Dr. Andreas Groll

Summer Term 2015

Lecture Notes on Different Aspects of Regression Analysis

Department of Mathematics, Workgroup Financial Mathematics, Ludwig-Maximilians-University Munich, Theresienstr. 39, 80333 Munich, Germany, andreas.groll@math.lmu.de
These lecture notes were written in order to support the students of the graduate course “Different Aspects of Regression Analysis” at the Mathematics Department of the Ludwig Maximilian University of Munich in their first approach to regression analysis.

Regression analysis is one of the most used statistical methods for the analysis of empirical problems in economic, social and other sciences. A variety of model classes and inference concepts exists, reaching from the classical linear regression to modern non- and semi-parametric regression. The aim of this course is to give an overview of the most important concepts of regression and to give an impression of its flexibility. Because of the limited time the different regression methods cannot be explained very detailed, but their overall ideas should become clear and potential fields of application are mentioned. For more detailed information it is referred to corresponding specialist literature whenever possible.
The aim is to model characteristics of a response variable $y$ that is depending on some covariates $x_1, \ldots, x_p$. Most parts of this chapter are based on Fahrmeir et al. (2007). The response variable $y$ is often also denoted as the dependent variable and the covariates as explanatory variables or regressors. All models that are introduced in the following primarily differ in the different types of response variables (continuous, binary, categorial or counting variables) and the different types of covariates (also continuous, binary or categorial).

One essential characteristic of regression problems is that the relationship between the response variable $y$ and the covariates is not given as an exact function $f(x_1, \ldots, x_p)$ of $x_1, \ldots, x_p$, but is overlain by random errors, which are random variables. Consequently, also the response variable $y$ becomes a random variable, whose distribution is depending on the covariates.

Hence, a major objective of regression analysis is the investigation of the influence of the covariates on the mean of the response variable. In other words, we model the (conditional) expectation $E[y|x_1, \ldots, x_p]$ of $y$ in dependency of the covariates. Thus, the expectation is a function of the covariates:

$$E[y|x_1, \ldots, x_p] = f(x_1, \ldots, x_p).$$

Then the response variable can be decomposed into

$$y = E[y|x_1, \ldots, x_p] + \varepsilon = f(x_1, \ldots, x_p) + \varepsilon,$$

where $\varepsilon$ denotes the random variation from the mean, which is not explained by the covariates. Often, $f(x_1, \ldots, x_p)$ is denoted as the systematic component, the random variation $\varepsilon$ is also denoted as stochastic component or error term. So a major objective of regression analysis is the estimation of the systematic component $f$ from the data $y_i, x_{i1}, \ldots, x_{ip}, i = 1, \ldots, n$, and to separate it from the stochastic component $\varepsilon$. 

1

Introduction
1.1 Ordinary Linear Regression

The most famous class is the class of linear regression models
\[ y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon, \]
which assumes that the function \( f \) is linear, so that
\[ E[y|x_1, \ldots, x_p] = f(x_1, \ldots, x_p) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p \]
holds. For the data we get the following \( n \) equations
\[ y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, \ldots, n, \]
with unknown regression parameters or regression coefficients, respectively, \( \beta_0, \ldots, \beta_p \). Hence, in the linear model each covariate has a linear effect on \( y \) and the effects of single covariates aggregate additively. The linear regression model is particularly reasonable, if the response variable \( y \) is continuous and (ideally) approximately normally distributed.

Example 1.1.1 (Munich rent levels). Usually the average rent of a flat depends on some explanatory variables such as type, size, quality etc. of the flat, which hence is a regression problem. We use the so-called net rent as the response variable, which is the monthly rent income, after deduction of all operational and incidental costs. Alternatively, the net rent per square meter (sm) could be used as response variable.

