

Numerical results of that outprint are used for a statistical inference performed in the Example 1.

* * * * MULTIPLE REGRESSION * * * *

EQUATION NUMBER 1.

DEPENDENT VARIABLE.. PROD Crop production

BEGINNING BLOCK NUMBER 1. METHOD: ENTER AREA IMSEED FERT CRDAM AGEH SIZEH

VARIABLE(S) ENTERED ON STEP NUMBER

1..	SIZEH	Size of House Hold
2..	AGEH	Age of Holder
3..	FERT	Fertilizer Applied
4..	IMSEED	Improved Seed
5..	AREA	Total Area Under Crop
6..	CRDAM	Observed Crop Damage

MULTIPLE R	.70799
R SQUARE	.50125
ADJUSTED R SQUARE	.48395
STANDARD ERROR	2.78655

ANALYSIS OF VARIANCE

	DF	SUM OF SQUARES	MEAN SQUARE
REGRESSION	6	1350.06065	225.01011
RESIDUAL	173	1343.32433	7.76488

F = 28.97792 SIGNIF F = .0000

---------- VARIABLES IN THE EQUATION ----------

VARIABLE	B	SE B	BETA	T	SIG T
SIZEH	.10633	.09148	.06298	1.162	.2467
AGEH	-.05449	.27942	-.01059	-.195	.8456
FERT	.62528	.57202	.05943	1.093	.2759
IMSEED	1.49038	.65548	.12863	2.274	.0242
AREA	8.13629	.66945	.66438	12.154	.0000
CRDAM	-.62737	.51196	-.06916	-1.225	.2221
(CONSTANT)	.10900	.79284		.137	.8908

Examples:

1. Based on the numerical results of estimation of the model given in the original computer printout reproduced on the previous page, let's perform a statistical inference for that model(assuming normality of random disturbance distribution).

The sample size (cross-sectional type of empirical data) was n=180(of agricultural households).

The coefficient of Determination R^2 for the estimated linear model was calculated as $R^2 = 0.501$, which means that 50.1% of the total empirical dispersion of wheat production in the sampled agricultural households can be explained by the regression on given 6 explanatory variables.

The Adjusted R^2 is $\bar{R}^2 = 0.48395$.

For the test of significance of R^2 , we obtained the value of F-Snedecor statistic as $F = \frac{R^2}{1-R^2} \cdot \frac{n-k-1}{k} = 28.98$, while the critical value $F_\alpha = 2.15$ for the significance level $\alpha = 0.05$ and for 6 and 173 degrees of freedom. Since $F > F_\alpha$, the R^2 is statistically significant.

In the reproduced computer printout we have not only the residual variance $S^2 = 7.76$, $S=2.786$ (printed as STANDARD ERROR), estimates b_i(printed as B), their standard errors S_{b_i} (printed as SEB), but also t-Student's statistics (T).

Let's rewrite results of estimation of the model in the form:

$$\hat{y} = \underset{(0.091)}{0.106\, x_1} - \underset{(0.279)}{0.054\, x_2} + \underset{(0.572)}{0.625\, x_3} + \underset{(0.655)}{1.490\, x_4} + \underset{(0.669)}{8.136\, x_5} \underset{(0.512)}{-0.627 x_6} + 0.109$$

$$\underset{}{1.16} \qquad \underset{}{-0.195} \qquad \underset{}{1.09} \qquad \underset{}{2.27} \qquad \underset{}{12.15} \qquad \underset{}{-1.22}$$

where x_1 is the first explanatory variable(size of household) etc, figures in brackets are the standard errors S_{b_i} and in the last (lower) row there are values of t-Student's statistics $t_i = \frac{b_i}{S_{b_i}}$

Assuming the significance level $\alpha = 0.05$, we have the critical value $t_\alpha = 1.96$ for df = 173. Because we have obtained $|t| > t_\alpha$ only for two cases(for x_4, t=2.27 and for x_5, t=12.15) we can conclude, that the regression coefficients are statistically significant only for the variable x_4(Improved Seed) and for the variable x_5 (Area under wheat). Remaining four explanatory variables have non-significant regression coefficients.

Let's assume that we are also interested in obtaining interval estimate for the parameter β_5 for the explanatory variable x_5(Area under wheat).

From the printout we obtained estimate $b_5 = 8.136$ and its standard error $S_{b_5} = 0.669$. For the confidence coefficient $1-\alpha = 0.95$ and for df = 173, we have (from Statistical Tables) $t_\alpha = 1.96$.

So, we obtain the following confidence interval for the parameter β_5 (marginal wheat yield in q/ha):

$$8.136 - 1.96 \cdot 0.669 < \beta_5 < 8.136 + 1.96 \cdot 0.669$$

i.e. $6.8 < \beta_5 < 9.4$.

Based on numerical results given in the Example 2 from previous Chapter 8 (page 96) for the linear model of wheat production in Arsi region of Ethiopia, let's perform a statistical inference for that model (assuming normality of random disturbances).

For the variable x_1 Area under wheat its parameter β_1 has been estimated as $b_1 = 11.70$ with the standard error $S_{b_1} = 1.14$

So, the value of Student's statistic $t_1 = \frac{11.70}{1.14} = 10.26$, while the critical value $t_\alpha = 1.98$ for the critical level $\alpha = 0.05$ and for the appropriate df = 120. Since $|t_1| = 10.26 > t_\alpha$ we may conclude that the estimate b_1 is statistically significant.

Similarly for the remaining three variables we have:

- for x_2(use of improved seed), $t_2 = \frac{2.82}{1.30} = 2.17$
- for x_3(fertilizer applied), $t_3 = \frac{1.62}{0.94} = 1.72$

- for x_4(crop damage observed), $t_4 = \frac{-3.39}{0.97} = -3.49$

So, the variables x_2 and x_4 have statistically significant parameters, while the variable x_3 has its parameter statically not significant. We can add, that for that model, $R^2 = 0.527$ and $F = 21.57$, which means, that R^2 is statistically significant.

3. Based on numerical results from the Example 1 from previous Chapter 8 (page 95-96) for the estimated linear model of the monthly expenditures for non-food commodities in households, let's perform a statistical inference for that model (assuming normality of random disturbances).

The model was estimated as $\hat{y} = \underset{(0.048)}{0.453x_1} - \underset{(0.091)}{0.267x_2} + 0.053$

where x_1 - household's income, x_2 - expenditure on food commodities. The sample size was n=5, and let's take the significance level $\alpha = 0.10$, so that the critical value for df= 2 is $t_\alpha = 2.92$. Empirical values of the Student's - t statistic are respectively:

$$t_1 = \frac{0.453}{0.048} = 9.44 \quad \text{and} \quad t_2 = \frac{-0.267}{0.091} = -2.93$$

Since $|t| > t_\alpha$ for both cases, we may conclude that both explanatory variables in the model have statistically significant parameters.

We can also calculate the coefficient of Determination R^2 as: $R^2 = 1 - \frac{0.0044}{1.048} = 0.998$ obviously it is statistically significant, since $F = 249.0$ and $F_\alpha = 19.0$ for $\alpha = 0.05$ and df = 2 and 2.

In addition let us assume that we are interested in obtaining the forecast $\hat{y}_T$ of the monthly expenditures for non-food commodities y_T assuming the following vector of future values of explanatory variables $X^T_{(T)} = (7\ 4\ 1)$. Using formulas from paragraph 9.4

and the matrix $(X^TX)^{-1}$ from page 95, we obtain the forecast

$\hat{y}_T = (7\ 4\ 1) \cdot \begin{pmatrix} 0.453 \\ -0.267 \\ 0.053 \end{pmatrix} = 2.16$, with its standard error estimated as:

$$S_{\hat{y}_T} = 0.05\sqrt{1 + 1.4} \approx 0.07$$

It means, that if in any household its monthly income increases up to 700 Birr and its expenditure for food commodities increases up to 400 Birr we can predict the monthly expenditure for non-food commodities of that household as 216 Birr (with standard error of this forecast as 7 Birr).

CHAPTER 10. SOME SPECIAL PROBLEMS.

10.1. Changing the Scale.

Sometimes it is convenient to change the original scale(units) of variables included in the model into new one. For example a production of crop measured in quintals should be now measured in metric tons, etc. Let's denote scaling factors by:

p - for the dependent variable Y

p_i - for the i-th explanatory variable X_i

Usually p and p_i are equal to the following numbers: 0.1 0.01 0.001 10,100 etc.

After changing the scale we can denote the empirical data as:

$$\begin{cases} y_t^* = p\, y_t \\ x_{ti}^* = p_i\, x_{ti} \quad \text{for } i = 1,2,\ldots, k+1 \end{cases}$$

or in matrix notation $\mathbb{y}^* = \mathbb{y} \cdot p$

$$\mathbb{X}^* = \mathbb{X} \cdot \mathbb{P}$$

where $\mathbb{P}$ is the diagonal, symmetric matrix defined as

$$\mathbb{P} = \begin{pmatrix} p_1 & 0 & \ldots & 0 \\ 0 & p_2 & \ldots & 0 \\ \vdots & & & \vdots \\ 0 & 0 & \ldots & p_{k+1} \end{pmatrix}$$

and its inverse matrix is

$$\mathbb{P}^{-1} = \begin{pmatrix} \frac{1}{p_1} & 0 & \ldots & 0 \\ 0 & \frac{1}{p_2} & \ldots & 0 \\ \vdots & & & \vdots \\ 0 & 0 & \ldots & \frac{1}{p_{k+1}} \end{pmatrix}$$

Let's assume, that the model was estimated based on new scale data, i.e. $\mathbb{y}^*$ and $\mathbb{X}^*$, so we have the OLS estimate $\mathbb{b}^* = (\mathbb{X}^{*T}\mathbb{X}^*)^{-1}\mathbb{X}^{*T}\mathbb{y}^*$ with the covariance matrix estimated as $\hat{\mathbb{V}}(\mathbb{b}^*) = s_*^2 (\mathbb{X}^{*T}\mathbb{X}^*)^{-1}$.

For comparison purposes we are interested however to come back to the prior(original) scale, i.e. we would like to obtain $\mathbb{b}$ from given $\mathbb{b}^*$. The connection between these two estimates are as follows:

$$\mathbb{b}^* = (\mathbb{X}^{*T}\mathbb{X}^*)^{-1}\mathbb{X}^{*T}\mathbb{y}^* = (\mathbb{P}\mathbb{X}^T\mathbb{X}\mathbb{P})^{-1}\mathbb{P}\mathbb{X}^T\mathbb{y}\cdot p =$$

$$= p\,\mathbb{P}^{-1}(\mathbb{X}^T\mathbb{X})^{-1}\underbrace{\mathbb{P}^{-1}\mathbb{P}}_{\mathbb{I}}\mathbb{X}^T\mathbb{y} = p\,\mathbb{P}^{-1}\mathbb{b}$$

It means, that for each element of the vector b^* we have:

$$b_i^* = p \cdot \frac{1}{p_i} \cdot b_i \quad \text{or} \quad b_i = \frac{1}{p} \cdot p_i \cdot b_i^* \quad \text{for } i=1,2,\ldots,k+1.$$

In the same way we obtain for the standard errors of estimates:

$$S_{b_i^*} = p \cdot \frac{1}{p_i} \cdot S_{b_i} \quad \text{or} \quad S_{b_i} = \frac{1}{p} \cdot p_i \cdot S_{b_i^*}$$

But changing the scale will not have any effect on t-Student's statistic:

$$t = \frac{b_i^*}{S_{b_i^*}} = \frac{p \cdot \frac{1}{p_i} b_i}{p \cdot \frac{1}{p_i} S_{b_i}} = \frac{b_i}{S_{b_i}} \quad \text{as for the original scale.}$$

Values of the t-Student's statistic have no "names", since the common "name" of b_i and S_{b_i} can be cancelled.

It is useful sometimes to have similar property for the regression coefficient. For that purpose so called "<u>standardized regression coefficient Beta</u>" has been defined as:

$$\text{Beta}_i = b_i \cdot \frac{S_{x_i}}{S_y}$$

where b_i is the estimate of the regression coefficient β_i

S_{x_i} is the standard deviation of the explanatory variable X_i

S_y is the standard deviation of the explained variable Y.

Let's denote the name of given dependent variable Y by $[Y]$, and the name of the explanatory variable X_i by $[X_i]$.
So, "the name" of b_i is obviously $\frac{[Y]}{[X_i]}$, because $[Y] = [b_i] \cdot [x_i]$

"The name" of S_{x_i} is $[X_i]$, and "the name" of S_y is $[Y]$.

Therefore the standardized regression coefficient "Beta_i" has no "name", since :

$$\text{Beta}_i = \frac{[Y]}{[X_i]} \cdot \frac{[X_i]}{[Y]} = 1$$

Since the standardized regression coefficient Beta has no "name" it can be used as a measure of importance of explanatory variables i.e. as a measure of the share of independent variables in a value of the dependent variable.

We can not compare "b_i", but we can compare "Beta i" .

Example:

In the computer outprint reproduced on page 111 in paragraph 9.5 for regression analysis of the model for wheat production in Shoa region there is one column entitled "Beta". Elements of that column are the standardized regression coefficients for all explanatory variables.

So, we have: $\text{Beta}_1 = -0.063$, $\text{Beta}_2 = -0.011$, $\text{Beta}_3 = 0.059$, $\text{Beta}_4 = 0.129$, $\text{Beta}_5 = 0.664$, $\text{Beta}_6 = -0.069$.

It follows, that the most important variable is X_5(area under wheat), then X_4(improved seed), then X_6(crop damage), etc.

10.2. Multicollinearity Problem.

Sometimes it hapens, that for the estimated multiple regression model its determination coefficient R^2 is very high, while all regression coefficients are statistically non-significant. One of the main reasons for such rather unpleasant situation is existence of so called Multicollinearity between explanatory variables. To explain it, let's comeback to the Assumption 2 of our model: $r(X) = k+1$. That assumption means, that the matrix X of observations on explanatory variables has full rank, i.e. all its column vectors are linearly independent. This is essential for obtaining the inverse matrix $(X^TX)^{-1}$, recalling that for Gram'matrix $r(X^TX) = r(X)$. If rank $r(X) < k+1$, than rank $r(X^TX) < k+1$ and $\det(X^TX) = 0$, which means that the Gram' matrix X^TX is singular (its inverse matrix $(X^TX)^{-1}$ does not exist) and the OLS method can not be applied.

Two examples of matrices with non full rank are:

$$X = \begin{pmatrix} 10 & 5 & 1 \\ 8 & 4 & 1 \\ 6 & 3 & 1 \\ 4 & 2 & 1 \end{pmatrix}$$, where for the first two columns we have $x_1 = 2x_2$,

and

$$X = \begin{pmatrix} 2 & 1 & 2 & 1 \\ 4 & 3 & 5 & 1 \\ 6 & 2 & 5 & 1 \\ 10 & 4 & 9 & 1 \end{pmatrix}$$, where for the first three columns we have $\frac{1}{2} x_1 + x_2 = x_3$.

A full linear dependency between columns of the observations matrix X like in above examples, is very rare in an econometric practice, but often we are facing with high correlation between explanatory variables i.e. approximate linear dependency between columns of matrix X, like in the following example:

$$X = \begin{pmatrix} 2 & 1 & 1.8 & 1 \\ 4 & 3 & 5.1 & 1 \\ 6 & 2 & 4.9 & 1 \\ 10 & 4 & 9.2 & 1 \end{pmatrix}$$

where approximately we have $\frac{1}{2} x_1 + x_2 \approx x_3$

Definition.

An approximate linear dependency between columns of the observations matrix X (reflecting high correlation between explanatory variables) is called a multicollinearity.

Existence of multicollinearity has an important negative effect on results of estimation of econometric models based on the multiple linear regression, because of $\det(X^T X) \approx 0$ for a multicollinearity case. In consequence, we obtain very large diagonal elements d_{ii} of the inverse matrix $(X^T X)^{-1} = \frac{(X^T X)^D}{\det(X^T X)}$, where $(X^T X)^D$ is the adjoint matrix.

Then, estimated standard errors $S_{b_i} = S\sqrt{d_{ii}}$ are to large, and in consequence values of t-Student's statistic $t = \frac{b_i}{S_{b_i}}$ are to small for b_i to be statistically significant.

But a multicollinearity does not influence the value of determination coefficient R^2 which may by high, while all regression coefficients in the model are statistically non-significant, due to multicollinearity. To cope with the multicollinearity problem first of all we have to check its significance by means of special significance test for multicollinearity.

The most often used <u>significance test for multicollinearity</u> is known in literature as <u>the Farrar-Glauber-Haitovski</u> test based on some Chi-square statistic obtained from the determinant of the explanatory variables correlation matrix

$$\mathbb{R} = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1k} \\ r_{21} & 1 & \cdots & r_{2k} \\ \vdots & & & \vdots \\ r_{k1} & r_{k2} & \cdots & 1 \end{pmatrix},$$

where r_{ij} is the correlation coefficient between x_i and x_j

In the Farrar-Glauber-Haitovski test we are testing the hypothesis H_o : $\det \mathbb{R} = 0$, against H_1 : $\det \mathbb{R} > 0$.

Because $\det \mathbb{R} = 0$ only if $\det (X^T X) = 0$, it follows, that rejection of the null hypothesis H_o, means that multicollinearity is statistically non-significant.

The χ^2 statistic used in the test is defined as:

$$\chi^2 = -\left[n-1-\frac{1}{6}(2k+5)\right] \cdot \ln (1-\det \mathbb{R}),$$

and it has Chi-square distribution (under normality assumption 4) with $df = \frac{1}{2} k(k-1)$, where k is the dimension of the correlation matrix $\mathbb{R}$, and n is the sample size.

The critical region is constructed as:

$Q = \{\chi^2 : \chi^2 \geq \chi^2_\alpha\}$, where χ^2_α is the critical value taken from Statistical Tables of χ^2 distribution for $df = \frac{1}{2} k(k-1)$ and for given significance level α in such a way, that $P\{\chi^2 \geq \chi^2_\alpha\} = \alpha$.

So, based on the determinant of correlation matrix $\mathbb{R}$, an empirical value of χ^2 statistic should be computed and then compared with the critical region Q.

If $\chi^2 \in Q$, the hypothesis H_o : $\det \mathbb{R} = 0$ is rejected (multicollinearity is statistically non-significant).

If $\chi^2 \notin Q$ we can not reject the hypothesis H_o about multicollinearity. In such a case multicollinearity should be considered as significant and some action should be undertaken for improving the estimation results spoiled by the multicollinearity.

Generally speaking an econometrician can apply one out of two different approaches in case when he discovered that multicollinearity is significant in the estimated model.

(i) If the model is big, one can remove some variables involved in multicollinearity, and estimate again the new, smaller model.

(ii) If the model is small, one shall find some additional (prior) information on the estimated parameters vector β and then apply one out of non-classical methods of estimation, like Non-linear Programming Technique or Baysian Technique of estimation described in Chapter 14.

Example:

During estimation procedure (based on sample size n=12) applied to some linear model with k=3 explanatory variables the following correlation matrix has been obtained:

$$\mathbb{R} = \begin{pmatrix} 1 & 0.8 & 0.2 \\ 0.8 & 1 & 0.4 \\ 0.2 & 0.4 & 1 \end{pmatrix}$$

For checking a significance of multicollinearity, the Farrar-Glauber-Haitovski test is applied.

So, we compute $\det \mathbb{R} = 0.864$, $1-\det \mathbb{R} = 0.136$,

$\ln 0.136 = -1.995$

The empirical value of Chi-square statistic is calculated as:

$$\chi^2 = -\left[11 - \frac{1}{6}(6+5)\right] \cdot (-1.995) = 9.17 \cdot 1.995 = 18.29$$

For the significance level $\alpha = 0.05$ and for $df = \frac{1}{2} \cdot 3 \cdot 2 = 3$ the critical value is $\chi^2_\alpha = 7.815$, so the critical region is defined as $Q = \{\chi^2 : \chi^2 \geq 7.815\}$

Since the empirical value $\chi^2 = 18.29 \in Q$, we have to reject the hypothesis (H_o : $\det \mathbb{R} = 0$) about multicollinearity and we can conclude that for the estimated model a multicollinearity is not significant.

10.3. Testing the Normality of Residuals.

All significance tests applied for checking different aspects of the estimated model(like t-test , F-test) are based on Assumption 4, that the disturbance term $\mathcal{E}$ has a normal distribution.
Therefore it is very important to remember that the analysis of residuals for the estimated model should always start with performing a significance test for the statistical hypothesis, that residuals $\hat{\varepsilon}_t$ (which can be treaten as estimates of random disturbances ε_t) have the normal distribution $N(0, \sigma^2)$. If that normality hypothesis has to be rejected, we can not continue analysis of the model, but we shall rebuild (change) it in such a way, that finally normality assumption can be proved.
Formally, the hypothesis on normality which is under testing can be written as $H_o : F(\hat{\varepsilon}_t) = F_N$, where $\hat{\varepsilon}_t$ stands for residuals in the estimated model, and F_N stands for distribution function of the normal distribution $N(0, \sigma^2)$.
To test the null hypothesis H_o we can apply one out of few tests of goodness of fit,known in Mathematical Statistics.

The Chi-square test.

The well known Chi-square test of goodness of fit can be applied as the normality test for rather large sample size ($n > 50$).
It consists of several activities ordered in the following sequence:

1° Individual residuals $\hat{\varepsilon}_t$ have to be classified into r classes (intervals) for obtaining an empirical distribution.

2° For each i-th class C_i the probability $p_i = P\{\hat{\varepsilon}_t \in C_i\}$ should be calculated from the normal distribution $N(0, \sigma^2)$. Instead of unknown variance σ^2 , the residual variance S^2 should be used.

3° For each i-th class its theoretical numbers np_i should be obtained in order to compare with empirical numbers n_i.

4° Value of the Chi-square statistic should be computed as:

$$\chi^2 = \sum_{i=1}^{r} \frac{(n_i - np_i)^2}{np_i}$$

and compared with the critical value χ^2_α (from Statistical Tables for df = r-2 and for given significance level α .

5° If $\chi^2 \geq \chi^2_\alpha$, the null hypothesis H_o is rejected, which means that residuals have different than normal distribution, and the model have to be changed.

If $\chi^2 < \chi^2_\alpha$ we cannot reject hypothesis H_o about normality distribution of residuals.

A computer program for the Chi-square test is available in SPSS package, as well as a histogram plot which facilitates obtaining the empirical distribution of residuals i.e. numbers n_i.

The Kolmogorov' test.

For rather small sample size, the Kolmogorov' test of goodness of fit can be applied as a normality test for residuals.

The null hypothesis is as before $H_o : F(\hat{\varepsilon}_t) = F_N$.

The sequence of work in this test is as follows:

1° Residuals $\hat{\varepsilon}_t$ should be ordered into non-decreasing sequence.

2° For each residual $\hat{\varepsilon}_t$ the value of empirical distribution function should be obtained as $F_n(\hat{\varepsilon}_t) = \frac{n_c}{n}$, where n_c is the cumulated number of residuals up to $\hat{\varepsilon}_t$.