Here, part of the data and variables of the 1999 Munich rent levels are used (compare Fahrmeir et al., 2007). More recent rent level data were either not publicly available or less suitable for illustration. The current rent levels for Munich is available at [http://mietspiegel-muenchen.de](http://mietspiegel-muenchen.de) Table 1.1 contains the abbreviations together with a short description for selected covariates that are used later on in our analysis. The data contain information of more than 3000 flats and have been collected in a representative random sample.

In the following only the flats with a construction year of 1966 or later are investigated. The sample is separated into three parts corresponding to the three different qualities of location. Figure 1.1 shows a scatterplot of the flats with normal location for the response variable rent and the size as the explanatory variable. The scatterplot indicates an approximately linear influence of the size of the flat on the rent:
\[ rent_i = \beta_0 + \beta_1 \cdot size_i + \varepsilon_i. \] (1.1.1)

The error terms \( \varepsilon_i \) can be interpreted as random variations of the straight line \( \beta_0 + \beta_1 \cdot size_i \). As systematic differences from zero are already captured by the parameter \( \beta_0 \), it is assumed that \( E[\varepsilon_i] = 0 \). An alternative formulation of the Model (1.1.1) is
\[ E[\text{rent}|size] = \beta_0 + \beta_1 \cdot size, \]
which means that the expected rent is a linear function of the size of the flat.
## 1.1 Ordinary Linear Regression

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>mean/frequency in %</th>
</tr>
</thead>
<tbody>
<tr>
<td>rent</td>
<td>net rent per month (in DM)</td>
<td>895.90</td>
</tr>
<tr>
<td>rentsm</td>
<td>net rent per month and sm (in DM)</td>
<td>13.87</td>
</tr>
<tr>
<td>size</td>
<td>living area in sm</td>
<td>67.37</td>
</tr>
<tr>
<td>year</td>
<td>year of construction</td>
<td>1956.31</td>
</tr>
<tr>
<td>loc</td>
<td>quality of location estimated by a consultant</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1=normal location</td>
<td>58.21</td>
</tr>
<tr>
<td></td>
<td>2=good location</td>
<td>39.26</td>
</tr>
<tr>
<td></td>
<td>3=perfect location</td>
<td>2.53</td>
</tr>
<tr>
<td>bath</td>
<td>equipment of the bathroom</td>
<td>93.80</td>
</tr>
<tr>
<td></td>
<td>0=standard</td>
<td>6.20</td>
</tr>
<tr>
<td>kit</td>
<td>equipment of the kitchen</td>
<td>95.75</td>
</tr>
<tr>
<td></td>
<td>0=standard</td>
<td>4.25</td>
</tr>
<tr>
<td>ch</td>
<td>central heating</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0=no</td>
<td>10.42</td>
</tr>
<tr>
<td></td>
<td>1=yes</td>
<td>89.58</td>
</tr>
<tr>
<td>dis</td>
<td>district of Munich</td>
<td></td>
</tr>
</tbody>
</table>

**Table 1.1:** Description of the variables of the Munich rent level data in 1999.

![Fig. 1.1: Scatterplot between net rent and size of the flat for flats with a construction year of 1966 or later and normal location (left). Additionally, in the right figure the regression line corresponding to Model (1.1.1) is illustrated.](image)

The example was an application of the *ordinary linear regression model*

\[ y = \beta_0 + \beta_1 x + \varepsilon, \]

or in more general form

\[ y = f(x) + \varepsilon = E[y|x] + \varepsilon, \]
respectively, where the function \( f(x) \) or the expectation \( E[y|x] \) are assumed to be linear, \( f(x) = E[y|x] = \beta_0 + \beta_1 \cdot x \).

In general, for the standard model of ordinary linear regression the following assumptions hold:

\[
y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1.1.2}
\]

where the error terms \( \varepsilon_i \) are independent and identically distributed (iid) with

\[ E[\varepsilon_i] = 0 \quad \text{and} \quad Var(\varepsilon_i) = \sigma^2, \quad \sigma > 0. \]

The property that all error terms have identical variances \( \sigma^2 \) is denoted as homoscedasticity. For the construction of confidence intervals and test statistics it is useful, if additionally (at least approximately) the normal distribution assumption

\[ \varepsilon_i \sim N(0, \sigma^2) \]

holds. Then, also the response variables are (conditionally) normally distributed with

\[ E[y_i|x_i] = \beta_0 + \beta_1 x_i, \quad Var(y_i|x_i) = \sigma^2, \]

and are (conditionally) independent for given covariates \( x_i \).

The unknown parameters \( \beta_0 \) and \( \beta_1 \) can be estimated according to the method of least squares (LS-method). The estimates \( \hat{\beta}_0 \) and \( \hat{\beta}_1 \) are obtained by minimization of the sum of the squared distances

\[
LS(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2,
\]

for given data \( (y_i, x_i), i = 1, \ldots, n \). The method of least squares is discussed in more detail in Section 2.2.1. Putting \( \hat{\beta}_0, \hat{\beta}_1 \) into the linear part of the model, yields the estimated regression line \( \hat{f}(x) = \hat{\beta}_0 + \hat{\beta}_1 x \). The regression line can be considered as an estimate \( E[y|x] \) of the conditional expectation of \( y \), given \( x \), and thus can be used for the prediction of \( y \), which is defined as \( \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \).

**Example 1.1.2 (Munich rent levels - ordinary linear regression).** We illustrate the ordinary linear regression with the data shown in Figure 1.1 using the corresponding Model (1.1.1). A glance on the data raises doubts that the assumption of identical variances \( Var(\varepsilon_i) = Var(y_i|x_i) = \sigma^2 \) is justified, because the variability seems to rise with increasing living area of the flat, but this is initially ignored.

Using the LS-method for the Model (1.1.1) one obtains the estimates \( \hat{\beta}_0 = 253.95, \hat{\beta}_1 = 10.87 \). This yields the estimated linear function

\[ \hat{f}(\text{size}) = 253.95 + 10.87 \cdot \text{size} \]

in Figure 1.1 (on the right). The slope parameter \( \hat{\beta}_1 = 10.87 \) can be interpreted as follows: if the flat size increases by 1 sm, then the average rent increases by 10.87 DM. \( \triangle \)
1.2 Multiple Linear Regression

The standard model of ordinary linear regression from Equation (1.1.2) is a special case \((p = 1)\) of the multiple linear regression model

\[
y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, \ldots, n,
\]

with \(p\) regressors or covariates, respectively, \(x_{i1}, \ldots, x_{ip}\). Here, \(x_{ij}\) denotes the \(j\)-th covariate of observation \(i\), with \(i = 1, \ldots, n\). The covariates can be metric, binary or multicategorical (after suitable encoding). Similar to the ordinary linear regression model new variables can be extracted from the original ones by transformation. Also for the error terms the same assumptions are required. If the assumption of normally distributed error terms holds, then the response variables, given the covariates, are again independent and normally distributed:

\[
y_i | x_{i1}, \ldots, x_{ip} \sim N(\mu_i, \sigma^2),
\]

with \(\mu_i = E[y_i | x_{i1}, \ldots, x_{ip}] = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}\).
Standard Model of Multiple Linear Regression

**Data**

\((y_i, x_{i1}, \ldots, x_{ip}), i = 1, \ldots, n\), for a metric variable \(y\) and metric or binary encoded categorical regressors \(x_1, \ldots, x_p\).

**Model**

\[ y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, \ldots, n. \]

The errors \(\varepsilon_1, \ldots, \varepsilon_n\) are independent and identically distributed (iid) with

\[ E[\varepsilon_i] = 0 \quad \text{and} \quad \text{Var}(\varepsilon_i) = \sigma^2. \]

The estimated linear function

\[ \hat{f}(x_1, \ldots, x_p) = \hat{\beta}_0 + \hat{\beta}_1 x_{1} + \ldots + \hat{\beta}_p x_{p} \]

can be considered as an estimate \(\hat{E}[y|x_1, \ldots, x_p]\) of the conditional expectation of \(y\), given \(x_1, \ldots, x_p\), and can be used for the prediction of \(y\), which is defined as \(\hat{y}\).