3° For each residual $\hat{\varepsilon}_t$ the value of theoretical (normal) distribution function $F(\hat{\varepsilon}_t) = F(u_t)$ where $u_t = \frac{\hat{\varepsilon}_t}{S}$ should be obtained.

4° Value of the Kolmogorov' D-statistic should be taken as $D = \max \left| F_n(\hat{\varepsilon}_t) - F(\hat{\varepsilon}_t) \right|$ and compared with the critical value D_α for given n and significance level α .

5° If $D \geqslant D_\alpha$, the null hypothesis H_o about normality is rejected (which means that residuals have different, but not normal distribution).

If $D < D_\alpha$, we can not reject the hypothesis H_o about normality distribution of residuals.

Because Statistical Tables for getting oritical values D_α of the Kolmogorov' test are not so often printed, some critical values D_α for the significance level $\alpha = 0.05$ are given below:

n	5	6	7	8	9	10	11	12	13	14
D_α	0.563	0.519	0.483	0.454	0.430	0.409	0.391	0.375	0.361	0.349
n	15	17	19	20	25	30	35	40	45	50
D_α	0.338	0.318	0.301	0.294	0.264	0.242	0.224	0.210	0.198	0.188

The Hellwig' test based on empty cells.

For small sample size the Hellwig' test of normality can be also applied. It is very simple test based on empty cells and on the following Theorem known from Probability Theory:

If a random variable X has the distribution function F(x), then the transformed random variable $Y = F(X)$ has the uniform distribution within $[0, 1]$ interval.

The idea of the Hellwig' test is therefore based on transformation using normal distribution function F(u) where $u \sim N(0,1)$, and then checking if values of those transformed values have uniform distribution within $[0,1]$ interval which is split into sub-intervals, called cells.

If the original distribution is really normal, it should be not to many empty cells(i.e. cells without any observation). So, the number of empty cells can be used to construct the critical region for the null hypothesis $H_o : F(\hat{\varepsilon}_t) = F_N$

The sequence of work in the Hellwig' test of normality is as follows:

1° Residuals $\hat{\varepsilon}_t$ should be ordered into non-decreasing sequence.

2° Residuals are standardized into $u_t = \frac{\hat{\varepsilon}_t}{S}$ and from Statistical Tables values $F(u_t)$ of normal distribution function should be obtained.

3° Unit interval $[0, 1]$ is divided into n sub-intervals I_j (cells) of equal length $\frac{1}{n}$, $I_j = \left[\frac{j-1}{n}, \frac{j}{n}\right)$ for $j=1,2,\ldots, n$.
So each value $F(\hat{\varepsilon}_t) = F(u_t)$ belongs to some cell I_j.

4° The statistic h_o defined as the number of empty cells is observed. It has discrete probability distribution from which such critical value h_α can be obtained, that $P\{h_o \geq h_\alpha\} = \alpha$ for given sample size n and given significance level α.

5° If $h_o \geq h_\alpha$ the hypothesis about normality has to be rejected.
If $h_o < h_\alpha$ we cannot reject the hypothesis about normality distribution of residuals.

Critical values h_α of the Hellwig' test for the significance level $\alpha = 0.05$ and for given sample size n are as follows:

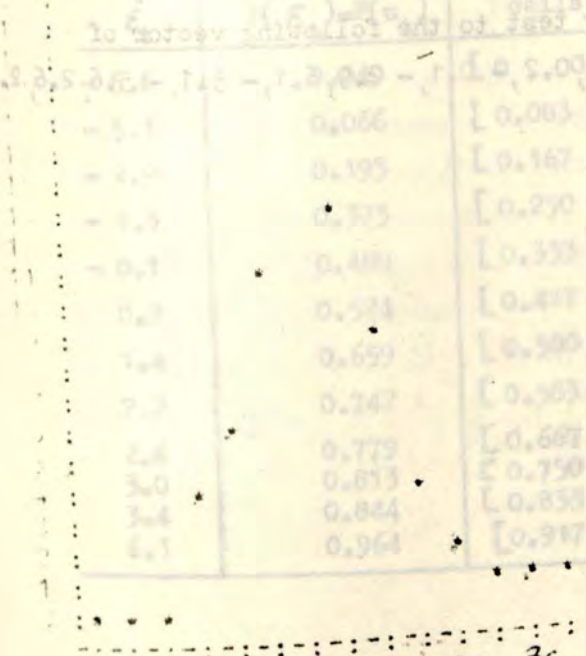
n	5	6	7	8	9	10	11	12	13	14	15	17	20
h_α	2	3	3	4	4	5	5	6	6	6	7	8	9

Examples:

1. After estimation of some linear model by the OLS method we obtained the following computer histogram plot of residuals using ISEA package:

RESIDUALS : HISTOGRAM (ISEA)

LOWEST INTERVAL = -14.720 , INTERVALS = $\frac{1}{2}S = 2.45340$, $S = 4.9$



i	Interval	n_i
1	$[-2s, -s)$	10
2	$[-s, 0)$	31
3	$[0, s)$	22
4	$[s, 2s)$	6
5	$[2s, +\infty)$	3
		72 = n

9

Based on the histogram we were able to construct the empirical distribution (given above on the right-hand side) of residuals with the length of intervals equal to S=4.9

Let's apply now the Chi-square test to check if the normal distribution $N(0, S^2)$ fits to that empirical distribution of residuals.

It is convenient to carryout all calculations in the following working table (recalling the standardization $u_t = \frac{\hat{\varepsilon}_t}{S}$):

i	$\hat{\varepsilon}_t$	n_i	u_t	$F(u_t)$	p_i	n_{pi}	$(n_i - np_i)^2 / np_i$
1	(-9.8, -4.9)	10	-1.0	0.159	0.159	11.5	0.20
2	(-4.9, 0)	31	0	0.500	0.341	24.5	1.72
3	(0, 4.9)	22	1.0	0.841	0.341	24.5	0.26
4	(4.9 , 9.8)	9	-	-	0.159	11.5	0.54
5	(9.8 , $+\infty$)						
		72				72	2.72= χ^2

So, the empirical value of Chi-square statistic is $\chi^2 = 2.72$

For df = r-2 = 4-2 = 2 and for the significance level $\alpha = 0.05$ we have the critical value $\chi^2_\alpha = 5.991$.

Since $\chi^2 = 2.72 < \chi^2_\alpha$, we can not reject the null hypothesis about normality distribution of residuals obtained from the estimated model.

2. Let's apply the Kolmogorov' normality test to the following vector of residuals $\hat{\varepsilon}^T = (1.4, 3.4, 3.0, -1.1, 0.2, -0.1, -2.9, 6.1, -5.1, -5.6, 2.6, 2.2)$ with standard deviation S= 3.38.

The working table is as follows:

$\hat{\varepsilon}_t$	n_c	$F_n(\hat{\varepsilon}_t)$	u_t	$F(\hat{\varepsilon}_t)$	$\lvert F_n(\hat{\varepsilon}_t) - F(\hat{\varepsilon}_t)\rvert$
-5.6	1	0.083	-1.65	0.049	0.034
-5.1	2	0.167	-1.51	0.066	0.101
-2.9	3	0.250	-0.86	0.195	0.055
-1.1	4	0.333	-0.32	0.375	0.042
-0.1	5	0.417	-0.03	0.488	0.071
0.2	6	0.500	0.06	0.524	0.024
1.4	7	0.583	0.41	0.659	0.076
2.2	8	0.667	0.65	0.742	0.075
2.6	9	0.750	0.77	0.779	0.029
3.0	10	0.833	0.89	0.813	0.020
3.4	11	0.917	1.01	0.844	0.073
6.1	12	1.000	1.80	0.964	0.036

So, we obtained the value of the D-Kolmogorov' statistic equal to D=0.101
For n=12 and for the significance level α=0.05 the critical value of the Kolmogorov' statistic is $D_\alpha = 0.375$.
Since $D = 0.101 < D_\alpha$, we can not reject the null hypothesis about normality distribution of residuals.

3. Based on the residuals vector $\hat{\varepsilon}^T$ given in the previous Example 2 let's apply the Hellwig' test of normality.
The working table is as follows(m_j stands for number of observations within j-th cell I_j):

$\hat{\varepsilon}_t$	$F(\hat{\varepsilon}_t)=F(u_t)$	cells I_j	m_j
- 5.6	0.049	[0 , 0.083)	2
- 5.1	0.066	[0.083 , 0.167)	0
- 2.9	0.195	[0.167 , 0.250)	1
- 1.1	0.375	[0.250 , 0.333)	0
- 0.1	0.488	[0.333 , 0.417)	1
0.2	0.524	[0.417 , 0.500)	1
1.4	0.659	[0.500 , 0.583)	1
2.2	0.742	[0.583 , 0.667)	1
2.6	0.779	[0.667 , 0.750)	1
3.0	0.813	[0.750 , 0.833)	2
3.4	0.844	[0.833 , 0.917)	1
6.1	0.964	[0.917 , 1]	1

So, we obtained the number of empty cells $h_o = 2$.

For n=12 and for the significance level $\alpha = 0.05$, the critical value in the Hellwig test is $h_\alpha = 6$.

Since $h_o = 2 < h_\alpha$, we can not reject the null hypothesis about normality distribution of given residuals.

CHAPTER 11. USE OF THE GENERALIZED LEAST SQUARES METHOD FOR MODELS WITH HETEROSCEDASTICITY.

11.1. The Generalized Least Squares Method.

Up to here, we discussed models with so-called classic assumptions which included assumption 3 about the simple diagonal covariance matrix of the random disturbance term $\mathcal{E}$, i.e. $V(\mathcal{E}) = \sigma^2 I$.

Sometimes that assumption is rejected during the stage of verification of the model, and other, more general form of the covariance matrix $V(\mathcal{E})$ should be accepted. In consequence, instead of the Ordinary Least Squares (OLS) method, other so-called Generalized Least Squares method of estimation should be applied.

The Generalized Least Squares (GLS) method was proposed by Aitken in 1935 for estimation of linear models with the general covariance matrix of the random disturbance term $\mathcal{E}$, i.e. $V(\mathcal{E}) = \sigma^2 \Omega$, where Ω is nxn symmetric, positive definite matrix.

So now, we have to change assumption 3 into more general assumption 3':

ASS.1. $y = X\beta + \mathcal{E}$

ASS.2. Matrix X is nonrandom and its rank $r(X) = k+1 \leq n$

ASS.3'. $E(\mathcal{E}) = 0$ and covariance matrix $V(\mathcal{E}) = \sigma^2 \Omega$.

There are two important consequencies of that more general assumption 3':

(i) The OLS estimator b is now not the BLUE (i.e. it is not the most efficient estimator) of the parameters vector β.

(ii) The residual variance S^2 is now not an unbiased estimator of σ^2.

In order to obtain the best linear estimator of the parameters vector β of the model with the general assumption 3' we shall apply Aitken's Generalized Least Squared (GLS) method based on minimization of the general quadratic form of the random disturbance vector $\mathcal{E}$, i.e.

(min) $\tilde{S} = \mathcal{E}^T \Omega^{-1} \mathcal{E}$ instead of (min) $S = \mathcal{E}^T \mathcal{E}$, which is the criterion in the OLS method.

The criterion function $\tilde{S}$ used in the GLS method can be written as:

$$\tilde{S} = (y - X\beta)^T \Omega^{-1} (y - X\beta) = y^T \Omega^{-1} y - 2 y^T \Omega^{-1} X \beta + \beta^T X^T \Omega^{-1} X \beta.$$

By application of the necessary condition for minimum(with respect to β) of the function $\tilde{S}$, we obtain the following equation:

$$\frac{\partial \tilde{S}}{\partial \beta} \equiv -2X^T\Omega^{-1}y + 2X^T\Omega^{-1}X\beta = 0$$

or $$X^T\Omega^{-1}X\beta = X^T\Omega^{-1}y$$

The solution of that equation will be denoted by $\tilde{b}$ and it is known as the GLS estimator defined as $\tilde{b} = (X^T\Omega^{-1}X)^{-1}X^T\Omega^{-1}y$.
Obviously the OLS estimator $b = (X^TX)^{-1}X^Ty$ is the special case of the GLS estimator $\tilde{b}$ for $\Omega = I$.
It is easy to check that the GLS estimator $\tilde{b}$ is unbiased:

$$E(\tilde{b}) = (X^T\Omega^{-1}X)^{-1}X^T\Omega^{-1}\cdot E(y) =$$

$$= \underbrace{(X^T\Omega^{-1}X)^{-1}X^T\Omega^{-1}\cdot X}_{I}\beta = \beta.$$

By similar calculations which we applied for the OLS estimator, we can obtain the covariance matrix of the GLS estimator $\tilde{b}$ as

$$V(\tilde{b}) = \sigma^2(X^T\Omega^{-1}X)^{-1},$$

and the unbiased estimator of the unknown σ^2, as the generalized residuals variance $\tilde{s}^2 = \frac{\tilde{\varepsilon}^T\Omega^{-1}\tilde{\varepsilon}}{n-k-1} = \frac{y^T\Omega^{-1}y - \tilde{b}^TX^T\Omega^{-1}y}{n-k-1}$

where $\tilde{\varepsilon} = y - X\tilde{b}$

There are two equivalent ways for getting the GLS estimator $\tilde{b}$.
The first one is straightforward, i.e. having matrix Ω we have to compute its inverse matrix Ω^{-1}, and then apply the formula $\tilde{b} = (X^T\Omega^{-1}X)^{-1}X^T\Omega^{-1}y$.
The second way is to find such a transformation applied to the original model, that the new, transformed model has a disturbance term with classic assumption, i.e. with covariance matrix $V(\varepsilon) = \sigma^2 I$, and therefore the OLS method can be applied.

The transformation utilizes the appropriate matrix P, according to the following Theorem:

For each quadratic and positive definite matrix Ω exists such a matrix P, that $P\Omega P^T = I$, and $P^T P = \Omega^{-1}$.

Let's apply that theorem to our linear model $y = X\beta + \varepsilon$ with general assumption 3', that $E(\varepsilon) = 0$ and $V(\varepsilon) = \sigma^2\Omega$, where Ω is symmetric, positive definite matrix. So, exists such a matrix P, that $P\Omega P^T = I$ and $P^T P = \Omega^{-1}$.

Let's consider such a linear transformation of our model, which is just the left-hand side multiplication of the model by the matrix P:

$$Py = PX\beta + P\varepsilon$$

$$\text{or } y^* = X^*\beta + \varepsilon^*, \quad \text{where } Py = y^*, \; PX = X^*, \; P\varepsilon = \varepsilon^*$$

The disturbance term of the new model (with *) has the following parameters:

the mean vector $E(\varepsilon^*) = E(P\varepsilon) = P \cdot E(\varepsilon) = P \cdot 0 = 0$

and the covariance matrix $V(\varepsilon^*) = E[\varepsilon^* \cdot \varepsilon^{*T}] = E[P\varepsilon \cdot \varepsilon^T P^T] =$

$$= PV(\varepsilon)P^T = P\sigma^2\Omega P^T = \sigma^2 P\Omega P^T = \sigma^2 I.$$

Because of this classic covariance matrix $V(\varepsilon^*) = \sigma^2 I$, we can apply the OLS method for estimation β in the new model (*), i.e. we can use the criterion function

$$S^* = \varepsilon^{*T}\varepsilon^* = (P\varepsilon)^T(P\varepsilon) = \varepsilon^T P^T P \varepsilon = \varepsilon^T \Omega^{-1}\varepsilon = \tilde{S}$$

So, (min) S^* = (min) $\tilde{S}$, and therefore the OLS estimator b^* for the new model (*) is equal to the GLS estimator $\tilde{b}$ for the original model:

$$b^* = (X^{*T}X^*)^{-1}X^{*T}y^* = (X^T P^T P X)^{-1} X^T P^T P y =$$

$$= (X^T\Omega^{-1}X)^{-1}X^T\Omega^{-1}y = \tilde{b}$$

So, the second way for getting the GLS estimator $\tilde{b}$ is as follows: for given matrix Ω , we shall find the proper matrix P for the linear transformation of the original model into new model (with *) , for which the OLS estimator b^* is applied (and gives the same result as the GLS estimator $\tilde{b}$ computed directly).

In econometric practice the Generalized Least Squares (GLS) method is applied for two special cases of the matrix Ω connected with:

(i) the heteroscedasticity of random disturbances(usually for cross-sectional statistical data)

(ii) the autocorrelation of random disturbances(usually for time-series data)

The first case i.e. the heteroscedasticity is considered in next paragraphs of this chapter, while the autocorrelation will be discussed in the next Chapter 12.

11.2. Heteroscedasticity of Random Disturbances.

The random disturbances ε_t of the model $y = X\beta + \varepsilon$ are called heteroscedastic, if they have different variance for each t, i.e. $D^2(\varepsilon_t) = \sigma_t^2 = \sigma^2 \cdot k_t$.

The covariance matrix of a heteroscedastic random disturbance vector ε can be written as:

$$V(\varepsilon) = \sigma^2 \Omega, \quad \text{where} \quad \Omega = \begin{pmatrix} k_1 & 0 & \dots & 0 \\ 0 & k_2 & \dots & 0 \\ \vdots & & & \vdots \\ 0 & 0 & \dots & k_n \end{pmatrix}$$

It is clear that the model with heteroscedastic random disturbances should be estimated by means of the GLS method, which for this special case is called sometimes the Weighted Least Squares, because its criterion function is now:

$$(\min)\ \tilde{S} = \varepsilon^T \Omega^{-1} \varepsilon = \sum_{t=1}^{n} \frac{\varepsilon_t^2}{k_t}$$, where the weights $\frac{1}{k_t}$ are elements of the inverse matrix Ω^{-1} .

That means, that the value ε_t^2 is discounted by the factor k_t which is large for the large variance σ_t^2 .

Knowing K_t , i.e. for given Ω we can estimate the parameters vector β

of the model by applying the GLS estimator $\tilde{b}$ directly.
Alternatively, we can apply the OLS estimator b^* to the transformed model with the transformation matrix P defined as:

$$P = \begin{pmatrix} \frac{1}{\sqrt{k_1}} & 0 & \dots & 0 \\ 0 & \frac{1}{\sqrt{k_2}} & \dots & 0 \\ \vdots & & & \\ 0 & 0 & \dots & \frac{1}{\sqrt{k_n}} \end{pmatrix}, \text{ with the property } P^T P = \Omega^{-1}.$$

Let's consider the transformation $X^* = PX$, $y^* = Py$, $\varepsilon^* = P\varepsilon$
or for each individual observation:

$$y_t^* = \frac{y_t}{\sqrt{k_t}}, \quad x_{ti}^* = \frac{x_{ti}}{\sqrt{k_t}} \text{ for } i=1,2,\dots,k+1; \quad \varepsilon_t^* = \frac{\varepsilon_t}{\sqrt{k_t}}$$

After such a transformation the variance of new random disturbances will be constant: $D^2(\varepsilon_t^*) = D^2\left(\frac{\varepsilon_t}{\sqrt{k_t}}\right) = \frac{1}{k_t} D^2(\varepsilon_t) = \frac{\sigma^2 \cdot k_t}{k_t} = \sigma^2$.

which means the classic (homoscedastic) case, for which the OLS estimator b^* can be applied.
However still the main question is about the matrix Ω, i.e. how we can know its elements k_t.
We can consider a few typical situations which often hapen in econometric practice and for which it is easy to assume or to estimate coefficients k_t.
These main types of heteroscedasticity can be listed as:

(i) proportional to one out of explanatory variables,
(ii) proportional to the expected value of dependent variable Y,
(iii) grouped heteroscedasticity.

In the first type (i) of heteroscedastic random disturbances it is assumed that $D(\varepsilon_t) = \sigma \cdot x_t$, or $D^2(\varepsilon_t) = \sigma^2 \cdot x_t^2$, where x_t are observations on some appropriate explanatory variable X.

It follows, that $k_t = x_t^2$ and $\sqrt{k_t} = x_t$.

So, the transformation(*) of the original model with that type of heteroscedasticity is very simple: each individual t-th observation should be divided by $\sqrt{k_t} = x_t$.

A typical example of this type of heteroscedasticity is the linear savings function $y_t = \beta x_t + \alpha + \varepsilon_t$, where y_t stands for monthly savings of the t-th household, x_t stands for monthly income of the t-th household, β stands for the marginal propensity to savings, estimated based on cross-sectional data from a households income and expenditures sample survey.

It seems reasonable assumption for this model, that $D(\varepsilon_t) = \sigma \cdot x_t$ or $D^2(\varepsilon_t) = \sigma^2 \cdot x_t^2$, which means that for households with larger disposable income a larger dispersion of savings is observed (and therefore larger variance of random disturbances), while for households with small disposable income we may expect a small dispersion in savings (i.e. smaller variance $D^2(\varepsilon_t)$ due to restricted pattern of expenditures in poor households.

So, the given savings function should be estimated by the OLS method applied not to the original model, but to the new, transformed model:

$$\frac{y_t}{x_t} = \beta + \frac{1}{x_t}\alpha + \frac{\varepsilon_t}{x_t} \qquad \text{for } t = 1,2, \ldots, n$$

or

$$y_t^* = \alpha x_t^* + \beta + \varepsilon_t^*, \quad \text{where } y_t^* = \frac{y_t}{x_t}, \; x_t^* = \frac{1}{x_t}, \; \varepsilon_t^* = \frac{\varepsilon_t}{x_t}$$

because the new model (with *) has homoscedastic random disturbances i.e. with constant variance $D^2(\varepsilon_t^*) =$

$$= \frac{1}{x_t^2} D^2(\varepsilon_t) = \frac{\sigma^2 \cdot x_t^2}{x_t^2} = \sigma^2.$$

We shall notice here, that the marginal propensity to savings i.e. the parameter β became the constant term in the finally estimated (by the OLS) model.

In the second type (ii) of heteroscedastic random disturbances it is assumed that $D(\varepsilon_t) = \sigma \cdot E(y_t)$ or $D^2(\varepsilon_t) = \sigma^2 [Ey_t]^2$, where Ey_t is the expected value of observations on the explained variable.

It follows, that $k_t = [Ey_t]^2$, or $\sqrt{k_t} = Ey_t$, which can be estimated as $\hat{y}_t = b_1 x_{t1} + b_2 x_{t2} + \ldots + b_{k+1}$ (by the OLS method).

So, for this type of heteroscedasticity we have to deal with two stages of estimation:

Stage I - apply the OLS method for getting $\hat{y}_t$ as estimates of $\sqrt{k_t}$

Stage II - apply the GLS method for getting final estimate $\tilde{b}$ directly, or by the OLS estimate b^* applied to the transformed (*) model i.e. after dividing all observations by $\sqrt{k_t} = \hat{y}_t$.

That type of heteroscedasticity appears sometimes during estimation of the econometric production function $Q_t = f(K_t, L_t, \varepsilon_t)$ based on the cross-sectional type of data (e.g. agricultural holdings or enterprises).

Here Q_t stands for the production scale,
L_t stands for the labour,
K_t stands for the capital(fixed assets) engaged in production.

We may expect, that big enterprises differ more than small enterprises (with small production scale Q_t), therefore it is reasonable to assume heteroscedasticity as $D(\varepsilon_t) = \sigma \cdot E(Q_t)$, with $E(Q_t) = \sqrt{k_t}$.

In the third type (iii) of heteroscedasticity it is assumed, that the whole set of n observations can be split into r homogeneous groups (with homoscedasticity within each group), but with different variances (heteroscedasticity) between groups.

Such kind of the grouped heteroscedasticity is usually connected with an administrative or organizational division of the sampled data (e.g. groups may be taken as administrative regions of the Country).

For the grouped heteroscedasticity we have:

$D^2(\varepsilon_t) = \sigma^2 \cdot k_i$ if the t-th observation belongs to the i-th group $(i = 1,2, \ldots, r)$

The full covariance matrix of the random disturbance vector ε can be written as:

$$V(\varepsilon) = \sigma^2 \Omega, \text{ where } \Omega = \begin{pmatrix} k_1 & 0 & 0 & \cdots & 0 \\ 0 & \ddots & 0 & 0 & 0 \\ 0 & 0 & k_1 & 0 & 0 \\ \vdots & 0 & 0 & k_2 & \vdots \\ \vdots & 0 & 0 & 0 & \ddots \\ 0 & 0 & 0 & 0 & k_r \end{pmatrix}$$

The coefficients k_i for $i=1,2,\ldots,r$ can be estimated for each i-th group after checking by the Bartlett's test that the grouped heteroscedasticity is statistically significant. This is explained in the next paragraph.

11.3. Bartlett's Test for Grouped Heteroscedasticity.

Statistical significance of the grouped heteroscedasticity can be checked by means of the Bartlett's test assuming normal distribution of random disturbances.

The Bartlett's test was design to test the statistical hypothesis, that the variances of r normal distributions are equal.

Let us assume that we have r random samples drawn from r normal distributions $N(m_i, \sigma_i^2)$.

Sample observations are denoted by x_{ij} $(i = 1,2,\ldots, r;\ j = 1,2,\ldots,n_i)$ and sample averages are denoted by $\overline{x}_i$.

The null hypothesis which is under testing is about homoscedasticity;

$$H_o : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_r^2$$

against the alternative hypothesis H_1 : variances σ_i^2 are not equal (heteroscedasticity).

The Bartlett's test for the above null hypothesis is based on the Chi-square statistic defined as:

$$\chi^2 = \frac{1}{c}\left[(n-r)\ln S^2 - \sum_{i=1}^{r}(n_i - 1)\ln S_i^2\right.$$

where:

$$S_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2$$

$$S^2 = \frac{1}{n-r}\sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2, \quad n = \sum_{i=1}^{r} n_i$$

$$C = 1 + \frac{1}{3(r-1)}\left(\sum_{i=1}^{r}\frac{1}{n_i - 1} - \frac{1}{n-r}\right)$$

When the null hypothesis H_o about homoscedasticity is true, the statistic χ^2 has the Chi-square distribution with df = r-1 . The critical region is constructed as $Q = \{\chi^2 : \chi^2 \geq \chi^2_\alpha\}$, where χ^2_α is such a critical value, that for given significance level α and for df = r-1, the probability $P\{\chi^2 \geq \chi^2_\alpha\} = \alpha$. If the empirical value $\chi^2 \in Q$, the hypothesis H_o about homoscedasticity is rejected, which means that the heteroscedasticity between r populations is statistically significant.

If the empirical value $\chi^2 \notin Q$, i.e. $\chi^2 < \chi^2_\alpha$ we can not reject the null hypothesis H_o about homoscedasticity, which means that the heteroscedasticity is statistically non-significant.

The Bartlett's test can be applied now to residuals of the estimated multiple linear regression model for checking the significance of the grouped heteroscedasticity, assuming given r group of sample observations and the normality of residuals.

The sequence of work for testing the grouped heteroscedasticity in a given estimated linear model is as follows:

1° The model should be estimated by the OLS method i.e. estimator b is applied, and the list of individual residuals $\hat{\varepsilon}_t$ should be obtained.

2° The whole set of sample observations, and consequently set of residuals $\hat{\varepsilon}_t$ should be split into r groups according to the given criterion (e.g. administrative regions).

3° For obtained r samples with each sample having n_i residuals $\hat{\varepsilon}_t$, the Bartlett's test is applied.

4° If $\chi^2 < \chi^2_\alpha$ has been obtained, the grouped heteroscedasticity is statistically non-significant and the OLS estimates obtained in 1° are good.

5° If $\chi^2 \geqslant \chi^2_\alpha$ has been obtained, the grouped heteroscedasticity is statistically significant, and the GLS method should be used, with the elements K_i of the matrix Ω being estimated as follows:

$K_1 = 1$ for elements from the first group,

$K_2 = \frac{S_2^2}{S_1^2}$ for elements from the second group,

$\vdots$

$K_r = \frac{S_r^2}{S_1^2}$ for elements from the r-th group.

Obviously having estimated K_t in such a way, we can obtain the final estimate of the parameters vector β of the model, by applying the GLS estimator b directly, or indirectly by applying the OLS estimator b^* to the transformed model (with the transformation $*$, i.e. $\frac{y_t}{\sqrt{k_t}}$ etc).

Example:

In agricultural season 1981/82 the research experiment was performed in CSO Addis Ababa for studying a relationship between the monthly rainfall distribution and the crop yield. The cross sectional statistical data from the Farmers Associations were used in that experiment for estimation the linear model:

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 + \varepsilon$$

where y stands for crop yield,
x_1 stands for rainfall observed in June,
x_2 stands for rainfall observed in July,
x_3 stands for rainfall observed in August,
x_4 stands for rainfall observed in September,

For the basic cereal crop teff, the model estimated by the OLS method based on n=24 observations (FAs) in Ilubabor region gave rather large standard errors of estimates b_i and the determination coefficient $R^2 = 0.452$.

We may suspect here the significant grouped heteroscedasticity, since the rainfall data were taken from only 5 Meteorological Stations reporting systematically the rainfall for their nearest FAs.

Let's apply the Bartlett's test to residuals of that estimated model for checking the grouped heteroscedasticity.

The whole set of residuals $\hat{\varepsilon}_t$ was split into 5 groups (according to the appropriate MetStation) with size $n_1 = 4$, $n_2 = 10$, $n_3 = 3$, $n_4 = 2$, $n_5 = 5$.

For each group, S_i^2 was calculated, and we obtained:

$S_1^2 = 0.81$, $S_2^2 = 2.67$, $S_3^2 = 0.04$, $S_4^2 = 0.34$, $S_5^2 = 12.60$,

while $S^2 = 4.06$.

Next, we obtained $(n-r)\cdot \ln S^2 = (24-5)\cdot \ln 4.06 = 19 \cdot 1.40 = 26.60$,

$$\sum_{i=1}^{5} (n_i - 1) \ln S_i^2 = 10.79$$

and $c = 1 + \frac{1}{3.4}\left(\frac{1}{3} + \frac{1}{9} + \frac{1}{2} + 1 + \frac{1}{4} - \frac{1}{19}\right) = 1.178$, $\frac{1}{c} = 0.85$

So, the empirical value of Chi-square statistic is

$$\chi^2 = 0.85(26.60 - 10.79) = 13.44 ,$$

while for df = r-1 = 4, and for the significance level $\alpha = 0.05$ we have the critical value $\chi^2_\alpha = 9.49$.

Since $\chi^2 = 13.44 > \chi^2_\alpha$, the hypothesis $H_o : \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = \sigma_5^2$ has to be rejected, which means that the grouped heteroscedasticity is statistically significant and should be taken into account during estimation of the model.

For application of the GLS method, elements k_i (for i= 1,2,3,4,5) of the matrix Ω can be estimated now as: $k_1 = 1$, $k_2 = \frac{2.67}{0.81} = 3.3$,

$k_3 = \frac{0.04}{0.81} = 0.05$, $k_4 = \frac{0.34}{0.81} = 0.42$, $k_5 = \frac{12.60}{0.81} = 15.6$.

Using the appropriate transformation (i.e. dividing each observation by its appropriate $\sqrt{k_i}$) and then applying the OLS estimator b^* (equivalent to the GLS estimator $\tilde{b}$), we obtain the estimated model with much larger determination coefficient $R^2 = 0.996$ and with smaller than before standard errors (written below estimates):

$$\hat{y} = \underset{(0.014)}{0.026x_1} + \underset{(0.012)}{0.013x_2} - \underset{(0.005)}{0.003x_3} + \underset{(0.011)}{0.015x_4} - 3.58$$

CHAPTER 12. MODELS WITH AUTOCORRELATION OF RANDOM DISTURBANCES.

12.1. The Autoregressive Stochastic Process of Disturbances.

The second special case in which the Generalized Least Squares (GLS) method is applied for estimation of the model, is the case of autocorrelated disturbances (i.e. disturbances from the autoregressive stochastic process).

Let's consider the model with following assumptions:

ASS. 1. $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$

ASS. 2. $\mathbf{X}$ non-random, and $r(\mathbf{X}) = k+1 \leq n$

ASS. 3". The random variables u_t (i.e. elements of the random disturbance vector $\mathbf{u}$) form the stationary autoregressive first-order stochastic process defined as

$u_t = \rho u_{t-1} + \varepsilon_t$, where $|\rho| < 1$,

and ε_t is the stochastic process (called "the white noise") with classic assumptions:

$$\bigwedge_t \; E(\varepsilon_t) = 0, \quad D^2(\varepsilon_t) = \sigma_\varepsilon^2$$

$$\bigwedge_{t,s} \; COV(\varepsilon_t, \varepsilon_s) = 0$$

From Assumption 3" it follows, that

$$u_t = \rho u_{t-1} + \varepsilon_t = \rho(\rho u_{t-2} + \varepsilon_{t-1}) + \varepsilon_t = \rho^2 u_{t-2} + \rho\varepsilon_{t-1} + \varepsilon_t =$$

$$= \rho^2(\rho u_{t-3} + \varepsilon_{t-2}) + \rho\varepsilon_{t-1} + \varepsilon_t =$$

$$= \varepsilon_t + \rho\varepsilon_{t-1} + \rho^2\varepsilon_{t-2} + \dots \quad \text{(i.e. linear function of all } \varepsilon_t\text{)}$$

So, the expected value of our random disturbances will be

$$E(u_t) = E(\varepsilon_t + \rho\varepsilon_{t-1} + \rho^2\varepsilon_{t-2} + \dots) = 0, \text{ for each } t,$$

and the variance:

$$D^2(u_t) = E(u_t^2) = E\left[\varepsilon_t + \rho\varepsilon_{t-1} + \rho^2\varepsilon_{t-2} + \ldots\right]^2 =$$

$$= E(\varepsilon_t^2) + \rho^2 E(\varepsilon_{t-1}^2) + \rho^4 E(\varepsilon_{t-2}^2) + \ldots = \text{(since all covariances are zeros)}$$

$$= \sigma_\varepsilon^2 + \rho^2\sigma_\varepsilon^2 + \rho^4\cdot\sigma_\varepsilon^2 + \ldots =$$

$$= \sigma_\varepsilon^2 (1 + \rho^2 + \rho^4 + \ldots) =$$

$$= \sigma_\varepsilon^2 \cdot \frac{1}{1-\rho^2}$$

So, $D^2(u_t) = \sigma_u^2 = \dfrac{\sigma_\varepsilon^2}{1-\rho^2}$ for each t.

The covariance between two consecutive u_t will be equat to:

$$\operatorname{cov}(u_t, u_{t-1}) = E(u_t \cdot u_{t-1}) = E\left[(\varepsilon_t + \rho\varepsilon_{t-1} + \rho^2\varepsilon_{t-2} + \ldots)\cdot\right.$$

$$\left.\cdot(\varepsilon_{t-1} + \rho\varepsilon_{t-2} + \ldots)\right] =$$

$$= E\left\{\left[\varepsilon_t + \rho(\varepsilon_{t-1} + \rho\varepsilon_{t-2} + \ldots)\right]\left[\varepsilon_{t-1} + \rho\varepsilon_{t-2} + \ldots\right]\right\} =$$

$$= E\left\{\left[\varepsilon_t \cdot (\varepsilon_{t-1} + \rho\varepsilon_{t-2} + \ldots)\right]\right\} + \rho E\left[\varepsilon_{t-1} + \rho\varepsilon_{t-2} + \ldots\right]^2 =$$

$$= 0 + \rho E(u_{t-1})^2 = D^2(u_{t-1}) = \rho\sigma_u^2.$$

By a similar calculation we have

$$\operatorname{cov}(u_t, u_{t-2}) = \rho^2 \cdot \sigma_u^2$$

Generally, $\operatorname{cov}(u_t, u_{t-\tau}) = \rho^\tau \cdot \sigma_u^2 = \rho^\tau \dfrac{\sigma_\varepsilon^2}{1-\rho^2}$.

From the equation $\operatorname{cov}(u_t, u_{t-1}) = \rho\sigma_u^2$ it follows, that

$\dfrac{\operatorname{cov}(u_t, u_{t-1})}{\sigma_u^2} = \rho$, and now it is clear, that this essential parameter has two meanings: it is not only <u>the autoregression coefficient</u>, but also the <u>autocorrelation coefficient of the first</u>

order (i.e. correlation coefficient between to consecutive disturbances u_t and u_{t-1}).

The assumption 3) can be now written as:

$$\bigwedge_t E(u_t) = 0 , \qquad \bigwedge_t D^2(u_t) = \sigma_u^2 = \frac{\sigma_\varepsilon^2}{1-\rho^2} ,$$

$$\bigwedge_\tau \operatorname{cov}(u_t , u_{t-\tau}) = \sigma_u^2 \cdot \rho^\tau$$

or in matrix notation:

the mean vector $E(\mathbf{u}) = \mathbf{0}$,

the covariance matrix $V(\mathbf{u}) = \sigma_u^2 \Omega = \frac{\sigma_\varepsilon^2}{1-\rho^2} \Omega$,

where matrix

$$\Omega = \begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-2} \\ \vdots & & & & \vdots \\ \rho^{n-1} & \rho^{n-2} & \cdots & \rho & 1 \end{pmatrix}$$

The inverse matrix Ω^{-1} is defined as:

$$\Omega^{-1} = \frac{1}{1-\rho^2} \underbrace{\begin{pmatrix} 1 & -\rho & 0 & \cdots & 0 \\ -\rho & 1+\rho^2 & -\rho & \cdots & 0 \\ 0 & -\rho & 1+\rho^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \\ 0 & 0 & 0 & -\rho & 1 \end{pmatrix}}_{V} = \frac{1}{1-\rho^2} V$$

12.2. Application of the GLS Method.

It is clear now that the linear model $\mathbf{y} = X\beta + \mathbf{u}$ with the assumption 3): $V(\mathbf{u}) = \sigma_u^2 \Omega$, i.e. with autocorrelated disturbances, should be estimated by menas of the Generalized Least Squares (GLS) method. Assuming that the autocorrelation coefficient ρ is known, the matrix Ω and its inverse Ω^{-1} is known also, and the GLS estimator $\hat{\beta} = (X^T \Omega^{-1} X)^{-1} X^T \Omega^{-1} y$ can be utilized directly to estimate the parameters vector β of the model.

If a computer program for the GLS estimator $\tilde{b}$ is not available, we can utilize instead the program for the OLS estimator b^* applied to the transformed model (see paragraph 11.1).

The appropriate transformation of the original model with autoregressive disturbances u_t should give us the model with classic(independent) disturbances ε_t.

Since $u_t = \rho u_{t-1} + \varepsilon_t$, we obtain $u_t - \rho u_{t-1} = \varepsilon_t$.

Therefore the transformation which should be applied to all observations is defined as:

$$y_t^* = y_t - \rho y_{t-1}$$

$$x_{ti}^* = x_{ti} - \rho \cdot x_{(t-1)i} \quad \text{for} \quad i=1,2,\dots\ k+1,\ t=2,3,\dots n.$$

This is clear, if we consider the simplest linear model with one explanatory variable only, i.e. $y_t = \beta x_t + \alpha + u_t$

Obviously $y_{t-1} = \beta x_{t-1} + \alpha + u_{t-1}$, $\rho y_{t-1} = \rho\beta x_{t-1} + \rho\alpha + u_{t-1}$

and the difference:

$$y_t - \rho y_{t-1} = \beta(x_t - \rho x_{t-1}) + \alpha(1-\rho) + (u_t - \rho u_{t-1})$$

or $$y_t^* = \beta x_t^* + \alpha \cdot 1^* + \varepsilon_t$$

where $$x_t^* = x_t - \rho x_{t-1} \ , \ 1^* = 1 - \rho \ , \ \varepsilon_t = u_t - \rho u_{t-1}$$

Infact, because such a transformation can be applied for n-1 observations only (t= 2,3,... n) the additional transformation should be applied for the first observation.

It is defined as $y_1^* = y_1\sqrt{1-\rho^2}$, $x_{1i}^* = x_{1i}\sqrt{1-\rho^2}$ for $i=1,2,\dots k+1$.

So, in matrix notation the transformation applied to the original model $y = X\beta + u$ (and to the original observations) in order to estimate it by means of the OLS method (estimator b^*) is as follows:

$$y^* = Py, \quad X^* = PX, \quad \varepsilon = Pu$$

where $$P = \begin{pmatrix} \sqrt{1-\rho^2} & 0 & 0 & \dots & 0 \\ -\rho & 1 & 0 & \dots & 0 \\ 0 & -\rho & 1 & \dots & \vdots \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & -\rho & 1 \end{pmatrix}$$

with properties $\mathbf{P}^T\mathbf{P} = (1-\rho^2)\boldsymbol{\Omega}^{-1}$

$$\mathbf{P}\boldsymbol{\Omega}\mathbf{P}^T = (1-\rho^2)\mathbf{I}$$

By the left-hand side multiplication of the original model by the matrix $\mathbf{P}$ we obtain the new, transformed model $y^* = X^*\beta + \varepsilon$ with the classic disturbance vector, i.e. $E(\varepsilon) = 0$ and with covariance matrix

$$V(\varepsilon) = E(\varepsilon\varepsilon^T) = E[\mathbf{P}u \cdot u^T\mathbf{P}^T] = \mathbf{P} \cdot E(uu^T)\mathbf{P}^T =$$

$$= \mathbf{P} \cdot V(u) \cdot \mathbf{P}^T = \mathbf{P} \cdot \sigma_u^2 \boldsymbol{\Omega} \cdot \mathbf{P}^T = \sigma_u^2 \mathbf{P}\boldsymbol{\Omega}\mathbf{P}^T =$$

$$= \sigma_u^2(1-\rho^2)\mathbf{I} = \frac{\sigma_\varepsilon^2}{1-\rho^2} \cdot (1-\rho^2) \cdot \mathbf{I} = \sigma_\varepsilon^2 \mathbf{I}$$

Therefore to the transformed model theOLS method can be applied i.e. we can apply the estimator b^* which is equivalent to the GLS estimator $\tilde{b}$:

$$b^* = (X^{*T}X^*)^{-1}X^{*T}y^* = (X^T\mathbf{P}^T\mathbf{P}X)^{-1}X^T\mathbf{P}^T\mathbf{P}y =$$

$$= \left[X^T(1-\rho^2)\boldsymbol{\Omega}^{-1}X\right]^{-1}X^T(1-\rho^2)\boldsymbol{\Omega}^{-1}y =$$

$$= (X^T\boldsymbol{\Omega}^{-1}X)^{-1}X^T\boldsymbol{\Omega}^{-1}y = \tilde{b}.$$

For application of the GLS method(directly via $\tilde{b}$ or indirectly via b^*) a numerical value of the basic parameter ρ (autocorrelation coefficient) should be known.

However in econometric practice the autocorrelation coefficient ρ is usually unknown and it should be estimated as well.

Since residuals $\hat{u}_t$ can be treaten as estimates of unobservable random disturbances u_t , the autocorrelation coefficient ρ of disturbances can be estimated based on residuals $\hat{u}_t$ obtained after preliminary

estimation of the original model $y = X\beta + u$ by the OLS method, i.e. by application of the estimator b.

In econometric literature many estimators of the parameter ρ have been proposed (since it plays double role - as the autoregression coefficient and as the autocorrelation coefficient), but the most often used estimator $\hat{\rho}$, which assures the condition for the stationary autoregressive process i.e. $|\hat{\rho}| < 1$, is defined as:

$$\hat{\rho} = \frac{\sum_{t=2}^{n} \hat{u}_t \cdot \hat{u}_{t-1}}{\sqrt{\sum_{t=2}^{n} \hat{u}_t^2 \cdot \sum_{t=2}^{n} \hat{u}_{t-1}^2}},$$

where $\hat{u}_t$ and $\hat{u}_{t-1}$ are consecutive residuals of the original model preliminarly estimated by the OLS method i.e. by applying of the estimator $b = (X^T X)^{-1} X^T y$.

Based on the estimate $\hat{\rho}$, the estimate $\hat{\Omega}$ of the matrix Ω (or the estimate $\hat{P}$) can be obtained.

In order to check if the autocorrelation coefficient estimate $\hat{\rho}$ is significantly different from zero, the special significance test known as the Durbin-Watson test for serial correlation (i.e. autocorrelation) should be applied. It is explained in next paragraph 12.3.

If the autocorrelation coefficient $\hat{\rho}$ is statistically non-significant, the preliminary OLS estimate b can be taken as the final estimate of the parameters vector β of the model.

However, if by applying of the Durbin-Watson test the autocorrelation of disturbances has been proved as statistically significant, the final estimate of the parameters vector β of the model should be obtained by the Generalized Least Squares method (directly via b with utilization of the estimated matrix $\hat{\Omega}$, or indirectly via b^* for the transformed model with the estimated transformation matrix $\hat{P}$).

12.3. Use of the Durbin-Watson Test for Autocorrelation.

One can enumerate many sources of autocorrelation of random disturbances in the econometric models, especially those estimated based on time-series statistical data. The main sources are:

(i) Mis-specification of the mathematical form of the model (i.e. the wrong type of function "f"),

(ii) Omitted important explanatory variable,

(iii) Effects of important random events like war, drought, flood or other disaster which influenced economy for many years.

Since especially the first two sources are difficult for detection, it is better to perform the Durbin-Watson test for autocorrelation whenever the time-series data were used for estimation of the econometric model.

The Durbin-Watson test for serial correlation(autocorrelation) has been designed to test the hypothesis $H_o : \rho = 0$, against $H_1 : \rho > 0$ where ρ is the autocorrelation coefficient from the first order autoregressive stationary process of disturbances of the linear model. The test is based on residuals $\hat{u}_t$ obtained from the estimated (by the OLS) model. The basic Durbin-Watson (D-W) statistic d is constructed from residuals as:

$$d = \frac{\sum_{t=2}^{n} (\hat{u}_t - \hat{u}_{t-1})^2}{\sum_{t=1}^{n} \hat{u}_t^2}$$

Since residuals (and therefore also statistic d) depend on the given observations matrix X, it was difficult to find for any matrix X such a critical value d_α, that for the given significance level α, $P\{d \leq d_\alpha\} = \alpha$.

Assuming the normal distribution of disturbances, it was possible however to find such bounds d_L and d_u for the statistic d, that $\bigwedge_X \; d_L \leq d_\alpha \leq d_u$

If $d < d_L$ (the critical region) we shall reject the hypothesis $H_o : \rho = 0$ (because in that case $d \leq d_\alpha$ with small probability α equal to the given significance level).