The following examples illustrate how flexible the multiple linear regression model is, using suitable transformation and encoding of covariates.

*Example 1.2.1 (Munich rent levels - Rents in normal and good location).* We now incorporate the flats with good location and mark the data points in the scatterplot in Figure 1.2 accordingly. In addition to the regression line for flats with normal location a separately estimated regression line for flats with good location is plotted. Alternatively, one can analyze both location types jointly in a single model resulting in two regression lines that are parallel shifted. The corresponding regression model has the form

\[ rent_i = \beta_0 + \beta_1 size_i + \beta_2 gloc_i + \varepsilon_i. \quad (1.2.1) \]

Here, \(gloc\) is a binary *indicator variable*

\[ gloc_i = \begin{cases} 
1 & \text{if the } i\text{-th flat has good location} \\
0 & \text{if the } i\text{-th flat has normal location.} 
\end{cases} \]

Using the LS-method we obtain the estimated average rent

\[ \hat{rent} = 219.74 + 11.40 \cdot size + 111.66 \cdot gloc. \]

Due to the 1/0-encoding of the location, an equivalent representation is

\[ \hat{rent} = \begin{cases} 
331.40 + 11.40 \cdot size & \text{for good location} \\
219.74 + 11.40 \cdot size & \text{for normal location.} 
\end{cases} \]
Both parallel lines are shown in Figure 1.3. The regression coefficients can be interpreted as follows:

- Both in good and normal location an increase of the living area of 1 sm results in an increase of the average rent of 11.40 DM.
- For flats with the same size the average rent of flats with good location exceeds the rent of corresponding flats with normal location by 111.66 DM.

Fig. 1.2: Left: scatterplot between net rent and size for flats with normal (circles) and good (plus) location. Right: separately estimated regression lines for flats with normal (solid line) and good (dashed line) location.

Fig. 1.3: Estimated regression lines for the Model (1.2.1) for flats with normal (solid line) and good (dashed line) location.
Example 1.2.2 (Munich rent levels - Non-linear influence of the flat size). We now use the variable \( \text{rentsm} \) (net rent per sm) as response variable and transform the size of the flat into \( x = \frac{1}{\text{size}} \). The corresponding model is

\[
\text{rentsm}_i = \beta_0 + \beta_1 \cdot \frac{1}{\text{size}_i} + \beta_2 \cdot \text{gloc}_i + \varepsilon_i. \tag{1.2.2}
\]

The estimated model for average rent per sm is

\[
\hat{\text{rentsm}} = 10.74 + 262.70 \cdot \frac{1}{\text{size}_i} + 1.75 \cdot \text{gloc}.
\]

Both curves for the average rent per sm

\[
\hat{\text{rentsm}} = \begin{cases} 
12.49 + 262.70 \cdot \frac{1}{\text{size}_i} & \text{for good location} \\
10.74 + 262.70 \cdot \frac{1}{\text{size}_i} & \text{for normal location.}
\end{cases}
\]

are illustrated in Figure 1.4. The slope parameter \( \hat{\beta}_1 = 262.70 \) can be interpreted as follows: if the flat size increases by one sm to \( \text{size} + 1 \), then the average rent is reduced to

\[
\text{rentsm} = 10.74 + 262.70 \cdot \frac{1}{\text{size}_i + 1} + 1.75 \cdot \text{gloc}.
\]

\[\triangle\]