If $d \geq d_u$ we can not reject the hypothesis H_o (since in that case $d > d_\alpha$ with big probability $1 - \alpha$).

If $d_L < d < d_u$ (so called non-conclusive region) we do not know if we can reject H_o or not.

In Statistical Tables we can find the appropriate bounds d_L and d_u for given significance level α, given sample size n and given number of explanatory variables in the model K.

For the significance level $\alpha = 0.05$ these bounds are as follows:

n	K=1		K=2		K=3		K=4		K=5	
	d_L	d_u	d_L	d_u	d_L	d_u	d_L	d_u	d_L	d_u
15	1.08	1.36	0.95	1,54	0.82	1,75	0,69	1,97	0.56	2,21
16	1.10	1.37	0.98	1.54	0.86	1,73	0.74	1.93	0,62	2,15
17	1.13	1.38	1,02	1.54	0.90	1.71	0.78	1,90	0,67	2,10
18	1.16	1.39	1.05	1.53	0.93	1.69	0.82	1.87	0.71	2.06
19	1.18	1.40	1.08	1.53	0,97	1.68	0.86	1.85	0.75	2.02
20	1.20	1.41	1.10	1.54	1.00	1.68	0.90	1.83	0,79	1.99
22	1.24	1.43	1.15	1.54	1,05	1,66	0.96	1.80	0,86	1.94
24	1.27	1.45	1.19	1,55	1.10	1.66	1.01	1.78	0.93	1.90
26	1.30	1.46	1.22	1,55	1.14	1.65	1,06	1.76	0.98	1,88
28	1.33	1.48	1.26	1.56	1.18	1.65	1.10	1,75	1.03	1.85
30	1.35	1.49	1.28	1,57	1.21	1.65	1,14	1.74	1.07	1,83
35	1.40	1.52	1.34	1,58	1,28	1.65	1.22	1.73	1.16	1,80
40	1.44	1.54	1.39	1.60	1.34	1,66	1,29	1.72	1.23	1,79
45	1.48	1.57	1.43	1.62	1,38	1.67	1,34	1,72	1.29	1,78
50	1.50	1.59	1.46	1.63	1,43	1.67	1,38	1,72	1.34	1,77
60	1.55	1.62	1.51	1,65	1.48	1.69	1,44	1.73	1.41	1.77
70	1.58	1.64	1.55	1.67	1,52	1.70	1,49	1,74	1.46	1,77
80	1.61	1.66	1.59	1,69	1,56	1,72	1.53	1,74	1.51	1,77
90	1.63	1.68	1.61	1,70	1,59	1,73	1,57	1,75	1.54	1,78
100	1.65	1.69	1.63	1.72	1.61	1,74	1.59	1,76	1,57	1,78

If the null hypothesis was formulated as $H_o : \rho = 0$, against $H_1 : \rho < 0$, the statistic $d' = 4 - d$ should be used (instead of d) in the Durbin-Watson test. The Durbin-Watson statistic d is usually

included in printouts of regression analysis computer programs of many statistical packages (SPSS, ISEA, etc).

So, in case when we adopt assumption 3" (with autocorrelated disturbances) the final two-stage procedure of estimation of the model consists of the following sequence of activities:

1° Apply the OLS to the model $y = X\beta + u$ in order to obtain the residual vector $\hat{u} = y - Xb$ where $b = (X^T X)^{-1} X^T y$.

2° Apply the Durbin-Watson test for the hypothesis $H_o : \rho = 0$ based on the statistic d (in case of $H_1 : \rho > 0$), or on the statistic $d' = 4-d$ (in case of $H_1 : \rho < 0$).

3° If the result of testing is, that the hypothesis $H_o : \rho = 0$ can not be rejected, the preliminary OLS estimate b of vector β is good as final one.

4° If the hypothesis $H_o : \rho = 0$ has to be rejected (which means statistically significant autocorrelation), we have to calculate the estimate $\hat{\rho}$ of the autocorrelation coefficient ρ, obtain the estimate $\hat{\Omega}$ of the matrix Ω, and finally apply the GLS method either indirectly (via b^* applied to the transformed model with the transformation matrix $\hat{P}$), or directly (via b applied to the original model).

Example:

Let's assume that the following time-series data were collected to investigate the trend function of milk production (in thousand tons) in a some country.

year	t	1	2	3	4	5	6	7	8	9	10	11
Production	y_t	10	12	15	18	21	25	30	36	42	50	60

Based on these data the linear trend function $y_t = \beta t + \alpha + u_t$ was estimated by the OLS method and the result of that estimation was: $\hat{y}_t = \underset{(0.4)}{4.8}\, t + 0.2$, with $r^2 = 0.95$.

Let's perform the Durbin-Watson test for autocorrelation of disturbances. Assuming the first order outoregressive process $u_t = \rho u_{t-1} + \varepsilon_t$ with classic assumptions for ε_t (including normality of its distribution) we want to test the null hypothesis $H_o : \rho = 0$, against $H_1: \rho > 0$.

The working table for calculation of the Durbin-Watson statistic d, and the autocorrelation coefficient estimate $\hat{\rho}$ based on residuals $\hat{u}_t$ is as follows:

t	$\hat{y}_t$	$\hat{u}_t$	$\hat{u}_t^2$	$(\hat{u}_t - \hat{u}_{t-1})^2$	$\hat{u}_t \cdot \hat{u}_{t-1}$
1	5.0	5.0	25.00	-	-
2	9.8	2.2	4.84	7.84	11.00
3	14.6	0.4	0.16	3.24	0.88
4	19.4	- 1.4	1.96	3.24	- 0.56
5	24.2	- 3.2	10.24	3.24	4.48
6	29.0	- 4.0	16.00	0.64	12.80
7	33.8	- 3.8	14.44	0.04	15.20
8	38.6	- 2.6	6.76	1.44	9.88
9	43.4	- 1.4	1.96	1.44	3.64
10	48.2	1.8	3.24	10.24	-2.52
11	53.0	7.0	49.00	27.04	12.60
Σ			136.60	58.40	67.40

So, the D-W statistic d is:

$$d = \frac{\sum_{t=2}^{n} (\hat{u}_t - \hat{u}_{t-1})^2}{\sum_{t=1}^{n} \hat{u}_t^2} = \frac{58.40}{136.60} = 0.437 ,$$

while the critical value for $\alpha = 0.05$, $n=11$ and $k=1$ is $d_L \approx 1.0$. Since $d=0.44 < d_L$, we have to reject the null hypothesis $H_o: \rho = 0$, which means, that autocorrelation is statistically significant.

The estimate $\hat{\rho}$ of the outocorrelation coefficient is

$$\hat{\rho} = \frac{67.4}{\sqrt{108.6 \cdot 84.6}} = 0.7$$

The final estimate of the trend function model can now be obtained by the GLS method indirectly, i.e. by applying again the OLS estimator b^* to the transformed model.

For that purpose we shall transform all original observations (t, y_t) into (t^*, y_t^*), where $y_t^* = y_t - \hat{\rho} y_{t-1}$, $t^* = t - \hat{\rho}(t-1)$ for $t = 2, 3, \ldots n$

and $y_1^* = y_1 \cdot \sqrt{1-\hat{\rho}^2} = y_1 \cdot 0.71$,

$$t_1^* = t_1 \cdot 0.71 = 0.71$$

Based on such transformed observation, we can now apply the OLS method to the model:

$$y_t^* = \beta t^* + \alpha \cdot 1^* + \varepsilon_t = \beta t^* + \alpha(1-\rho) + \varepsilon_t =$$

$$= \beta t^* + \alpha' + \varepsilon_t, \quad \text{where } \alpha' = \alpha(1-\rho)$$

Numerical results are: $\hat{\beta} = 7.0$, $\hat{\alpha'} = -5.78$,

$$\hat{\alpha} = \frac{\hat{\alpha'}}{1-\hat{\rho}} = \frac{-5.78}{0.3} = -19.27$$

So, the final GLS estimate of the model is:

$$\hat{y}_t^* = \underset{(0.3)}{7.0}\, t^* - 19.27 \cdot 1^* = 7.0\, t^* - 19.27(1-\hat{\rho})$$

12.4. Forecasting Based on the Model with Autoregressive Disturbances.

In case of the linear model $y = X\beta + u$ with autocorrelation of disturbances u_t the formula for getting a forecast of the dependent variable is much more complicated than for model with classic

assumptions, because forecasts have to be obtained step by step forward (i.e. in each step, for one future period of time only).

Let's denote: $T = n+1$ - a one step forward future (forecasted) period of time

$X_{(n+1)}$ - vector of given future values of explanatory variables

$\tilde{\beta}$ - the GLS estimate of the parameters vector β

$\tilde{u}_n$ - the last residual of the model estimated by the GLS.

The formula for the (one period forward) forecast of the explained variable is given as:

$$\hat{y}_{n+1} = X_{(n+1)}^T \tilde{\beta} + \rho \cdot \tilde{u}_n ,$$

and the variance of this forecast is given by the formula:

$$D^2(\hat{y}_{n+1}) = \sigma_\varepsilon^2 \left[1 + (X_{(n+1)} - \rho X_{(n)})^T (X^T V X)^{-1} (X_{(n+1)} - \rho X_{(n)}) \right]$$

where $\sigma_\varepsilon^2 = \sigma_u^2 (1 - \rho^2)$

$X_{(n)}$ is the vector of last observations on explanatory variables,

$V = (1-\rho^2) \Omega^{-1}$ (see the end of paragraph 12.1)

It is easier to understand that formula for the forecast, if we consider the simplest linear model with one explanatory variable only, i.e. the model:

$$y_t = \beta x_t + \alpha + u_t \quad , \text{ where } u_t = \rho u_{t-1} + \varepsilon_t$$

Let $\tilde{\beta}$ and $\tilde{\alpha}$ be the GLS estimates for β and α, and let $\tilde{u}_n = y_n - (\tilde{\beta} x_n + \tilde{\alpha})$ be the last residual.

Then the one period forward forecast of the dependent variable is (accordingly to the formula given above):

$$\hat{y}_{n+1} = (x_{n+1} \quad 1)\cdot\begin{pmatrix}\tilde{\beta}\\ \tilde{\alpha}\end{pmatrix} + \rho(y_n - \tilde{\beta}x_n - \tilde{\alpha}) = \tilde{\beta}x_{n+1} + \tilde{\alpha} + \rho y_n - \tilde{\beta}\rho x_n - \rho\tilde{\alpha} =$$

$$= \tilde{\beta}(x_{n+1} - \rho x_n) + \tilde{\alpha}(1-\rho) + \rho y_n$$

or $\hat{y}_{n+1} - \rho y_n = \tilde{\beta}(x_{n+1} - \rho x_n) + \tilde{\alpha}(1-\rho)$

or $\hat{y}^*_{n+1} = \tilde{\beta}x^*_{n+1} + \tilde{\alpha}\,1^*$, which is equal to the classic forecast based on the transformed model, where $\hat{y}^*_{n+1} = \hat{y}_{n+1} - \rho y_n$,

$x^*_{n+1} = x_{n+1} - \rho x_n$, $1^* = 1 - \rho$.

For the simplest model $y_t = \beta x_t + \alpha + u_t$, the variance of the forecast can be written in the following simpler form:

$$D^2(\hat{y}_{n+1}) = \sigma_\varepsilon^2 \left[1 + \frac{(x_{n+1} - x_n)^2}{\sum_{t=2}^{n}(x_t - \rho x_{t-1})^2}\right]$$

The variance σ_ε^2 can be estimated as the residual variance in the estimated transformed model.

Example:

Based on the estimated trend function from the example given in previous paragraph 12.3, let us calculate the forecast of milk production for the coming year T = n+1 = 12.

We shall use the formula: $\hat{y}_{n+1} = \tilde{\beta}(x_{n+1} - \rho x_n) + \tilde{\alpha}(1-\rho) + \rho y_n$.

From the numerical results given in the example in former paragraph we have:

$\tilde{\beta} = 7.0$, $\tilde{\alpha} = -19.3$, $\hat{\rho} = 0.7$, $n=11, x_{n+1} = n+1 = 12$, $x_n = 11$, $y_n = 60.0$

So, the forecast of the milk production is calculated as:

$$\hat{y}_{12} = 7(12 - 0.7 \cdot 11) - 19.3(1-0.7) + 0.7 \cdot 60.0 =$$

$$= 7 \cdot 43 - 5.78 + 42 = 66.3 \text{ (thousand tons of milk)}$$

The standard error of this forecast is:

$$SE(\hat{y}_{12}) = D(\hat{y}_{12}) = \sqrt{D^2(\hat{y}_{12})} =$$

$$= \sqrt{0.98\left[1 + \frac{(12-0.7\cdot 11)^2}{77.7}\right]} = \sqrt{0.98 \cdot 1.238} =$$

$$= \sqrt{1.21} = 1.1 \text{ (thousand tons of milk).}$$

CHAPTER 13. ANALYSIS OF RESIDUALS - A SUMMARIZED PICTURE.

13.1. Flow-Chart of the Proper Sequence of Tests Applied.

Since almost all statistical tests applied during the stage of checking of the estimated model are based on residuals (as estimates of random disturbances) that stage of the econometric modelling is often called the analysis of residuals. The main aim of the analysis of residuals is to check (i.e. to support or reject) the whole set of assumptions adopted for the structure of the model in order to obtain its best estimation.

As we have seen in previous chapters there are many statistical tests which should be applied during the analysis of residuals for checking each particular aspect of the assumed stochastic structure of the model. But each out of these tests is based on its own assumptions with normal distribution of disturbances as essential common assumption.

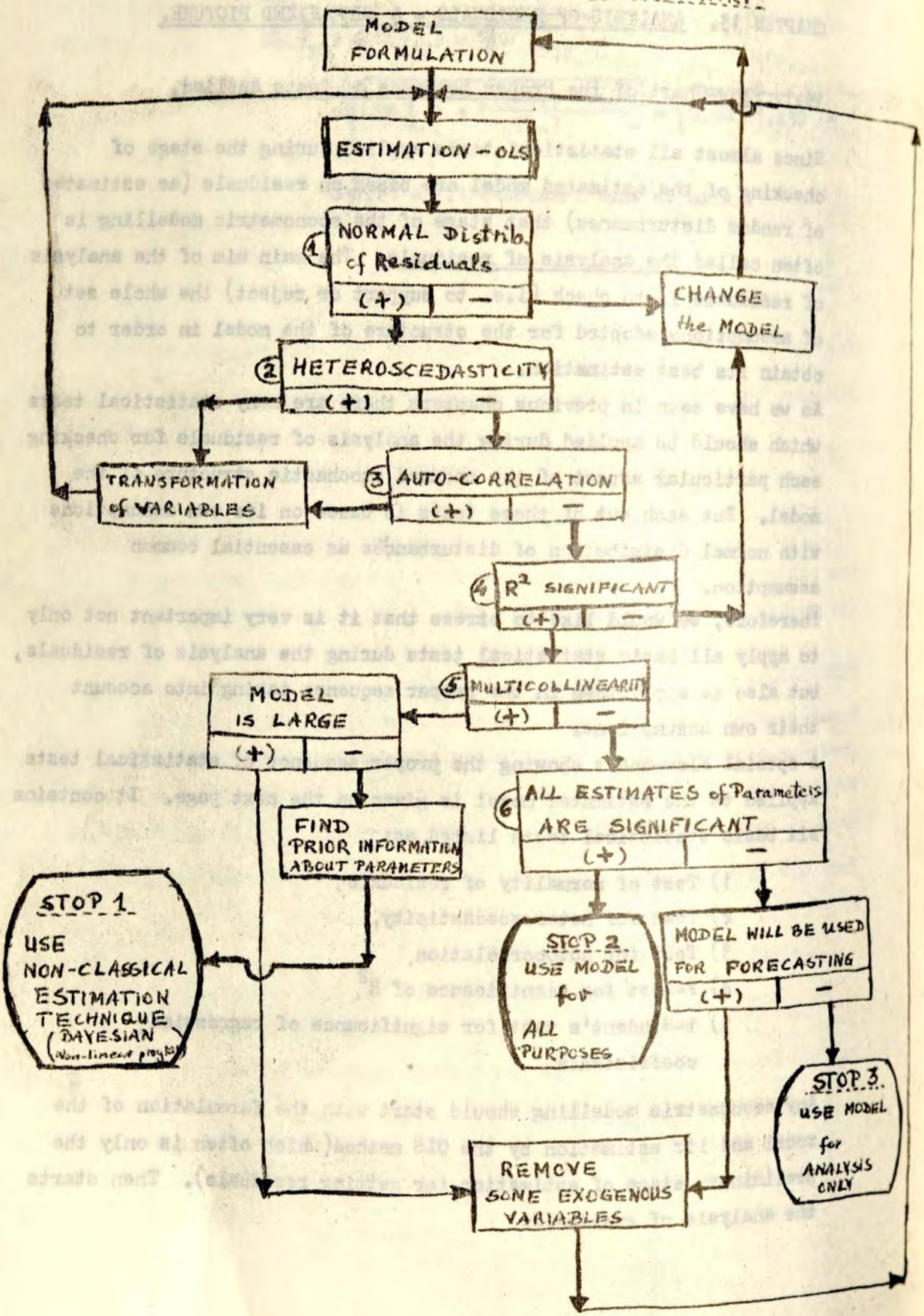
Therefore, we would like to stress that it is very important not only to apply all basic statistical tests during the analysis of residuals, but also to apply them in the proper sequence taking into account their own assumptions.

A special Flow-chart showing the proper sequence of statistical tests applied to the estimated model is given on the next page. It contains six basic statistical tests listed as:

1) Test of normality of residuals,
2) Test for heteroscedasticity,
3) Test for autocorrelation,
4) F-test for significance of R^2,
5) t-Student's test for significance of regression coefficients,

Any econometric modelling should start with the formulation of the model and its estimation by the OLS method(which often is only the preliminary stage of estimation for getting residuals). Then starts the analysis of residuals.

ANALYSIS OF RESIDUALS – FLOW-CHART



At the beginning, the normality test of residuals (1) should be applied, since the normality assumption is essential for other tests. For large sample size the Chi-square test can be used, while for a small sample the Kolmogorov' or the Hellwig' tests can be used as normality tests (see paragraph 10.3). If the hypothesis about normality is rejected (sign "—" in the Flow-chart) the model was formulated in a wrong way and it should be changed (by including some additional important explanatory variable or by changing the type of function). If the hypothesis about normality can not be rejected, we may continue the analysis of residuals from the estimated model.

Next we shall analyze the heteroscedasticity of random disturbances (2), especially in case when a cross-sectional type of statistical data were used for estimation of the model. Depends on which type of heteroscedasticity is assumed (if any), or is significant through the Bartlett's test (see Chapter 11) the appropriate transformation of variables should be applied and then we shall comeback to the estimation by the OLS method again.

If the heteroscedasticity is not significant (or it was neglected from the beginning for example in case of time-series data), which is denoted by the sign "—", we can continue the analysis.

The next step (3) concerns checking of the autocorrelation of residuals based on the Durbin-Watson test described in paragraph 12.3. That is especially important in case, when time-series data were used for estimation of the model.

If the autocorrelation is significant, which is denoted by the sign "+", the estimate $\hat{\rho}$ of the autocorrelation coefficient should be calculated and the proper transformation of variables should be applied to obtain the transformed model which should be again estimated by the OLS.

If the autocorrelation is not significant (sign "—") we can move to the next step of analysis (4) connected with the coefficient of determination R^2.

The F-test of significance for R^2 (see paragraph 9.3) should be applied here. The model can not be approved if its R^2 is statistically not significant(sign"—"). In this case the model should be changed(reformulated). If the coefficient of determination R^2 is significant(sign "+") we may continue the analysis.

Next, we shall analyse the multicollinearity problem (5) (in case when the model has many explanatory variables). The appropriate test of multicollinearity (see paragraph 10.2) should be applied.

Only if the multicollinearity is not significant (sign "-") we can move to the last t-Student test for regression coefficients.

If the multicollinearity is significant (sign"+") we have two ways to improve the estimation results. If the estimated model is large we can remove some exogenous variables engaged in multicollinearity and then we can estimate the new, smaller model. If the estimated model with significant multicollinearity is rather small, we shall try to find some additional, prior information about parameters and use one out of non-classical estimation techniques (see next chapter 14).

At the end of the analysis of residuals(and not at the beginning) the Student's t-test should be applied for testing a significance of regression coefficients (6).

If for all explanatory variables their regression coefficients are statistically significant(sign "+") the estimated model can be used for all purposes(i.e. for analysis as well as for forecasting). If not all regression coefficients are statistically significant and the model will be used for forecasting, we shall remove all non-significant explanatory variables and come back again to the estimation stage.

An example of full analysis based on that sequence is given in the next paragraph.

13.2. Example of a Full Analysis.

We shall illustrate now a full analysis of residuals on the following econometric model formulated for studying of the relationship between crop yields and utilization of fertilizers and machinery, based on statistical data from the Polish Agriculture.
The cross-sectional data from n=49 sub-regions called "the vojevodships" in Poland (which are like Awrajas in Ethiopia) were used for estimation of the linear model:

$$Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 + \varepsilon \quad , \quad \text{where}$$

Y is defined as the average yield of 4 main cereal crops (wheat, rye, barley, oats) measured in q/ha.

X_1 is defined as the quantity of mineral fertilizers applied, measured in quintals of the NPK per hectare.

X_2 is defined as a use of machinery, measured as the number of hectares of cultivated land per one tractor used in agriculture.

Numerical results of estimation of the model (by the OLS method) are as follows:

$$\hat{Y}_t = \underset{(1.23)}{3.50}\, x_{t1} - \underset{(0.07)}{0.16} x_{t2} + 25.91 \quad , \quad S = 2.46 \quad , \quad R^2 = 0.69$$

which means that the increment of mineral fertilizer application by 1 q/ha influences the increase of cereal crop yield by 3.50 q/ha, and the decrease by 1 hectare the number of hectares per one tractor influences the increase of cereal crop yield by 0.16 q/ha.
We start the analysis of residuals with application of the Kolmogorov' test of normality i.e. to test the hypothesis H_o: $F(\hat{\varepsilon}_t) = F_N$, where F_N is the distribution function of the normal distribution. The calculated (empirical) value of the Kolmogorov D statistic is D=0.064, while the critical value for $\alpha = 0.05$ and n=49, is $D_\alpha = 0.190$. Since $D < D_\alpha$, we can not reject the hypothesis H_o about normality

of residuals.

For testing the grouped heteroscedasticity using the Bartlett' test, the whole set of empirical observations(and consequently the set of residuals) was split into 4 groups according to agro-geografical zones in Poland. Within these groups residual variances were: $S_1^2 = 1.37$, $S_2^2 = 14.78$, $S_3^2 = 5.04$, $S_4^2 = 3.67$. Based on these values the appropriate Chi-square statistic was calculated as $\chi^2 = 7.056$ while the critical value for $\alpha = 0.05$ and df = 3, is $\chi^2_\alpha = 7.815$. Since $\chi^2 < \chi^2_\alpha$ we can not reject the hypothesis $H_o : \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$, which means that the grouped heteroscedasticity is statistically non-significant.

The autocorrelation of residuals was also non-significant(the Durbin-Watson statistic $d = 1.64 > d_u = 1.59$ for $\alpha = 0.05$, n=49) as we might expect, since the cross-sectional data were used for estimation of the model.

The next applied test is the F-test for testing if the coefficient of determination $R^2 = 0.69$ is statistically significant.

The appropriate value of F-Snedecor statistic was calculated as $F = \frac{0.69}{1-0.69} \cdot \frac{46}{2} = 51.2$, while the critical value (for the significance level $\alpha = 0.05$ and df= 2 and 46) is $F_\alpha = 3.20$.

Since $F > F_\alpha$, the $R^2 = 0.69$ is significant and it can be interpreted, that using those two explanatory variables (mineral fertilizers applied and machinery used) in the linear model, we can explain almost 70% of the cereal crop yield variation.

As the next, the Glauber -Farrar-Haitovski test for multicollinearity was applied, based on the determinant of the correlation matrix $\mathbb{R}$ constructed for the two explanatory variables X_1 and X_2.

The empirical correlation matrix has been obtained as

$$\mathbb{R} = \begin{pmatrix} 1 & -0.84 \\ -0.84 & 1 \end{pmatrix}$$ with its determinant $\det \mathbb{R} = 0.294$.

The appropriate Chi-square statistic(see paragraph 10.2) was $\chi^2 = 16.19$, while the critical value(for $\alpha = 0.05$ and df=1) is $\chi^2_\alpha = 3.841$

Since $\chi^2 > \chi^2_\alpha$, the null hypothesis H_o : $\det \mathbf{R} = 0$ (which means multicollinearity) has to be rejected, therefore in that case the multicollinearity is statistically non-significant and can be neglected.

Now, at the end of analysis, we may use the crucial t-Student test for testing significance of regression coefficients.

Since $b_1 = 3.50$ and $S_{b_1} = 1.23$, the appropriate value of t statistic is $t_1 = 2.84$.

For $b_2 = -0.16$ and $S_{b_2} = 0.074$, we have $t_2 = -2.16$.

The critical value of the t-Student's statistic(for $\alpha = 0.05$ and df = 46) is $t_\alpha = 2.01$.

Since for both cases we have $|t| > t_\alpha$, we may conclude, that both estimated regression coefficients are statistically significant.

If we are particularly interested in estimation of the regression coefficients β_i for the first explanatory variable X_1 (fertilizers applied)we can construct the confidence interval for it(assuming $1-\alpha = 0.95$) as: $3.50 - 2.01 \cdot 1.23 < \beta_1 < 3.50 + 2.01 \cdot 1.23$, which gives $1.03 < \beta_1 < 5.97$.

Since both estimates of regression coefficients are statistically significant the estimated model can be used also for forecasting.

Let's assume, that we are interested in obtaining a forecast of the cereal crop yield based on the estimated model and on the assumption, that the future value of the first explanatory variable will be $x_{T1} = 3.0$ (quintals of the NPK fertilizers per hectare), and the value of the second explanatory variable will be $x_{T2} = 20.0$ (hectares per one tractor).

So, we want to predict y_T for the given vector $\mathbf{X}^T_{(T)} = (3\ 20\ 1)$ of future values of explanatory variables. Using appropriate formulas

from the paragraph 9.4, we obtain:

$$\hat{y}_T = (\,3 \quad 20 \quad 1) \cdot \begin{pmatrix} 3.50 \\ -0.16 \\ 25.91 \end{pmatrix} = 33.21 (q/ha)$$

The standard error of this forecast is estimated as $SE(\hat{y}_T) = S_{\hat{y}_T} = 2.91$, So, the relative standard error is about 9%.

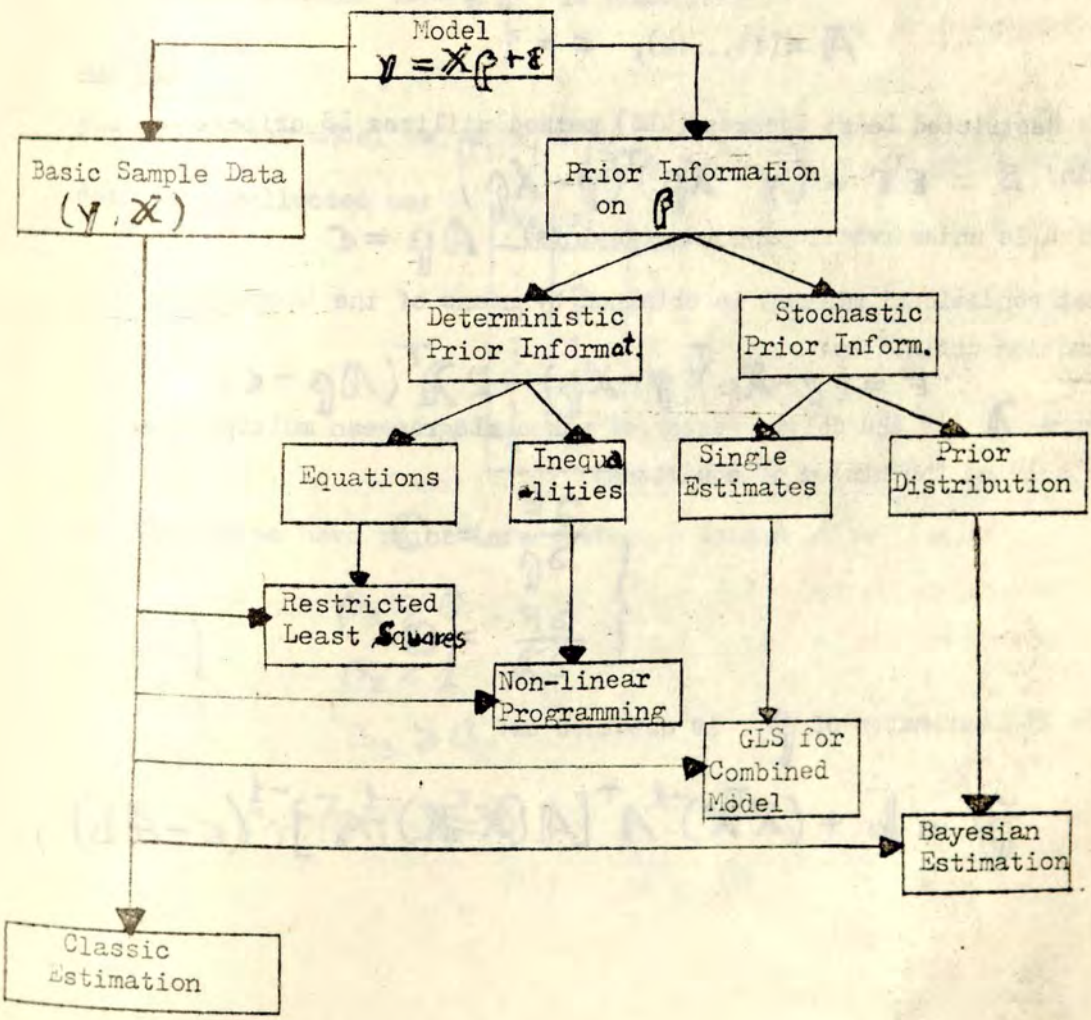
CHAPTER 14. USE OF PRIOR INFORMATION. NON-CLASSICAL ESTIMATION TECHNIQUES.

14.1. The Restricted Least Squares Method.

When some additional(or prior) information on the structural parameters β_i of the model are available we shall utilize them in addition to the basic sample data. That is especially advisable for small models with multicollinearity.

There are many so called Non-Classical Estimation Techniques which deal with utilization of prior information on estimated parameters. The proper applied technique depends on the type of prior information we have.

A general classification of all non-classical estimation techniques is given in the following scheme:



Let's start to make short review of those non-classical estimation techniques with the Restricted Least Squares (RLS) method. It is the best estimation technique in case when deterministic prior information in form of equations is available in addition to the basic sample data.

Let's assume that we have prior information in the form of l equations written in the matrix notation as:

$$A\beta = c \quad , \quad \text{where } l \times (k+1) \text{ matrix } A \text{ and } l \times 1 \text{ vector } c \text{ are given.}$$

For example, if we know in advance, that $\beta_1 = 2$, $\beta_3 = 1$

we have $$A = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \end{pmatrix}, \quad c = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$$

If we know, that $$\sum_{i=1}^{k} \beta_i = 1$$

$$A = (11 \dots 10), \quad c = 1.$$

The Restricted Least Squares (RLS) method utilizes LS criterion

$$(\min)\ S = \varepsilon^T \varepsilon = (y - X\beta)^T (y - X\beta)$$

which is under restriction (constraints): $A\beta = c$

That conditional minimum is obtained by means of the Lagrangean Function defined as:

$$F = (y - X\beta)^T (y - X\beta) - 2\lambda^T (A\beta - c)$$

where λ is the column vector of unknown Lagrangean multipliers.

By solving the system of equations:

$$\begin{cases} \dfrac{\partial F}{\partial \beta} = 0 \\ \dfrac{\partial F}{\partial \lambda} = 0 \end{cases}$$

the RLS estimator of β is obtained as:

$$\hat{\beta} = b + (X^T X)^{-1} A^T \left[A (X^T X)^{-1} A^T \right]^{-1} (c - Ab),$$

where b is the classic OLS estimator of β.

The RLS estimator $\hat{\beta}$ is unbiased, and its covariance matrix is given by the formula:

$$V(\hat{\beta}) = V - \underbrace{VA^T(AVA^T)^{-1}AV}_{Z} = V - Z \quad , \text{ where } V = V(b) = \sigma^2(X^TX)^{-1}$$

Since the matrix Z is positive definite, it follows, that

$D^2(\hat{\beta}_i) < D^2(b_i)$, which means that the RLS estimator is more efficient than the OLS estimator (because it utilizes more information than OLS estimator b).

Sometimes the deterministic prior information have an inequality form: $\beta_i \geq 0$ or $a < \beta_i < b$

For such inequalities restrictions, the solution of the RLS method can be obtained by means of the Non-Linear Programming (since the criterion function $S = \varepsilon^T\varepsilon$ is quadratic).

Example:

For the linear model $y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 + \varepsilon$, the basic sample data were collected as:

x_1	x_2	y
1	0	3
0	1	7
1	1	8
2	1	11
1	2	14
2	2	16

In addition we have prior information(in inequalities form):

$$\begin{cases} \beta_1 \geq 0 \\ \beta_2 \geq 1 \\ \beta_3 \geq 0 \\ \beta_1 + \beta_2 + \beta_3 \leq 8 \end{cases}$$

Using the Non-linear programming method (e.g. the Rosen's Gradient Projection Method) we obtain the RSL estimates as:

$\hat{\beta}_1 = 0$, $\hat{\beta}_2 = 6.33$, $\hat{\beta}_3 = 1.67$.

For this solution two constraints are fulfiled as equations:

$\hat{\beta}_1 = 0$ and $\hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3 = 8$, therefore it is equal to the RLS estimate $\hat{\beta}$ obtained for the matrix $A = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \end{pmatrix}$ and vector $C = \begin{pmatrix} 0 \\ 8 \end{pmatrix}$.

14.2. Use of the GLS Method of the Combined Model.

Sometimes estimation of a given econometric model is repeated year by year(or after few years) and former estimates are available. Such prior information has obviously a stochastic character, but it can be utilized to construct a combined model (with both sample data and those stochastic prior information) which can be estimated by the GLS method.

Let's consider the linear model $y = X\beta + \varepsilon$ with the basic sample information (y, X) and the prior stochastic information in the form of unbiased estimates c_1 (from previous estimation) for the sub-vector β_1 of the vector $\beta = \begin{pmatrix} \beta_1 \\ \hline \beta_2 \end{pmatrix}$

Those stochastic prior information can be written as

$$c_1 = \beta_1 + d \text{ , where } E(d) = 0, \; V(d) = W$$

$$\text{or } c_1 = A\beta + d \text{ , where } A = [I \,|\, 0]$$

It is also assumed, that random vectors d and ε are non-correlated i.e. $E(d \cdot \varepsilon^T) = 0$.

Now the Combined Model (for both basic sample data and stochastic prior information) can be constructed as:

$$\left(\begin{array}{c} y \\ \hline c_1 \end{array}\right) = \left(\begin{array}{c} X \\ \hline A \end{array}\right)\beta + \left(\begin{array}{c} \varepsilon \\ \hline d \end{array}\right)$$

The joint disturbance vector in the combined model has parameters

the mean vector $E\begin{pmatrix}\varepsilon\\ d\end{pmatrix} = 0$,

the covariance matrix

$$V\begin{pmatrix}\varepsilon\\ d\end{pmatrix} = E\left[\begin{pmatrix}\varepsilon\\ d\end{pmatrix}\begin{pmatrix}\varepsilon\\ d\end{pmatrix}^T\right] = E\left[\begin{pmatrix}\varepsilon\\ d\end{pmatrix}(\varepsilon^T \mid d^T)\right] =$$

$$= E\begin{pmatrix}\varepsilon\varepsilon^T & \varepsilon d^T\\ d\varepsilon^T & d d^T\end{pmatrix} = \begin{pmatrix}\sigma^2 I & 0\\ 0 & W\end{pmatrix} =$$

$$= \sigma^2\begin{pmatrix}I & 0\\ 0 & \frac{1}{\sigma^2}W\end{pmatrix} = \sigma^2\Omega.$$

So, the Generalized Least Squares (GLS) can be applied for estimation of the parameters vector β of the Combined Model.

If we denote the upper part of the matrix Ω by $A = [I \mid 0]$ as before, the GLS estimator $\tilde{\beta}$ based on the Combined Model is obtained as:

$$\tilde{\beta} = (X^TX + \sigma^2A^TW^{-1}A)^{-1}\cdot(X^Ty + \sigma^2A^TW^{-1}c_1)$$

and its covariance matrix is:

$$V(\tilde{\beta}) = \sigma^2(X^TX + \sigma^2A^TW^{-1}A)^{-1}.$$

It is worth noticing here, that since the matrix A has very simple form, it is easily to obtain the appendix $\sigma^2A^TW^{-1}A$, which can spoilt an eventual multicollinearity.

It means, that the Combined Model can be used in the case of observed multicollinearity in matrix X, because even for X^TX being singular matrix, the new matrix $(X^TX + \sigma^2A^TW^{-1}A)$ is not singular.

14.3. Bayesian Estimation Technique.

During the last 20 years the Bayesian inference has been depeloped in Statistics and Econometrics quite well. This idea is based on the Bayes theorem(from the Probability Theory) which gives the posterior probability distribution from the prior distribution and the likelihood function from the basic sample. The Statistical

Games Theory developed by A. Wald proves, that the Bayes estimators are optimal from the decision theory viewpoint, i.e. when the loss-function is involved in estimation procedure in order to measure consequences of the estimation error.

Let us esplain now briefly the main points of the Bayesian estimation technique.

We are interested in estimation of the parameter Θ of the given probability distribution $F(x \mid \Theta)$, based on random sample $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in \mathcal{X}$ (sample space).

The estimate "a" of the parameter Θ and the true value of the estimated parameter belong to the same parameter space Ω, i.e. $\Theta \in \Omega$, $a \in \Omega$.

The loss function $L(\Theta, a) \geq 0$ is defined for each pair (Θ, a) and its value measures the consequences of the estimation error.

A statistical decision function d(called an estimator) is defined as $d : \mathcal{X} \to \Omega$, $d(\mathbf{x}) = a$, $\mathbf{x} \in \mathcal{X}$, $a \in \Omega$.

The whole set of estimators d is denoted by D i.e. $d \in D$.

The <u>risk function $R(\Theta, d)$</u> is defined as the expected value of the loss function $L(\Theta, a)$, i.e.

$$R(\Theta, d) = E\left[L(\Theta, a)\right] = \int_{\mathcal{X}} L(\Theta, a)\, dF(\mathbf{x} \mid \Theta)$$

Now, the estimation problem is described as the <u>Statistical Game</u> which is defined as the triple (Ω, D, R), where the parameter space Ω is the set of all strategies of the first player (the Nature), D is the set of all strategies (estimators) of the second player (the Statistician), and the risk function $R(\Theta, d)$ is the pay-off function in this two-person game.

It is assumed, that the Nature is not a reasonable player (nor a malevolent) and the single strategy ($\Theta \in \Omega$) of the Nature is choosen randomly from given probability distribution $G(\Theta)$ defined on the parameter space Ω, and called <u>the prior distribution</u> of the parameter Θ.

The prior distribution $G(\Theta)$ reflects the whole prior information on parameter Θ we have before sample X is selected.

The Bayes risk $r(G, d)$ is defined as the expected value of the risk function $R(\Theta, d)$ with respect to the prior distribution $G(\Theta)$, i.e.

$$r(G, d) = E\left[R(\Theta, d)\right] = \int_{\Omega} R(\Theta, d)\, dG(\Theta).$$

It is obvious, that the best strategy for the Statistician is such an estimator $d(X)$, which minimizes the Bayes risk $r(G, d)$. It is called the Bayes estimator $d_o \in D$, if $r(G, d_o) = \min_{d \in D} r(G, d)$

It can be proved, that such an optimal (Bayes) estimator $d_o(X) = a$ can be obtained for a given sample X, by minimization of the expected loss function $L(\Theta, a)$ with respect to the Posterior Distribution $G(\Theta|X)$ obtained via the Bayes Theorem from the prior distribution $G(\Theta)$ and the likelihood of the sample X.

In Econometrics we are looking however for the vector-estimate a of the parameters vector Θ (which is in fact our vector β from the multiple regression model), and usually the loss function is defined as the quadratic form:

$$L(\Theta, a) = (\Theta - a)^T C (\Theta - a)$$, where C is the given symmetric, positive-definite matrix

It can be proved, that for such a loss function involved in estimation of the parameters vector Θ, the Bayes estimator $d_o(X) = a$ of the vector Θ, is equal to the mean vector of the multi-dimensional posterior distribution $G(\Theta|X)$, i.e. $d_o(X) = E(\Theta|X)$.

So, the sequence of work in the Bayesian Estimation Technique applied for econometric models is as follows:

1° The joint prior distribution $G(\beta)$ should be choosen as reflecting the whole prior knowlegde we have about the parameters vector β of the estimated econometric model. Usually the empirical prior distribution (estimated from the results of

previous estimation of the model) can be used, or theoretical prior distribution choosen from the conjugate family of distributions, for (both prior and posterior distributions belong to that family) can be used alternatively.

2° For given sample (y, X) for the estimated model, its likelihood $l(y, X, \beta)$ should be obtained. Usually the likelihood for the normal distribution of the sample vector y is used in econometric practice.

3° For given prior distribution with the pdf as $g(\beta)$ and for the likelihood $l(y, X, \beta)$, the joint posterior distribution described by the posterior joint density function $p(\beta | y, X)$ should be calculated (from the Bayes Theorem) as:

$$p(\beta | y, X) \propto g(\beta) \cdot l(y, X, \beta)$$

where the symbol $\propto$ means "is proportional to", i.e. neglecting all constants.

4° Assuming the quadratic loss function, the Bayes estimate $\hat{\beta}_B$ of the vector β should be calculated as the mean vector of the joint posterior distribution, i.e.

$$\hat{\beta}_B = E(\beta | y, X) = \int \cdots \int \beta \cdot p(\beta | y, X) d\beta,$$

where the multiple integral can be computed using a special computer program for multi-dimensional numerical integration.

Examples:

1. Mrs Rokaya Hassan (Egipt) working for her PhD in Poland under supervision of the author of this Handbook, obtained the following (OLS) estimation of the dynamic demand function for cereals and starches in Egipt based on time-series data (the power-type function after logarithmic transformation):

$$\hat{Y} = \underset{(0.011)}{0.460 X_1} - \underset{(0.006)}{0.270 X_2} - \underset{(0.012)}{0.012\ t} + 4.19$$

where X_1 stands for the logarithm of income (per capita),

X_2 stands for logarithm of price index,

t stands for time variable.

Alternatively, using the Bayesian estimation(with the beta type prior distribution for parameters) she obtained however more precise estimates for parameters(elasticities) of that demand function model and more realistic(positive) trend coefficient:

$$\hat{Y}_B = \underset{(0.003)}{0.359\ X_1} - \underset{(0.001)}{0.177\ X_2} + \underset{(0.011)}{0.027\ t}.$$

2. Based on the cross-sectional households consumption survey data we obtained the following estimate of the demand function for fish and fish-products in Poland using the Bayesian estimation technique (with the empirical beta type prior distribution for elasticities):

$$\hat{Y}_B = \underset{(0.02)}{0.84\ X_1} - \underset{(0.04)}{0.45\ X_2} - 6.02$$

where Y stands for the logarithm of household's expenditure for fish and fish-products,

X_1 stands for the logarithm of household's income

x_2 stands for the logarithm of price of the aggregate fish-products consumed by a household.

3. In 1964 M. Morishima and M. Saito tried to estimate the dynamic Cobb-Douglas production function (see paragraph 15.2):

$Y_t = L_t^{\alpha} K_t^{\beta} e^{\gamma t}$ for the USA economy, based on time-series data(1901-1952). The OLS method failed completely (the estimated elasticity of capital, $\hat{\beta}$ was negative), because of multicollinearity between the L_t and K_t (Labour and Capital).

In 1973 A.Zellner and J.-F. Richard using the same time-series data but the Bayesian estimation technique (with the beta type prior distribution for elasticities) obtained reasonable estimates for parameters of the same model:

$$\hat{\alpha}_B = 0.597\ , \qquad \hat{\beta}_B = 0.115\ , \qquad \hat{\gamma}_B = 0.018\ .$$

CHAPTER 15. MAIN DOMAINS OF APPLIED ECONOMETRIC MODELS.

15.1. Econometric Demand Functions and Elasticities.

The demand of consumers reflects their preferences for buying some quantity of commodities. Assuming the market equilibrium in the economy, the demand can be measured as expenditures of consumers for given commodity.

In order to explain the variation of demand, the econometric demand function is usually estimated and analyzed.

The econometric demand function for a particular commodity is defined as a stochastic equation which reflects the statistical relationship existing between the demand and different variables (factors) which influence the demand according to a regularity in consumers preferences.

A general model of the econometric demand function can be written as:

$$D = f(X_1, X_2, \dots, X_k, \varepsilon), \text{ where}$$

$X_1, X_2, \dots X_k$ are the different factors which influence the demand D for a particular commodity, ε is the random disturbance.

One can enumerate the various factors considered as explanatory variables in the econometric demand function like:

(i) the disposable income of consumers(per capita),
(ii) the real price of the commodity,
(iii) the price of a substitute (or complementary) commodity,
(iv) the demographic structure of consumers,
(v) the consumption pattern and fashion, etc.

The type "f" of demand functions is usually taken as linear or power-type.