Fig. 1.4: Left: Scatterplot between net rent per sm and size for flats with normal (circles) and good (plus) location. Right: Estimated regression curves for flats with normal (solid line) and good (dashed line) location for the Model 1.2.2.
**Example 1.2.3 (Munich rent levels - Interaction between flat size and location).** To incorporate an interaction between the size of the flat and its location into the Model (1.2.1), we define an interaction variable \( \text{inter} \) by multiplication of the covariates \( \text{size} \) and \( \text{gloc} \) with values

\[
\text{inter}_i = \text{size}_i \cdot \text{gloc}_i.
\]

Then

\[
\text{inter}_i = \begin{cases} 
\text{size}_i & \text{if the } i\text{-th flat has good location} \\
0 & \text{if the } i\text{-th flat has normal location},
\end{cases}
\]

and we extend the Model (1.2.1) by incorporating apart from the main effects of \( \text{size} \) and \( \text{gloc} \) also the interaction effect of the variable \( \text{inter} = \text{size} \cdot \text{gloc} \):

\[
\text{rent}_i = \beta_0 + \beta_1 \text{size}_i + \beta_2 \text{gloc}_i + \beta_3 \text{inter}_i + \varepsilon_i. \tag{1.2.3}
\]

Due to the definition of \( \text{gloc} \) and \( \text{inter} \) we get

\[
\text{rent}_i = \begin{cases} 
\beta_0 + \beta_1 \text{size}_i + \varepsilon_i & \text{for normal location} \\
(\beta_0 + \beta_2) + (\beta_1 + \beta_3) \text{size}_i + \varepsilon_i & \text{for good location}.
\end{cases}
\]

For \( \beta_3 = 0 \) no interaction effect is present and one obtains the Model (1.2.1) with parallel lines, i.e. the same slopes \( \beta_1 \). For \( \beta_3 \neq 0 \) the effect of the flat size, i.e. the slope of the straight line for flats with good location, is changed by the value \( \beta_3 \) compared to flats with normal location.

The LS-estimation is not done separately for both location types as in Figure 1.2 (on the right), but jointly for the data of both location types using the Model (1.2.3). We obtain

\[
\hat{\beta}_0 = 253.95, \quad \hat{\beta}_1 = 10.87, \quad \hat{\beta}_2 = 10.15, \quad \hat{\beta}_3 = 1.60,
\]

and both regression lines for flats with good and normal location are illustrated in Figure 1.5. At this point, it could be interesting to check, if the modeling of an interaction effect is necessary. This can be done by testing the hypothesis

\[
H_0 : \beta_3 = 0 \quad \text{versus} \quad H_1 : \beta_3 \neq 0.
\]

How such tests can be constructed will be illustrated during the course.
Fig. 1.5: Estimated regression lines for flats with normal (solid line) and good (dashed line) location based on the Interaction-Model (1.2.3).
This section deals with the conventional linear regression model \( y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \epsilon \) with iid error terms. In the first major part the model is introduced and the most important properties and asymptotics of the LS-estimators are derived in Section 2.2.1 and ???. Next, the properties of residuals and predictions are investigated. Another big issue of this chapter is the conventional test- and estimation theory of the linear model, which is described in Section ??.

2.1 Repetition

This section contains a short insertion recapitulating some important properties of multivariate random variables. For some definitions and properties of multivariate normally distributed random variables, consult Appendix D.1.

Multivariate Random Variables

- Let \( \mathbf{x} \) be a vector of \( p \) (univariate) random variables, i.e. \( \mathbf{x} = (x_1, \ldots, x_p)^\top \). Let \( E[x_i] = \mu_i \) be the expected value of \( x_i \), \( i = 1, \ldots, p \). Then we get

\[
E[\mathbf{x}] = \mathbf{\mu}, \quad \mathbf{\mu} = (\mu_1, \ldots, \mu_p)^\top,
\]

which is the vector of expectations.