Sometimes, due to lack of reliable data on various factors X_s, the econometric demand function is restricted to so called Engel' curve, which reflects the relationship between the demand for a particular commodity and the consumers' income X only. The type "f" of Engel' curves

depends on the importance of the given commodity for consumers, and it can be:

- linear $D = \beta X + \alpha + \varepsilon$
- power-type $D = \alpha X^{\beta} \cdot \varepsilon$
- logarithmic $D = \beta \log X + \alpha + \varepsilon$
- hyperbolic $D = \alpha + \frac{\beta}{X} + \varepsilon$

Engel' curves are ussually convex, i.e. their second derivative

$$D''(x) \leq 0 \quad \text{for} \quad x > 0 .$$

For forecasting purposes of demand, special parameters known as elasticities of demand can be derived from the demand function. The elasticity of a given function $y = f(x_1, x_2, \ldots, x_k)$ with respect to the particular independent variable x_i is defined as:

$$\eta_{x_i} = \lim_{\Delta x_i \to 0} \frac{\frac{\Delta y}{y}}{\frac{\Delta x_i}{x_i}} = \lim_{\Delta x_i \to 0} \left(\frac{\Delta y}{\Delta x_i} \cdot \frac{x_i}{y} \right) = \frac{x_i}{y} \quad y'_{x_i} = \frac{x_i}{y} \cdot \frac{\partial y}{\partial x_i}$$

The elasticity η_{x_i} is a measure of the relative change of the function caused by the unit relative(1%) change of the independent variable X_i.

It is clear, when we realize, that for small increment $\Delta x_i \approx 0$ the elasticity

$$\eta_{x_i} \approx \frac{\frac{\Delta y}{y}}{\frac{\Delta x_i}{x_i}}$$

or

$$\frac{\Delta y}{y} \approx \frac{\Delta x_i}{x_i} \cdot \eta_{x_i}$$

When $\frac{\Delta x_i}{x_i} = 1(\%)$, $\frac{\Delta y}{y} \approx \eta_{x_i}$ (in %)

So, the interpretation of the numerical value of the elasticity is as follows: the value of elasticity η_{x_i} expresses the relative (percentage) increment of the function y, caused by the increment of 1 % of the value of independent variable x_i, provided constant other factors (so called "ceteris paribus" condition).

For different types "f" of demand functions, final formulas for elasticities of demand can be obtained. Usually the elasticity of demand is a function of the independent variable x_i.

This can be illustrated by the following linear Engel' function:

$D = \beta x + \alpha$, where D stands for demand for a particular commodity, and x stands for income of consumers.

The formula for the income elasticity of demand based on that demand function is $\eta_{x_i} = \frac{x}{y} \cdot y' = \frac{x}{\beta x + \alpha} \cdot \beta = \frac{\beta x}{\beta x + \alpha}$ (which depends on x).

Let's assume that the value of income is x = 400 (E.Birr) , and D = 0.1 x + 200.

Then, the income elasticity is $\eta_x = \frac{0.1 \cdot 400}{0.1 \cdot 400 + 200} = \frac{40}{240} = 0.17$,

which means, that if the income of consumer increases by 10% (i.e. by 40 E. Birr, reaching 440 E. Birr) the demand for a given commodity will increase by 1.7% only (i.e. from 240 at the value of income 400, up to 244 at the value of income 440 E.Birr).

Only for the power-type demand function, the elasticities do not depend on independent variables, and they are constant.

Let's prove that statement using the following power-type demand function(with two explanatory variables: X_1 - consumers' income, X_2 - price of given commodity): $D = \alpha X_1^{\beta_1} X_2^{\beta_2}$

Let's calculate the income elasticity and the price elasticity of demand.

The income elasticity of demand :

$$\eta_{X_1} = \frac{X_1}{\alpha X_1^{\beta_1} X_2^{\beta_2}} \cdot \alpha \beta_1 X_1^{\beta_1 - 1} X_2^{\beta_2} = \frac{\alpha \beta_1 X_1^{\beta_1} X_2^{\beta_2}}{\alpha X_1^{\beta_1} X_2^{\beta_2}} =$$

$$= \beta_1$$

The price elasticity of demand:

$$\eta_{X_2} = \frac{X_2}{\alpha X_1^{\beta_1} X_2^{\beta_2}} \cdot \alpha X_1^{\beta_1} \beta_2 X_2^{\beta_2 - 1} =$$

$$= \frac{\alpha \beta_2 X_1^{\beta_1} X_2^{\beta_2}}{\alpha X_1^{\beta_1} X_2^{\beta_2}} = \beta_2$$

So, for the power-type demand function elasticities of demand are constant and equal to the appropriate power parameters.
It worth mentioning, that the income elasticity of demand is usually positive (and higher for luxury goods), while the price elasticity of demand is usually negative.

Example:

The power-type demand function for butter was estimated as:
$D = 0.3\, X_1^{0.8} X_2^{-0.5}$, where X_1 stands for income of consumers and X_2 stands for price of butter.
So, the income elasticity of demand for butter is $\eta_{X_1} = 0.8$, and it means, that an increase of the consumers' income by 10% will cause an increment of demand for butter by 8%.
The price elasticity of demand for butter is $\eta_{X_2} = -0.5$, and it means that an increase of the price for butter by 10% will cause the decrease of demand for butter by 5%.
The estimation of econometric demand functions and their elasticities for main food commodities is the main task of econometric modelling for forecasting of demand and for policy making purposes.
Numerical results of estimation of the price elasticity and the income elasticity of demand for some food commodities based on households

consumption survey data in Poland are given in the table below:

Commodity	Price Elasticity	Income Elasticity
Flour	- 1.09	0.53
Bread	- 0.24	0.05
Meat	- 0.29	0.71
Butter	- 0.37	1.18
Vegetables	- 0.38	0.12
Fruits	- 0.74	0.84
Sugar	- 0.33	0.52
Cakes	- 0.42	2.48
Beer	- 0.75	0.43
Cigarettes	- 0.43	0.66
Vodka	- 1.36	1.14

It is worth noticing, that such basic food commodities like bread or vegetables have very low price and income elasticities of demand.

15.2. Econometric Production Functions.

The second main application of the single-equation econometric models is an analysis of production processes by means of econometric production functions.

The econometric production function is defined as a stochastic equation which reflects the statistical relationship existing between the level of production (output) and different variables being the production factors (inputs) in the given process of production.

A general model of the econometric production function can be written as:

$$Y = f(X_1, X_2, \ldots, X_k, \varepsilon),$$

where Y is the scale (level) of production of a given product, $X_1, X_2, \ldots, X_k$ are different inputs (production factors), ε is the random disturbance.

One can enumerate various production factors, like:

(i) manpower (labour or employment),

(ii) capital assets (machinery installed and used in the process of production),

(iii) raw materials used in production,

(iv) tools and instruments,

(v) energy (or fuel)

At the macro-economy level usually only two main production factors are used as explanatory variables in the production function: capital (K) and labour (L) like in basic Cobb-Douglas production function, which has the power-type form: $Y = \alpha_0 K^{\alpha_1} L^{\alpha_2}$

The sum of parameters α_1 and α_2 (elasticities) in the Cobb-Douglas production function i.e. $r = \alpha_1 + \alpha_2$ is known as returns to scale parameter often assumed to be equal to 1.

To discuss the real meaning of the returns to scale parameter $r = \alpha_1 + \alpha_2$ in the Cobb-Douglas production function let's consider two different time points "t" and "t + 1" for the C-D production function written as

$$Y_t = \alpha_0 X_{t1}^{\alpha_1} X_{t2}^{\alpha_2}$$

Let's assume that each production factor (i.e. capital X_1 and Labour X_2) increases by the same percentage "p" during the period of time from t to t+1.

So, at the time point t+1, the level of production will be

$$Y_{t+1} = \alpha_0 \left[X_{t1} \left(1 + \frac{p}{100}\right)\right]^{\alpha_1} \cdot \left[X_{t2} \left(1 + \frac{p}{100}\right)\right]^{\alpha_2} =$$

$$= \alpha_0 X_{t1}^{\alpha_1} X_{t2}^{\alpha_2} \left(1 + \frac{p}{100}\right)^{\alpha_1 + \alpha_2} = Y_t \left(1 + \frac{p}{100}\right)^r .$$

Therefore the dynamic index of production is equal to

$$\frac{Y_{t+1}}{Y_t} = \left(1 + \frac{p}{100}\right)^r$$

Now it is clear, that:

(a) if $r = 1$, $\frac{Y_{t+1}}{Y_t} = 1 + \frac{p}{100}$, which means that the increase of production is also by p% , i.e. we have the constant effect (productivity) of production factors (inputs),

(b) if $r < 1$, $\frac{Y_{t+1}}{Y_t} < 1 + \frac{p}{100}$, which means that the increment of production is smaller(in percentage) than "p", i.e. we have the decreasing effect (productivity) of inputs,

(c) if $r > 1$, $\frac{Y_{t+1}}{Y_t} > 1 + \frac{p}{100}$, which means larger than p% increase of production, i.e. we have the increasing productivity of inputs.

So, the returns to scale parameter $r = \alpha_1 + \alpha_2$ in the Cobb-Douglas production function points the type of productivity of inputs.
For developing countries, it is rather typical situation that the returns to scale parameter $r > 1$, therefore during estimation of the C-D production function in those countries the prior condition $r=1$ should be often rejected.
The econometric production function can be estimated based on cross-sectional statistical data as well as on time-series data.
In the latter case, the C-D production function is often modified by introducing the dynamic factor(trend) $e^{\gamma t}$ proposed by J. Tinbergen for measurement of effects of the neutral technological progress.
The modified, dynamic Cobb-Douglas production function can be written as:

$$Y_t = \alpha_0 X_{t1}^{\alpha_1} X_{t2}^{\alpha_2} e^{\gamma t}$$

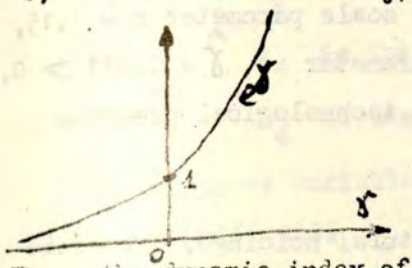
where γ is the trend parameter, which can be interpreted also from the effects of the neutral technological progress viepoint as follows:
Let's assume that between two different time points "t" and "t+1"

was nor structural change neigher increase of production factors (inputs) X_1 and X_2.

So, for t+1 we have $Y_{t+1} = \alpha_0 X_{t+1,1}^{\alpha_1} X_{t+1,2}^{\alpha_2} e^{\gamma(t+1)} =$

$= \alpha_0 X_{t1}^{\alpha_1} X_{t2}^{\alpha_2} e^{\gamma t} \cdot e^{\gamma} =$

$= Y_t e^{\gamma}$



Then, the dynamic index of production is equal to $\frac{Y_{t+1}}{Y_t} = e^{\gamma}$, and

(i) if $\gamma = 0$, $\frac{Y_{t+1}}{Y_t} = 1$ (constant production level)

(ii) if $\gamma > 0$, $\frac{Y_{t+1}}{Y_t} > 1$ (additional increment of production is observed, which must be due to the effect of neutral technological progress in the production process)

(iii) if $\gamma < 0$, $\frac{Y_{t+1}}{Y_t} < 1$ (decrease of production is observed, which means negative effects of technology, or regress)

So, an estimate of the trend parameter γ in the dynamic Cobb-Douglas production function gives us a useful measure of effects of the neutral technological progress in the process of production.
It is worth mentioning that in econometric literature many other types of production functions were also proposed, like:
the CES(Constant Elasticity of Substitution) function:

$$Y = \gamma [(1-\delta) X_1^{-\rho} + \delta X_2^{-\rho}]^{-\frac{\mu}{\rho}}$$

or the Newman-Read production function:

$$Y = \beta_0 X_1^{\beta_1} X_2^{\beta_2} e^{\alpha \ln X_1 \cdot \ln X_2}$$

Examples:

1. Based on time-series data for the Polish national economy, the following dynamic Cobb-Douglas production function was estimated

$$\hat{Y}_t = 3.44\, X_{t1}^{0.42}\, X_{t2}^{0.73}\, e^{0.031\, t}$$

where Y stands for the GDP , X_1 stands for Capital, X_2 stands for Labour(employment)

We can notice here that the returns to scale parameter r = 1.15, and since the estimate of the trend parameter is $\hat{\gamma} = 0.031 > 0$, a small positive effect of the neutral technological progress was observed.

2. Based on cross-sectional data (agricultural holdings) from some agricultural sub-region in Poland, the following agricultural production function was estimated:

$$\hat{Y} = 3929\, X_1^{0.162}\, X_2^{0.171}\, X_3^{9.278}\, X_4^{0.054}\, X_5^{0.081}$$

where Y is the value of agricultural production (crop production and livestock)

X_1 is the area under cultivation in an agricultural holding

X_2 is the manpower in an agricultural holding

X_3 is the livestick in an agricultural holding

X_4 is the cost of farm implements

X_5 is the cost of fertilizers applied

3. At the end of paragraph 8.3 we gave the example number 2 (on page 96) which in fact is an example of the agricultural linear production function (for wheat production) estimated for Arsi region of Ethiopia based on cross-sectional data from agricultural holdings. Let us remind those estimation results:

$$\hat{y} = 11.7\, x_1 + 2.82 x_2 + 1.62 x_3 - 3.39 x_4 + 2.72$$

y is the production of wheat in a private peasant holding

where x_1 is the area under wheat in a peasant holding

x_2 is the use of improved seed

x_3 is the fertilizer applied

x_4 is the crop damage observed in a peasant holding

The manpower variable was removed from the model since it was not significant.

CHAPTER 16. INTRODUCTION TO SIMULTANEOUS EQUATIONS MODELS.

16.1. Classification of Variables.

Up to now we discussed single-equation models for explanation some single economic phenomenon.

If the macroscale i.e. the whole national economy is under consideration the econometric modelling is much more complicated mainly because of interdependency between different variables, which leads to a dynamic multi-equational models. Because of that complexity and dynamics of macro-econometric models there is a need for the special classification of variables used in multi-equational models.

The classification system of variables for simultaneous equations models is as follows:

A) Explained and Explanatory variables.

This classification was used by us also before, since it utilizes equation by equation approach, i.e. for each equation separately a variable is explained (if it is on the left-hand side of the equation) or it is explanatory (if it is on the right-hand side of the equation).

B) Endogenous and Exogenous variables.

This classification takes into account the whole model viewpoint. The variable is called the endogenous if exists an equation in the model in which that variable is explained variable (however in other equations it can be used as explanatory variable). The variable is called exogenous if it is used in all equations only as explanatory variable.

C) Current and Lagged variables.

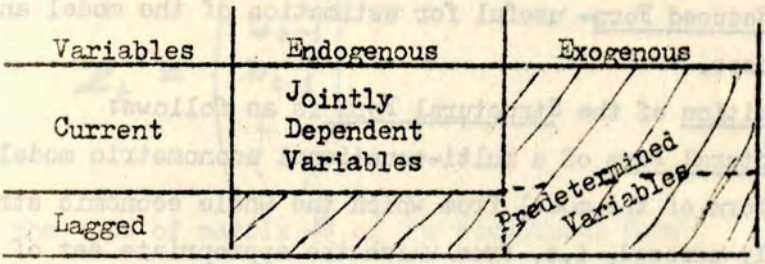
This classification considers the time(dynamics) viewpoint. A variable is called current if its value is from the current year "t", and a variable is called lagged if its value is from the past periods t-1, t-2, etc.

D) Jointly Dependent and Predetermined variables.

This classification is obtained by some kind of intersection of (B) and (C). The jointly dependent variables are the current endogenous variables only.

The predetermined variables are all exogenous variables and lagged endogenous variables.

That last classification can be illustrated by the following scheme:



Examples:

Let's consider 3-equations macro-econometric model with the following variables:

I_t - production investment in the current year t

L_t - employment in the production in the year t

D_t - national income (GDP) in the year t

t - time variable

The simultaneous equations model is:

$$\begin{cases} I_t = \beta_{11} D_t + \beta_{12} I_{t-1} + \beta_{10} + \varepsilon_{1t} \\ L_t = \beta_{21} L_{t-1} + \beta_{22} I_t + \beta_{23} t + \beta_{20} + \varepsilon_{2t} \\ D_t = \beta_{31} D_{t-1} + \beta_{32} I_t + \beta_{33} I_{t-1} + \beta_{34} (L_t - L_{t-1}) + \beta_{30} + \varepsilon_{3t} \end{cases}$$

Except the time variable t and "1" (for constant terms) which are

only the exogenous variables, all others variables are the endogenous.
The jointly dependent variables are: I_t , L_t , D_t .
The predetermined variables are: I_{t-1} , L_{t-1} , D_{t-1} , t, "1" .

16.2. Structural Form and Reduced Form of the Model.

There are two main forms of the multi-equations econometric models:
The Structural Form - used usually for economic analysis of the model, and the Reduced Form- useful for estimation of the model and forecasting purposes.
The definition of the Structural Form is as follows:
The Structural Form of a multi-equational econometric model is such a natural form of the model from which the whole economic structure is clearly exposed, i.e. from which the appropriate set of explanatory variables can be easy defined for each equation(for each explained variable).
It is useful to use the matrix notation for writing down the structural form of the model having G equations.
Let's denote by:

$\mathcal{Y}_t$ - (Gx1) vector of all jointly dependent variables of the model
$\mathcal{B}$ - (GxG) matrix of parameters standing by jointly dependent variables (shifted to the left-hand side of equations)
$\mathcal{Z}_t$ - (Kx1) vector of all predetermined variables of the model
Γ - (G xk) matrix of parameters standing by predetermined variables (shifted to the left-hand side of equations)
$\mathcal{E}_t$ - (Gx1) vector of disturbances in all equations of the model

Then, the Structural Form of the model can be written as:

$$\mathcal{B}\,\mathcal{Y}_t + \Gamma\,\mathcal{Z}_t = \mathcal{E}_t$$

Example:

The model of 3 equations given as an example in the previous

paragraph 16.1 was written in the structural form with $G = 3$, $K=5$,

$$y_t = \begin{pmatrix} I_t \\ L_t \\ D_t \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} 1 & 0 & -\beta_{11} \\ -\beta_{22} & 1 & 0 \\ -\beta_{32} & -\beta_{34} & 1 \end{pmatrix}, \quad \Gamma = \begin{pmatrix} -\beta_{12} & 0 & 0 & 0 & -\beta_{10} \\ 0 & -\beta_{21} & 0 & -\beta_{23} & -\beta_{20} \\ -\beta_{33} & \beta_{34} & -\beta_{31} & 0 & -\beta_{30} \end{pmatrix}$$

$$\varepsilon_t = \begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \\ \varepsilon_{3t} \end{pmatrix}, \quad z_t = \begin{pmatrix} I_{t-1} \\ L_{t-1} \\ D_{t-1} \\ t \\ 1 \end{pmatrix}$$

According to the type of matrix $\mathbf{B}$ of the structural form $\mathbf{B}y_t + \Gamma z_t = \varepsilon_t$, 3 groups of models can be defined: (a) Simple models, (b) Recursive models, (c) Models with interdependent equations.

(a) Simple models.

A multi-equations econometric model is called the Simple model, if its matrix $\mathbf{B}$ of the structural form is a diagonal matrix. It means, that it is no dependency at all between the jointly dependent variables in the model, because in each equation its explained variable depends on some predetermined variables only.

(b) Recursive models.

A model is called the Recursive model, if its matrix $\mathbf{B}$ of the structural form can be written (eventually after re-ordering its columns) as a triangular matrix. It means, that only one-way dependency between the jointly dependent variables is observed in the model, and that dependency has a recursive character, i.e.

Y_1 depends on predetermined variables only,

Y_2 depends on predetermined variables and on Y_1 only,

Y_3 depends on predetermined variables, on Y_1 and on Y_2 only, etc.

(c) Models with interdependent equations.

A model is called to be with interdependent equations if its matrix $\mathcal{B}$ can not be written neither as a diagonal nor a triangular matrix. It means that we have in the model two-ways relationship (a fit-back) between the jointly dependent variables.

That type of econometric model is most often used in practice at the national economy level, since it reflects existing real interdependency between economic variables in the national economy.

But from the other side, the models with interdependent equations are very troublesome for estimation because of so-called identification problem.

The identification problem is prior to the estimation problem and in order to explain it we have to introduce the second form of the model i.e. the Reduced form, defined as follows:

The Reduced Form of the multi-equational econometric model is such a form of the model, in which equations reflect the direct relationship between jointly dependent variables and predetermined variables only.

The reduced form of the model can be obtained from the structural form of the model, by solving the system of equations with respect to the jointly dependent variables(vector $\mathcal{Y}_t$).

Assuming that matrix $\mathcal{B}$ of the structural form is non-singular($\det\mathcal{B}\neq 0$), the structural form of the given model can be transformed into the reduced form in the following way:

$$\mathcal{B}\mathcal{y}_t + \Gamma\mathcal{z}_t = \mathcal{\varepsilon}_t \;\Big|\; \cdot\mathcal{B}^{-1} \text{(left-hand side multiplication)}$$

$$\underbrace{\mathcal{B}^{-1}\mathcal{B}}_{I}\mathcal{y}_t + \mathcal{B}^{-1}\Gamma\mathcal{z}_t = \mathcal{B}^{-1}\cdot\mathcal{\varepsilon}_t$$

$$\mathcal{y}_t = -\mathcal{B}^{-1}\Gamma\mathcal{z}_t + \mathcal{B}^{-1}\cdot\mathcal{\varepsilon}_t$$

or $$\underbrace{\mathcal{y}_t = \Pi\mathcal{z}_t + \mathcal{u}_t}$$ where, $\Pi = -\mathcal{B}^{-1}\Gamma$, $\mathcal{u}_t = \mathcal{B}^{-1}\mathcal{\varepsilon}_t$

The last equation gives the reduced form of the model.

Example:

Famous Keynes' consumption model is defined as:

$$\begin{cases} C_t = \alpha + \beta Y_t + \varepsilon_t \\ Y_t = C_t + I_t \end{cases}$$

where

C_t - consumption at the national level in the year t

Y_t - national income (GDP) in the year t

I_t - investment in the national economy in the year t

ε_t - random disturbance term

This is in fact the structural form of the model and it can be written in the matrix notation as $\mathbf{B}\mathbf{y}_t + \Gamma \mathbf{z}_t = \boldsymbol{\varepsilon}_t$, where $\mathbf{y}_t = \begin{pmatrix} C_t \\ Y_t \end{pmatrix}$, $\mathbf{B} = \begin{pmatrix} 1 & -\beta \\ -1 & 1 \end{pmatrix}$, $\mathbf{z}_t = \begin{pmatrix} 1 \\ I_t \end{pmatrix}$,

$$\Gamma = \begin{pmatrix} -\alpha & 0 \\ 0 & -1 \end{pmatrix}, \quad \boldsymbol{\varepsilon}_t = \begin{pmatrix} \varepsilon_t \\ 0 \end{pmatrix}$$

Since matrix $\mathbf{B}$ is neither diagonal nor triangular, the Keynes' model can be classified as the model with interdependent equations.

In order to obtain the reduced form $\mathbf{y}_t = \Pi \mathbf{z}_t + \mathbf{u}_t$ of the model, first of all the inverse matrix $\mathbf{B}^{-1}$ should be computed.

$\mathbf{B} = \begin{pmatrix} 1 & -\beta \\ -1 & 1 \end{pmatrix}$, $\det \mathbf{B} = 1 - \beta \neq 0$ if $\beta \neq 1$ (that assumption means that not full GNP is consumed).

$$\mathbf{B}^{-1} = \frac{1}{1-\beta} \begin{pmatrix} 1 & 1 \\ \beta & 1 \end{pmatrix}^T = \begin{pmatrix} \frac{1}{1-\beta} & \frac{\beta}{1-\beta} \\ \frac{1}{1-\beta} & \frac{1}{1-\beta} \end{pmatrix}$$

$$\Pi = -\mathbf{B}^{-1}\Gamma = -\begin{pmatrix} \frac{1}{1-\beta} & \frac{\beta}{1-\beta} \\ \frac{1}{1-\beta} & \frac{1}{1-\beta} \end{pmatrix} \cdot \begin{pmatrix} -\alpha & 0 \\ 0 & -1 \end{pmatrix} = \begin{pmatrix} \frac{\alpha}{1-\beta} & \frac{\beta}{1-\beta} \\ \frac{\alpha}{1-\beta} & \frac{1}{1-\beta} \end{pmatrix}$$

$$u_t = B^{-1} \cdot \varepsilon_t = \begin{pmatrix} \frac{\varepsilon_t}{1-\beta} \\ \frac{\varepsilon_t}{1-\beta} \end{pmatrix}$$

So, the reduced form of the Keynes' consumption model is as follows:

$$\begin{pmatrix} C_t \\ Y_t \end{pmatrix} = \begin{pmatrix} \frac{\alpha}{1-\beta} & \frac{\beta}{1-\beta} \\ \frac{\alpha}{1-\beta} & \frac{1}{1-\beta} \end{pmatrix} \cdot \begin{pmatrix} 1 \\ I_t \end{pmatrix} + \begin{pmatrix} \frac{\varepsilon_t}{1-\beta} \\ \frac{\varepsilon_t}{1-\beta} \end{pmatrix}$$

$$\text{or} \begin{cases} C_t = \frac{\alpha}{1-\beta} + \frac{\beta}{1-\beta} I_t + \frac{\varepsilon_t}{1-\beta} \\ Y_t = \frac{\alpha}{1-\beta} + \frac{1}{1-\beta} I_t + \frac{\varepsilon_t}{1-\beta} \end{cases}$$

Obviously, the reduced form is not so interesting from the economic analysis viewpoint as the structural form, but it is more useful for estimation. However, the structual form which is essential for analysis can not be estimated directly, because explanatory variables are correlated with the disturbance term, as we can see for the Keynes' model given above. The first equation of its structural form is $C_t = \alpha + \beta Y_t + \varepsilon_t$. From the reduced form we have, that

$$Y_t = \frac{\alpha}{1-\beta} + \frac{1}{1-\beta} I_t + \frac{\varepsilon_t}{1-\beta} \text{. So, } E(Y_t) = \frac{\alpha}{1-\beta} + \frac{1}{1-\beta} I_t,$$

$$\text{and} \quad Y_t - E(Y_t) = \frac{\varepsilon_t}{1-\beta}$$

So, the covariance between the explanatory variable Y_t and the disturbance ε_t in the consumption (C_t) equation is:

$$\text{COV}(Y_t, \varepsilon_t) = E\left\{\left[Y_t - E(Y_t)\right] \cdot \varepsilon_t\right\} = E\left[\frac{\varepsilon_t}{1-\beta} \cdot \varepsilon_t\right] = \frac{1}{1-\beta} E(\varepsilon_t^2) =$$

$$= \frac{\sigma^2}{1-\beta} \neq 0,$$

therefore Y_t and ε_t are correlated.

Since the structural form of the model can not be estimated directly we can estimate it using the reduced form, which is easy to estimate (because the reduced form can be considered as the simple model). Then having an estimate of the matrix Π of the reduced form we would like to obtain estimates for matrices B and Γ (parameters in the structural form). But that is possible only if the model is identified.

16.3. Identification of the model

A simultaneous equations model written in the structural form is identified, when its equations differ one from another in such a way that we can distinguish between their parameters. In other words, the model is identified, if exists a reverse transformation from parameters in the matrix Π of the reduced form into parameters in matrices B and Γ of the structural form of the model.
If the model is not identified we can not estimate its parameters at all, as it is illustrated by the following market model:

$$\begin{cases} D_t = \alpha P_t + \beta + \varepsilon_{1t} \\ P_t = \gamma S_t + \delta + \varepsilon_{2t} \\ S_t = D_t \end{cases}$$

, where D_t - demand for a given commodity
P_t - price of that commodity
S_t - supply (production) of that commodity

Assuming the market equilibrium $S_t = D_t$, we can rewrite the model as:

$$\begin{cases} D_t = \alpha P_t + \beta + \varepsilon_{1t} \\ P_t = \gamma D_t + \delta + \varepsilon_{2t} \end{cases}$$

From the second (price) equation we have

$$D_t = \frac{1}{\gamma} P_t - \frac{\delta}{\gamma} - \frac{\varepsilon_{2t}}{\gamma}, \qquad \gamma \neq 0.$$

while the first (demand) equation is

$$D_t = \alpha P_t + \beta + \varepsilon_{1t}$$

Now, it is easy to realize, that the structure of these two (price and demand) equations is the same, and therefore one can not distinguish between their parameters (i.e. between $\frac{1}{\delta}$ and α, as well as between $-\frac{\delta}{\gamma}$ and β).

It follows, that even for a large sample we can not estimate those parameters(e.g. we do not know if the estimate of the regression coefficient is valid for $\frac{1}{\delta}$ of for α).

We may conclude, that the above market model is not identified and it needs to be reformulated in such a way, that finally it would be identified (and therefore statistically estimated).

There exist special theorems for checking(prior to estimation) if the model is identified or not. Those theorems concern solution of the system of equations $-B^{-1}\Gamma = \Pi$ with respect to parameters involved in matrices B and Γ (assuming given matrix Π of the reduced form). Using those theorems we can check the identification of each equation separately. The whole model is identified if all its equations are identified.

It follows from the definition of identification, that simple models and recursive models are always identified, so the problem of identification is essential only for models with interdependent equations.

The two useful theorems for checking the identification of a given equation(out of G interdependent equations model) are as follows:

Theorem 1.

A necessary and sufficient condition for the given equation out of the full (G interdependent equations) model to be identified is the rank condition $r(A) = G-1$, where matrix A is constructed from parameters standing by variables excluded in the given equation,but included in other equations.

Theorem 2.

A necessary condition for the given equation (out of G interdependent equations) to be identified is the inequality

$$1 \geqslant G - 1$$

where l is the number of variables excluded from the given equation.
Based on those two theorems the following conclusions can be drawn:

1° If $l < G-1$, the given equation is not identified.

2° If $l \geqslant G-1$, the given equation can be identified or not, depends on the rank of the matrix $\mathcal{A}$ i.e on $r(\mathcal{A})$

3° If $r(\mathcal{A}) = G-1$, and at the same time $l = G-1$, the given equation is called as just identified (or exactly identified), which means, that there exists the unique reverse transformation from the matrix Π of the reduced form into appropriate vectors β and γ of the matrix $\mathcal{B}$ and Γ of the structural form.

4° If $r(\mathcal{A}) = G-1$, and at the same time $l > G-1$, the given equation is called as over-identified, which means that exist many (not the unique) solutions of the system of equations $\Pi = -\mathcal{B}^{-1}\Gamma$ (or $\mathcal{B}\Pi = -\Gamma$) with respect to elements of matrices $\mathcal{B}$ and Γ of the structural form.

The proper technique applied for estimation of parameters of the structural form depends on which type of identification exists for the model.
For the simpler case, when all equations of the model are just identified, the Indirect Least Squares (ILS) method of estimation can be applied, based on the unique solution of the equations $\mathcal{B}\Pi = -\Gamma$ with respect to $\mathcal{B}$ and Γ (see the next paragraph 17.1).
For models with over-identified equations we can use the Two-Stage Least Squares (2 SLS) which is described in the paragraph 17.2.

Example:
Let's check the identification of the first equation of Keynes' consumption model:

$$\begin{cases} C_t = \alpha + \beta Y_t + \varepsilon_t \\ Y_t = C_t + I_t \end{cases}$$

The complete list of variables is here: C_t, Y_t, I_t, 1.
In the first equation only I_t is excluded, so $l=1 = G-1$,
which means that the first equation can be identified.
Matrix $\mathcal{A} = \{-1\}$, $r(\mathcal{A})=1 = G-1$, and therefore we conclude that the first equation is just identified.

CHAPTER 17. ESTIMATION OF MULTI-EQUATIONS MODELS.

17.1. Indirect Least Squares Method (ILS).

If the whole multi-equations model is just (exactly) identified (i.e. all its equations are just identified) we can estimate its structural form parameters by so called Indirect Least Squares Method(ILS), which is based on two facts:

(i) The reduced form of the model can be always estimated based on given statistical data,since it is a simple model always identified.

(ii) Because the model is just identified, exists the unique reverse transformation from the reduced form parameters Π to the parameters B and Γ of the structural form.

So, a general idea of the Indirect Least Squares Method for estimation of the just identified econometric models is rather simple: firstly to estimate the reduced form of the model and then to solve the system of equation $B\Pi=-\Gamma$ in order to obtain the unique estimates $\hat{\Pi}$ and $\hat{\Gamma}$ indirectly from the given estimates $\hat{\Pi}$.

The main steps of work in the ILS method are as follows:

1° Checking, if the model is really just identified and obtaining the reduced form $y_t = \Pi z_t + u_t$.

2° Collecting the statistical data i.e. empirical observations in the form of matrix Y_{Gxn} for all jointly dependent variables, and matrix Z_{Kxn} for all predetermined variables.

3° Apply the OLS method for obtaining estimates $\hat{\Pi}$ of the reduced form parameters.

4° Solving the system of equations $\hat{B}\hat{\Pi}=-\hat{\Gamma}$ with respect to $\hat{B}$ and $\hat{\Gamma}$, assuming given $\hat{\Pi}$.

The ILS estimators of the structural form parameters B and Γ are

usually biased but they are asymptotically unbiased and consistent. It is not easy to obtain the standard errors (SE) of the ILS estimates, since the relationship between the estimates of Π and B, Γ is usually non-linear. Therefore only some approximate method for getting those standard errors can be applied.

Example 1.

Let's discuss the estimation of the Keynes' consumption model.

Its structural form is: $\begin{cases} C_t = \alpha + \beta Y_t + \varepsilon_t \\ Y_t = C_t + I_t \end{cases}$ where C_t - consumption, Y_t - GDP, I_t - investment

The reduced form of the model is:

$$\begin{pmatrix} C_t \\ Y_t \end{pmatrix} = \begin{pmatrix} \frac{\alpha}{1-\beta} & \frac{\beta}{1-\beta} \\ \frac{\alpha}{1-\beta} & \frac{1}{1-\beta} \end{pmatrix} \cdot \begin{pmatrix} 1 \\ I_t \end{pmatrix} + \begin{pmatrix} \frac{\varepsilon_t}{1-\beta} \\ \frac{\varepsilon_t}{1-\beta} \end{pmatrix}$$

Let's check the identification of the first consumption equation (since the unknown parameters α and β of the structural form are in the first equation only)

One variable I_t is excluded from the first equation, therefore $1 = 1 = G-1$. Since the matrix $A = \{-1\}$ and its rank $r(A) = 1 = G-1$ it follows (see Theorems 1 and 2 in paragraph 16.3) that the first equation of the model is just identified.

So, we can use the ILS method for estimation its unknown parameters α and β.

The matrix Π of the reduced form is:

$$\Pi = \begin{pmatrix} \frac{\alpha}{1-\beta} & \frac{\beta}{1-\beta} \\ \frac{\alpha}{1-\beta} & \frac{1}{1-\beta} \end{pmatrix} = \left(\begin{array}{c|c} \Pi_{11} & \Pi_{12} \\ \hline \Pi_{21} & \Pi_{22} \end{array}\right)$$

with properties: $\Pi_{11} = \Pi_{21}$ and $\Pi_{12} + 1 = \Pi_{22}$.

Because of these identities, we need only 2 equations to obtain the ILS estimates for α and β .

Let's assume, that the reduced form of the model has been estimated by the OLS method and we obtain estimates $\hat{\pi}_{11}$, $\hat{\pi}_{12}$, $\hat{\pi}_{21}$, $\hat{\pi}_{22}$ of the elements of matrix π .

Now we shall solve the system of 2 equations:

$$\begin{cases} \dfrac{\hat{\alpha}}{1-\hat{\beta}} = \hat{\pi}_{11} \\ \dfrac{1}{1-\hat{\beta}} = \hat{\pi}_{22} \end{cases}$$ with respect to $\hat{\alpha}$ and $\hat{\beta}$.

and we obtain the ILS estimates as: $\hat{\alpha} = \dfrac{\hat{\pi}_{11}}{\hat{\pi}_{22}} = \dfrac{\hat{\pi}_{11}}{\hat{\pi}_{12}+1}$

$$\hat{\beta} = 1 - \frac{1}{\hat{\pi}_{22}} = \frac{\hat{\pi}_{12}}{\hat{\pi}_{12}+1}$$

Assuming, that the OLS estimate of the reduced form consumption equation is $\hat{C}_t = \underset{\hat{\pi}_{11}}{40} + \underset{\hat{\pi}_{12}}{4} I_t$,

we obtain the following ILS estimates for α and β:

$$\hat{\alpha} = \frac{40}{4+1} = 8 \quad , \hat{\beta} = \frac{4}{4+1} = 0.8$$

So, the consumption equation in the structural form of the model has been estimated by the ILS method as: $\hat{C}_t = 8 + 0.8Y$

Example 2.

In paragraph 16.3 we gave the following market model as an example of non-identified model(D_t - demand, P_t - price, S_t - supply)

$$\begin{cases} D_t = \alpha P_t + \beta + \varepsilon_{1t} \\ P_t = \gamma S_t + \delta + \varepsilon_{2t} \\ S_t = D_t \text{ (equilibrium)} \end{cases} \quad \text{or} \quad \begin{cases} D_t = \alpha P_t + \beta + \varepsilon_{1t} \\ P_t = \gamma D_t + \delta + \varepsilon_{2t} \end{cases}$$

The 2-equational model is really not identified, since all variables are included in each equation ($1 = 0 < 1 = G-1$).

Let's modify the first equation(adding the new variable Z_t - consumer's income) in order to obtain an identified model.

So, the new model is :

$$\begin{cases} D_t = \alpha P_t + \beta Z_t + \varepsilon_{1t} \\ P_t = \gamma D_t + \delta + \varepsilon_{2t} \end{cases}$$

Now, for the first equation we have $l=1=G-1$, $A=\{-\delta\}$, $r(A)=1$ if $\delta \neq 0$.

For the second equation is $l=1=G-1$, $A=\{-\beta\}$, $r(A)=1=G-1$ if $\beta \neq 0$.

If follows, that the new modified model is just identified and it can be estimated by the ILS method.

The structural form of the model is $By_t + \Gamma z_t = \varepsilon_t$, where

$$y_t = \begin{pmatrix} D_t \\ P_t \end{pmatrix} \quad B = \begin{pmatrix} 1 & -\alpha \\ -\gamma & 1 \end{pmatrix} \quad z_t = \begin{pmatrix} z_t \\ 1 \end{pmatrix}, \Gamma = \begin{pmatrix} -\beta & 0 \\ 0 & -\delta \end{pmatrix}, \quad \varepsilon_t = \begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{pmatrix}$$

$$\det B = 1 - \alpha\gamma \neq 0 \quad \text{if } \alpha \neq \frac{1}{\gamma}$$

$$B^{-1} = \begin{pmatrix} \frac{1}{1-\alpha\gamma} & \frac{\alpha}{1-\alpha\gamma} \\ \frac{\gamma}{1-\alpha\gamma} & \frac{1}{1-\alpha\gamma} \end{pmatrix}$$

The reduced form of the model is $y_t = \Pi z_t + u_t$, where

$$\Pi = B^{-1}\Gamma = \begin{pmatrix} \frac{\beta}{1-\alpha\gamma} & \frac{\alpha\delta}{1-\alpha\gamma} \\ \frac{\beta\gamma}{1-\alpha\gamma} & \frac{\delta}{1-\alpha\gamma} \end{pmatrix}$$

$$u_t = B^{-1}\varepsilon_t = \begin{pmatrix} \frac{\varepsilon_{1t} + \alpha\varepsilon_{2t}}{1-\alpha\gamma} \\ \frac{\gamma\varepsilon_{1t} + \varepsilon_{2t}}{1-\alpha\gamma} \end{pmatrix}$$

Assuming, that the parameters of the reduced form have been estimated by the OLS method as $\hat{\Pi} = \begin{pmatrix} \hat{\pi}_{11} & \hat{\pi}_{12} \\ \hat{\pi}_{21} & \hat{\pi}_{22} \end{pmatrix}$ the ILS estimates for the structural form parameters $\alpha, \beta, \gamma, \delta$ can be obtained by solving the following system of equations:

$$\frac{\beta}{1-\alpha\gamma} = \hat{\pi}_{11}$$

$$\frac{\beta\gamma}{1-\alpha\gamma} = \hat{\pi}_{21}$$

$$\frac{\alpha\delta}{1-\alpha\gamma} = \hat{\pi}_{12}$$

$$\frac{\delta}{1-\alpha\gamma} = \hat{\pi}_{22}$$

The solution which gives the ILS estimates of the structural form parameters is:

$$\hat{\alpha} = \frac{\hat{\pi}_{12}}{\hat{\pi}_{22}}, \qquad \hat{\beta} = \pi_{11}\left(1 - \frac{\hat{\pi}_{12} \cdot \hat{\pi}_{21}}{\hat{\pi}_{11} \cdot \hat{\pi}_{22}}\right)$$

$$\hat{\gamma} = \frac{\hat{\pi}_{21}}{\hat{\pi}_{11}}, \qquad \hat{\delta} = \pi_{22}\left(1 - \frac{\hat{\pi}_{12} \cdot \hat{\pi}_{21}}{\hat{\pi}_{11} \cdot \hat{\pi}_{22}}\right)$$

17.2. The Two-Stage Least Squares Method.

In econometric practice such a simultaneous equations models which are over-identified(i.e. $r(A) = G-1$ and $1 > G-1$) are very often in use. For such a models many estimation methods have been proposed in econometric literature, but the most often used method was proposed by H. Theil and it is called: The Two-Stage Least Squares (2-SLS) method.

It is based on the following two assumptions:

1° Each equation out of G equations of the model is estimated separately.

2^{o} For each particular m-th equation it is assumed, that one parameter $\beta_{mm} = 1$, i.e. it is known in advance, and that in this m-th equation only G^{*} jointly dependent variables and K^{*} predetermined variables were included.

In order to explain the application of the 2-SLS method to the single (let say-the first) equation of the G-equational model, we shall introduce the following symbols:

γ_1 - vector(nx1) of observations on the jointly dependent variable Y_1 explained by the first equation of the model,

Y_2 - matrix $(n \times (G^{*}-1))$ of observations on the remaining $G^{*}-1$ jointly dependent variables included in the first equation,

β_2 - vector $((G^{*}-1)x1)$ of parameters by those $G^{*}-1$ jointly dependent variables in the first equation,

Z_* - matrix (nxK^{*}) of observations on K^{*} predetermined variables included in the first equation,

γ_* - vector$(K^{*}x1)$ of parameters by those K^{*} predetermined variables in the first equation,

ε_1 - vector (nx1) of random disturbances in the first equation,

Z - matrix (nxK) of observations on all K predetermined variables included in all equations.

Now, for n sample observations, the first equation which is under estimation can be written as:

$$(*) \quad y_1 = -Y_2\beta_2 - Z_*\gamma_* + \varepsilon_1$$

A general idea of the Two-Stage Least Squares (2 SLS) method of estimation of the given first equation consists of an estimation(by the OLS) of the reduced form $Y_2 = Z\Pi + U$ in order to obtain estimates $\hat{Y}_2$ which can replace Y_2 in the estimated equation (*), and then again apply the OLS method to the equation (*).

It is clear now, that the name of the 2-SLS method is taken from the double use of the OLS method: firstly to the auxiliary reduced form $Y_2 = Z\Pi + U$, and secondly to the estimated first equation (*) of the structural form in which Y_2 has been replaced by its estimates $\hat{Y}_2$ obtained in the first stage.
One can distinguish 3 main steps in the 2-SLS method.

1° Application of the OLS method for estimation the auxiliary reduced form $Y_2 = Z\Pi + U$ in order to obtain estimates $\hat{Y}_2$.

2° Replacement of Y_2 in the first equation (*) by its OLS estimates $\hat{Y}_2$, using formulas $Y_2 - \hat{Y}_2 = V$ (residuals matrix) and $Y_2 = \hat{Y}_2 + V$.

3° Application of the OLS method again to the first equation(*) for getting estimates of its parameters β_2 and γ_*.

The theoretical background for the main trick used in the 2-SLS method i.e. for replacement of Y_2 by its OLS estimates $\hat{Y}_2$ from the estimated reduced form $Y_2 = Z\Pi + U$ lies in utilization of the following Theorem for the OLS estimates:

Theorem.

If the linear regression model $y = X\beta + \varepsilon$ (with classic assumptions on disturbances) was estimated by the OLS method, then the residuals $\hat{\varepsilon}$ are uncorrelated with the estimated values $\hat{y}$ of the explained variable..

Based on the above Theorem applied to the auxiliary (reduced form) model $Y_2 = Z\Pi + U$ estimated by the OLS we may conclude, that residuals V and the estimated values $\hat{Y}_2$ are uncorrelated. Because of that, it is worth to replace Y_2 by its estimates $\hat{Y}_2$ in the estimated first equation (*) of the structural form of the model. Because $Y_2 = \hat{Y}_2 + V$, such replacement gives the following form of the estimated first equation (*):

$$y_1 = -(\hat{Y}_2 + V)\beta_2 - Z_* \gamma_* + \varepsilon_1 \qquad (*)$$

or $$y_1 = -\hat{Y}_2 \beta_2 - Z_* \gamma_* + (\varepsilon_1 - V\beta_2)$$

or $$y_1 = -(Y_2 - V \,\vdots\, Z_*)\begin{pmatrix}\beta_2\\ \gamma_*\end{pmatrix} + (\varepsilon_1 - V\beta_2) \qquad (**)$$

Now, because of above theorem, $\hat{Y}_2$ and V are uncorrelated and therefore the OLS method can be applied to the new form(**) of the first equation in order to estimate parameters β_2 and γ_*.

By application of the OLS estimator to the linear model (**) we obtain finally the 2-SLS estimator as:

$$\begin{pmatrix}\hat{\beta}_2\\ \hat{\gamma}_*\end{pmatrix} = -\begin{bmatrix}(Y_2 - V)^T(Y_2 - V) & (Y_2 - V)^T Z_* \\ Z_*^T (Y_2 - V) & Z_*^T Z_*\end{bmatrix}^{-1}\cdot\begin{pmatrix}Y_2^T - V^T\\ Z_*^T\end{pmatrix}\cdot y_1$$

or, in another equivalent form (more useful for computation):

$$\begin{pmatrix}\hat{\beta}_2\\ \hat{\gamma}_*\end{pmatrix} = -\begin{bmatrix}Y_2^T Z (Z^T Z)^{-1} Z^T Y_2 & Y_2^T Z_* \\ Z_*^T Y_2 & Z_*^T Z_*\end{bmatrix}^{-1}\cdot\begin{pmatrix}Y_2^T Z (Z^T Z)^{-1} Z^T y_1\\ Z_*^T y_1\end{pmatrix}$$

Obviously, we can obtain the same 2-SLS estimates $\begin{pmatrix}\hat{\beta}_2\\ \hat{\gamma}_*\end{pmatrix}$ by double utilization of the OLS method i.e. at the first stage to estimate the auxiliary reduced form $Y_2 = Z\Pi + W$ in order to obtain $\hat{Y}_2$, and then at the second stage the OLS method should be used again to estimate the first equation $y_1 = -\hat{Y}_2 \beta_2 - Z_* \gamma_* + (\varepsilon_1 - V\beta_2)$.

That second way of the 2-SLS application is more useful in the case when the computer program for the 2-SLS is not available.
As inputs for computer we need: $[Y_1, Y_2, Z, Z_*]$, and obviously such an input information should be prepared for each estimated equation (out of G equations) separately.
H. Theil proved, that the 2-SLS estimator $\begin{pmatrix}\hat{\beta}_2 \\ \hat{\gamma}_*\end{pmatrix}$ is asymptotically unbiased and consistent.
Its covariance matrix can be estimated by the formula:

$$V\begin{pmatrix}\hat{\beta}_2 \\ \hat{\gamma}_*\end{pmatrix} = s^2 H^{-1}, \quad \text{where } H = \left(\begin{array}{c|c} Y_2 Z(Z^T Z)^{-1} Z^T Y_2 & Y_2^T Z_* \\ \hline Z_*^T Y_2 & Z_*^T Z_* \end{array}\right),$$

$$s^2 = \frac{1}{n - G^* - K^*}\left[Y_1^T Y_1 - (\hat{\beta}_2^T \,|\, \hat{\gamma}_*^T)\left(\begin{array}{c} Y_2^T Z (Z^T Z)^{-1} Z^T Y_1 \\ \hline Z_*^T Y_1 \end{array}\right)\right]$$

Example:

Using the 2-SLS method we would like to estimate the first equation of the following model:

$$\begin{cases} I_t = \beta_{11} D_t + \beta_{12} I_{t-1} + \beta_{10} + \varepsilon_{1t} \\ L_t = \beta_{21} L_{t-1} + \beta_{22} I_t + \beta_{23} t + \beta_{20} + \varepsilon_{2t} \\ D_t = \beta_{31} D_{t-1} + \beta_{32} I_t + \beta_{33} I_{t-1} + \beta_{34} L_t + \varepsilon_{3t} \end{cases}$$

where I_t - investment
L_t - employment
D_t - national income (GDP)
t - time variable

First of all let's check the identification of the first (investment) equation, from which the variables L_t, L_{t-1}, D_{t-1}, t, are excluded.
So, $1 = 4 > 2 = G - 1$

Matrix $A = \begin{pmatrix} 1 & -\beta_{21} & 0 & -\beta_{23} \\ -\beta_{34} & 0 & -\beta_{31} & 0 \end{pmatrix}$, $r(A) = 2 = G-1$, because

$$\begin{vmatrix} -\beta_{21} & 0 \\ 0 & -\beta_{31} \end{vmatrix} = \beta_{21} \cdot \beta_{31} \neq 0 \quad \text{if} \quad \beta_{21} \neq 0 \,, \; \beta_{31} \neq 0 \,.$$

It follows that the first equation is over-identified $(r(A) = G-1, \; l > G-1)$

For estimation of the first equation by the 2-SLS we shall have the statistical data in form of (y_1, Y_2, Z, Z_*).

Let's assume that the time-series data for the jointly dependent variables I_t, L_t, D_t are as follows:

t	I_t	L_t	D_t
0	7.0	31.1	36.6
1	7.4	31.9	38.1
2	8.1	32.8	39.3
3	8.6	33.7	40.5
4	9.0	34.6	41.8
5	9.5	35.6	43.0
6	10.2	36.6	44.7
7	11.0	37.7	46.0
8	12.4	38.8	48.1
9	13.5	40.0	50.4
10	14.8	41.1	52.6

For the estimated first (investment) equation

$$I_t = \beta_{11} D_t + \beta_{12} I_{t-1} + \beta_{10} + \varepsilon_{1t}$$

we have:

$$Y_1 = \begin{pmatrix} I_t \\ 7.4 \\ 8.1 \\ 8.6 \\ 9.0 \\ 9.5 \\ 10.2 \\ 11.0 \\ 12.4 \\ 13.5 \\ 14.8 \end{pmatrix}, \quad Y_2 = \begin{pmatrix} D_t \\ 38.1 \\ 39.3 \\ 40.5 \\ 41.8 \\ 43.0 \\ 44.7 \\ 46.0 \\ 48.1 \\ 50.4 \\ 52.6 \end{pmatrix}, \quad Z = \begin{pmatrix} I_{t-1} & L_{t-1} & D_{t-1} & t \\ 7.0 & 31.1 & 36.6 & 1 \\ 7.4 & 31.9 & 38.1 & 2 \\ 8.1 & 32.8 & 39.3 & 3 \\ 8.6 & 33.7 & 40.5 & 4 \\ 9.0 & 34.6 & 41.8 & 5 \\ 9.5 & 35.6 & 43.0 & 6 \\ 10.2 & 36.6 & 44.7 & 7 \\ 11.0 & 37.7 & 46.0 & 8 \\ 12.4 & 38.8 & 48.1 & 9 \\ 13.5 & 40.0 & 50.4 & 10 \end{pmatrix}, \quad Z_* = \begin{pmatrix} I_{t-1} \\ 7.0 \\ 7.4 \\ 8.1 \\ 8.6 \\ 9.0 \\ 9.5 \\ 10.2 \\ 11.0 \\ 12.4 \\ 13.5 \end{pmatrix}$$

The 2-SLS method can be applied here by means of performing two times estimation by the OLS method.

At the first stage, the auxiliary reduced form model $Y_2 = \Pi Z + U$ was estimated by the OLS method as:

$$\hat{D}_t = 0.76\ I_{t-1} + 2.32\ L_{t-1} - 0.20\ D_{t-1} - 0.94\ t - 31.25$$

The numerical values of estimates $\hat{D}_t$ were (for t=1,2,...10):

(38.2 39.1 40.5 41.8 43.0 44.6 46.1 48.1 50.4 52.6).

At the second stage the basic investment equation (after replacement of D_t by $\hat{D}_t$): $I_t = \beta_{11}\hat{D}_t + \beta_{12} I_{t-1} + \beta_{10} + (\varepsilon_{1t} - \hat{u}_t \beta_{11})$, (where $\hat{u}_t = D_t - \hat{D}_t$) was estimated by the OLS again. The numerical results of that (final) estimation are:

$$\hat{I}_t = 0.275\ D_t + 0.493\ I_{t-1} - 6.59 \quad , \quad \text{with } R^2 = 0.994 .$$

CHAPTER 18. FORECASTING BASED ON MULTI-EQUATIONS MODELS.

18.1. Use of the Reduced Form for Forecasting.

The structural form $\mathbf{B}\mathbf{y}_t + \mathbf{\Gamma}\mathbf{z}_t = \boldsymbol{\varepsilon}_t$ of the model with interdependent equations (see paragraph 16.2) can not be used for forecasting purposes since it contains as explanatory variables also endogenous variables which are themself under forecasting. Therefore the reduced form of the model $\mathbf{y}_t = \mathbf{\Pi}\mathbf{z}_t + \mathbf{u}_t$ (which is a simple model) should be used for forecasting of the future values of G jointly dependent variables $\mathbf{y}_t$.

For the random disturbance term $\mathbf{u}_t$ of the reduced form it is assumed, that its mean vector $E(\mathbf{u}_t) = \mathbf{0}$, and its covariance matrix $V(\mathbf{u}_t) = \mathbf{\Psi}$.

In order to get forecast of the vector $\mathbf{y}_T$ for the future period of time T, based on the reduced form of the model we have to estimate the parameters $\mathbf{\Pi}$ of the reduced form and we shall assume (or predict using for example trend functions) the vector $\mathbf{z}_T$ of future values of K predetermined variables.

Since the reduced form $\mathbf{y}_t = \mathbf{\Pi}\mathbf{z}_t + \mathbf{u}_t$ has been classified as a simple model, its parameters $\mathbf{\Pi}$ can be easily estimated by the OLS method, assuming given statistical data in the form of the matrix $\mathbf{Z}_{nxK}$ (observations on K predetermined variables) and the matrix $\mathbf{Y}_{nxG}$ (observations on G jointly-dependent variables).

The OLS estimates of the reduced form parameters $\mathbf{\Pi}$ from all equations of the model can be written as

$$\hat{\mathbf{\Pi}}^T = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{Y}$$

Let's denote by $\mathbf{y}_T$ the forecasted vector of values of G jointly dependent variables for the future period of time T.

Let's denote by $\mathbf{u}_T$ the future random disturbance term with assumptions:

$$E(\mathbf{u}_T) = \mathbf{0}, \qquad V(\mathbf{u}_T) = \mathbf{\Psi}, \qquad E(\mathbf{u}_T \mathbf{u}_t^T) =$$

for all t=1,... n.

So, the forecast $\hat{y}_T$ of the forecasted vector $y_T = \Pi z_T + u_T$ can be obtain from the formula

$$\hat{y}_T = \hat{\Pi} z_T \quad ,$$

where $\hat{\Pi}$ is the OLS estimate of Π

z_T is the vector of given future K values of predetermined variables

In order to obtain the covariance matrix of that vector of forecasts $\hat{y}_T$ it will be useful to write it in the equivalent form as:

$$\hat{y}_T = F^T \hat{\pi} \quad \text{where the matrix } F_{(GK \times G)} \text{ is defined as:}$$

$$F = \begin{pmatrix} z_T & 0 & \cdots & 0 \\ 0 & z_T & \cdots & 0 \\ \vdots & & & \\ 0 & 0 & \cdots & z_T \end{pmatrix} ,$$

and the vector $\hat{\pi}_{(GKx1)}$ is a very long column vector constructed from the OLS estimates $\hat{\pi}_i$ being the row vectors of the matrix $\hat{\Pi}$.

$$\hat{\pi} = \begin{pmatrix} \hat{\pi}_1 \\ \hat{\pi}_2 \\ \vdots \\ \hat{\pi}_G \end{pmatrix}$$

and its covariance matrix is $V(\hat{\pi}) = \Phi_{GKxGK}$ (it is known by the OLS formulas for each vector $\hat{\pi}_i$)

Now, the covariance matrix of the vector of forecasts $\hat{y}_T$ can be obtained as:

$$V(\hat{y}_T) = V(F^T \hat{\pi}) + V(u_T) =$$

$$= F^T \Phi F + \Psi \quad , \text{ where } \Psi = V(u_T)$$

It is easy to realize, that a special case of this formula for $V(\hat{y}_T)$, is the formula given in paragraph 9.4 on page 109 for the variance

of forecast based on one equation model $y = X\beta + \varepsilon$. That variance was $D^2(\hat{y}_T) = \sigma^2 X_{(T)}^T (X^T X)^{-1} X_{(T)} + \sigma^2$, and it can be obtain as the special case of the above formula for $V(\hat{y}_T)$ taking:

$$F = X_{(T)}, \quad \Phi = V(b) = \sigma^2 (X^T X)^{-1}, \quad \Psi = D^2(\varepsilon_T) = \sigma^2.$$

18.2. Ex.Post Analysis of Forecasts.

The analysis of precision of forecasts by means of standard errors of forecasts is called often the Ex-Ante analysis of accuracy of forecasting, since it is performed before obtaining actual realizations y_T of forecasted variable. But, if the forecast had been prepared many years ago, and now the actual realizations of a forecasted variable are available, it is possible to make comparative analysis of forecasts $\hat{y}_T$ and actual realizations y_T of endogenous variables.
Such kind of analysis of the accuracy of forecasting based on comparison of actual statistical data with former forecasts is called the Ex - Post analysis of accuracy of forecasting.
The Ex -Post analysis was a special research interest of H. Theil, who proposed the divergency coefficient and its decomposition for Ex-Post analysis purpose.
The Theil's divergency coefficient is defined as:

$$u^2 = \frac{1}{n} \sum_{i=1}^{n} (P_i - A_i)^2$$

where P_i stands for the forecast (prediction) of a given variable for the i-th year

A_i stands for the actual realization of that variable for i-th year

n is the number of years for which the pairs (P_i, A_i) are available.

It is clear that if $\delta_i = \frac{u}{P_i}$, $\left(\text{where } u = \sqrt{u^2}\right)$ is small (in percentages) the forecast P_i was accurate.

By the special decomposition proposed by H. Theil for the divergency coefficient u^2 , we can obtain additional important information about the model which had been used for forecasting.

Let's denote by $\bar{P} = \frac{1}{n} \sum_{i=1}^{n} P_i$ the mean value of forecasts

$\bar{A} = \frac{1}{n} \sum_{i=1}^{n} A_i$ the mean value of actual realizations

$S_p^2 = \frac{1}{n} \sum_{i=1}^{n} (P_i - \bar{P})^2$ the variance of forecasts

$S_A^2 = \frac{1}{n} \sum_{i=1}^{n} (A_i - \bar{A})^2$ the variance of actual realizations

$r = \frac{\frac{1}{n} \sum_{i=1}^{n} (P_i - \bar{P})(A_i - \bar{A})}{S_P \, S_A}$ the correlation coefficient between forecasts and actual realizations.

Using these symbols we can obtain the following decomposition of the Theil's divergency coefficient:

$$U^2 = \frac{1}{n} \sum_{i=1}^{n} (P_i - A_i)^2 = \frac{1}{n} \sum_{i=1}^{n} \left[(P_i - \bar{P}) - (A_i - \bar{A}) + (\bar{P} - \bar{A}) \right]^2$$

$$= (\bar{P} - \bar{A})^2 + S_p^2 - 2r \, S_p \, S_A + S_A^2 =$$

$$= (\bar{P} - \bar{A})^2 + (S_p - S_A)^2 + 2 \, S_p \, S_A (1-r)$$

Each separate part of that decomposition gives some information and it has its specific name accordingly:

$U_M = \frac{(\bar{P} - \bar{A})^2}{u^2}$ is called the share of biasedness of prediction

$U_m = 0$ if $\bar{P} = \bar{A}$ (empirical unbiased prediction)

If U_m is large, it means that a wrong estimation method was used, since the predictor is biased.

$U_S = \frac{(S_p - S_A)^2}{u^2}$ is called the share of dispersion of prediction

$U_S = 0$ if $S_p = S_A$ (homoscedasticity of prediction)

If U_S is large (because of $S_P^2 > S_A^2$) it means small efficiency of predictor.

$U_c = \frac{2(1-r)S_p S_A}{u^2}$ is called the share of correlation between forecasts and actual realizations

$U_c = 0$ if r=1 (full linear correlation P_i and A_i)

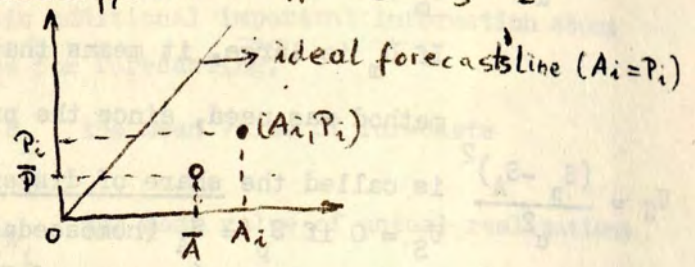
If U_c is large, the correlation P_i and A_i is weak.

Obviously $U_M + U_S + U_c = 1$

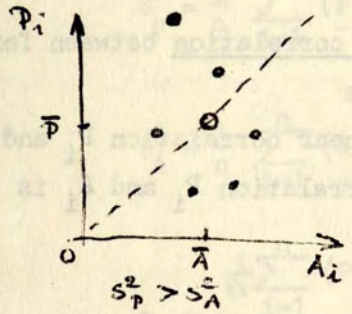
So, if in Ex-Post analysis of prediction we obtained rather large values of $\delta_i = \frac{u}{P_i}$, we can find the main source of that big divergency by looking of the calculated share of biasedness U_M, the share of dispersion U_S and the share of correlation U_c. Such an analysis can give us important information on possible way for improvement of the forecasting model.

For quick Ex-Post analysis with utilization of U_M, U_S, U_C the six typical situations illustrated geometrically are given on the next page.

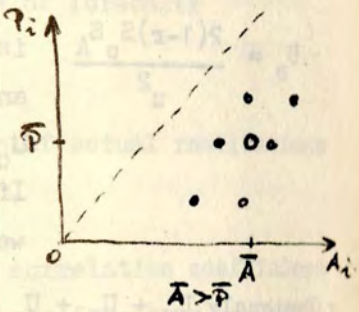
Decomposition of Theil's coefficient: $U_M + U_C + U_S = 1$



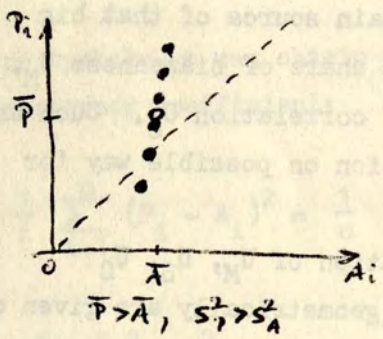
(1) $U_M = 0$, $U_S \neq 0$, $U_C \neq 0$



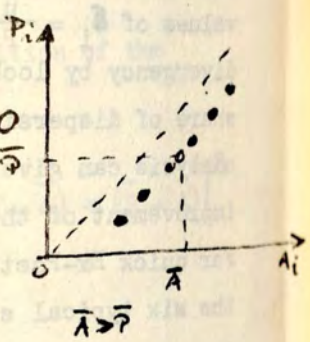
(2) $U_S = 0$, $U_M \neq 0$, $U_C \neq 0$



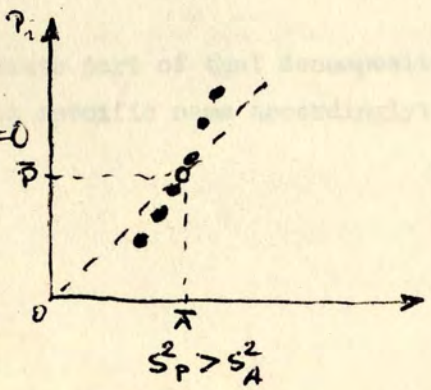
(3) $U_C = 0$, $U_M \neq 0$, $U_S \neq 0$



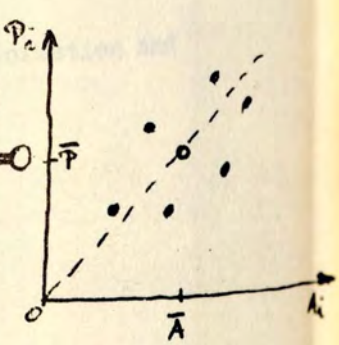
(4) $U_M \neq 0$, $U_S = U_C = 0$



(5) $U_S \neq 0$, $U_M = U_C = 0$



(6) $U_C \neq 0$, $U_M = U_S = 0$



Example:

The production of maize in a given agricultural region was forecasted based on some econometric model and after few years the actual maize production figures (in thousand tons) were obtained:

Year	1	2	3	4	5	Symbol
Forecast	30	35	40	45	50	P_i
Actual Production	34	36	38	42	45	A_i

The Ex-Post analysis of that prediction can now be performed based on decomposition of the Theil's divergeny coefficient.

Numerical results of calculation are as follows:

$$U^2 = \frac{1}{5}\sum_{i=1}^{5}(P_i - A_i)^2 = \frac{1}{5} \cdot 55 = 11 \quad , \quad u = \sqrt{u^2} = 3.3$$

For i=1, $\delta_1 = \frac{u}{P_1} = \frac{3.3}{30} = 0.11 = 11\%$ (which means rather big divergency of prediction).

Mean value of forecasts is $\bar{P} = 40$, while the mean value of actual maize production is $\bar{A} = 39$ (rather small difference).

The Variance of forecasts is $S_P^2 = 50$, $S_P = 7.07$, while the variance of actual production is $S_A^2 = 16$, $S_A = 4.0$

The correlation coefficient between P_i and A_i is high r= 0.99.

The share of biasedness is

$$U_M = \frac{(\bar{P} - \bar{A})^2}{u^2} = \frac{1}{11} = 9\%$$

The share of dispersion is $U_S = \frac{(S_P - S_A)^2}{u^2} = \frac{(7.07 - 4.0)^2}{11} = 86\%$

The share of correlation is $U_c = \frac{2(1-r)\,S_P S_A}{u^2} = \frac{2 \cdot 0.01 \cdot 7.07 \cdot 4.0}{11} = 5\%$

It follows, that the main source of big divergency of prediction is to high variance of forecasts in comparisen with the variance of actual maize production. Therefore we can improve the accuracy of maize production forecasts mainly by degreasing the variance of predictor. This can be done by reestimation of the forecasting econometric model using more efficient estimation method(e.g. using the GLS method instead of the OLS method).

Example

The production of maize in a given agricultural region was forecasted based on some econometric model and after few years the actual maize production figures (in thousand tons) were ascertained:

Year	1	2	3	4	5	Symbol
Forecast	30	35	40	45	50	z_i
Actual Production	34	36	38	42	49	a_i

The ex-post analysis of that prediction can now be performed based on decomposition of the Theil's divergency coefficient.

Numerical results of calculation are as follows:

$$U_d = \frac{1}{n}\sum_{i=1}^{n}(z_i - a_i)^2 = \frac{1}{5}\cdot 55 = 11 \quad , \quad v = \sqrt{U_d} = 3.3$$

for $|m|$, $\delta_1 = \frac{\bar{z}}{\bar{a}} = \frac{40}{39} = 1.026$ (which means rather low divergency of prediction).

Mean value of forecasts is $\bar{z} = 40$, while the mean value of actual maize production is $\bar{a} = 39$ (rather small difference).

The variance of forecasts is $S_z^2 = 50$, $S_z = 7.07$, while the variance of actual production is $S_a^2 = 16$, $S_a = 4.0$.

The correlation coefficient between z_i and a_i is $r = 0.99$.

The share of bias/errors is

$$U_M = \frac{(\bar{z} - \bar{a})^2}{U_d} = \frac{1}{11} = 9\%$$

The share of dispersion is $U_S = \frac{(S_z - S_a)^2}{U_d} = \frac{(7.07 - 4.0)^2}{11} = 0.9\%$

The share of correlation is $U_0 = \frac{2(1-r)\,S_z S_a}{U_d} = \frac{2 \cdot 0.01 \cdot 7.07 \cdot 4.0}{11} = 9\%$

It follows, that the main source of big divergency of prediction is high variance of forecasts in comparison with the variance of actual maize production. Therefore we can improve the accuracy of maize production forecasts mainly by decreasing the variance of predictor.

This can be done by respecification of the forecasting econometric model or more efficient estimation method (e.g. using the GLS instead of the OLS method).