- A closed representation of the variation parameters (variances and covariances) of all \( p \) random variables would be desirable. We have

\[
\text{Variance: } \quad \text{Var}(x_i) = E[(x_i - E(x_i))^2] = E[(x_i - E(x_i))(x_i - E(x_i))]
\]

\[
\text{Covariance: } \quad \text{Cov}(x_i, x_j) = E[(x_i - E(x_i))(x_j - E(x_j))]
\]

In general, there are \( p \) variances and \( p(p - 1) \) covariances, altogether \( p + p(p - 1) = p^2 \) parameters that contain information concerning the
variation. These are summarized in the \((p \times p)\)-covariance matrix (also called variance-covariance matrix):

\[
\Sigma := \text{Cov}(\mathbf{x}) = E[(\mathbf{x} - E(\mathbf{x}))(\mathbf{x} - E(\mathbf{x}))^\top]
\]

Example \(p = 2\):

\[
\Sigma = E \left[ \begin{pmatrix} x_1 - E[x_1] \\ x_2 - E[x_2] \end{pmatrix} (x_1 - E[x_1], x_2 - E[x_2]) \right] \\
= E \left[ \begin{pmatrix} x_1 - E[x_1] & x_2 - E[x_2] \\ x_1 - E[x_1] & x_2 - E[x_2] \end{pmatrix} \begin{pmatrix} x_1 - E[x_1] \\ x_2 - E[x_2] \end{pmatrix} \right] \\
= \begin{pmatrix} \text{Var}(x_1) & \text{Cov}(x_1, x_2) \\ \text{Cov}(x_2, x_1) & \text{Var}(x_2) \end{pmatrix}.
\]

- Properties of \(\Sigma\):
  (i) quadratic
  (ii) symmetric
  (iii) positive-semidefinite (recall: a matrix \(A\) is positive-semidefinite \(\iff\) \(x^\top A x \geq 0, \forall x \neq 0\))

### Multivariate Normal Distribution

For more details, see Appendix D.1. The general case:

\[
\mathbf{x} \sim N_p(\mu, \Sigma)
\]

**Remark 2.1.1.** For independent random variables \(\Sigma\) is a diagonal matrix.

Figures 2.1 to 2.4 show the density functions of two-dimensional normal distributions.

**Fig. 2.1:** Density function of a two-dimensional normal distribution for uncorrelated factors, \(\rho = 0\), with \(\mu_1 = \mu_2 = 0, \sigma_1 = \sigma_2 = 1.0\)
Fig. 2.2: Density function of a two-dimensional normal distribution for uncorrelated factors, $\rho = 0$, with $\mu_1 = \mu_2 = 0$, $\sigma_1 = 1.5$, $\sigma_2 = 1.0$

Fig. 2.3: Density function of a two-dimensional normal distribution, $\rho = 0.8$, $\mu_1 = \mu_2 = 0$, $\sigma_1 = \sigma_2 = 1.0$

Fig. 2.4: Density function of a two-dimensional normal distribution, $\rho = -0.8$, $\mu_1 = \mu_2 = 0$, $\sigma_1 = \sigma_2 = 1.0$
2.2 The Ordinary Multiple Linear Regression Model

Let data be given by \( y_i \) and \( x_{i1}, \ldots, x_{ip} \). We collect the covariates and the unknown parameters in the \((p+1)\)-dimensional vectors \( x_i = (1, x_{i1}, \ldots, x_{ip})^\top \) and \( \beta = (\beta_0, \ldots, \beta_p)^\top \). Hence, for each observation we get the following equation

\[
y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i = x_i^\top \beta + \varepsilon_i, \quad i = 1, \ldots, n. \tag{2.2.1}
\]

By definition of suitable vectors and matrices we get a compact form of our model in matrix notation. With

\[
y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{11} & \ldots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \ldots & x_{np} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix},
\]

we can write the Model (2.2.1) in the simpler form

\[
y = X\beta + \varepsilon,
\]

where

- \( y \): response variable
- \( X \): design matrix or matrix of regressors, respectively
- \( \varepsilon \): error term
- \( \beta \): unknown vector of regression parameters

Assumptions:

- \( \varepsilon \sim N(0, \sigma^2 I_n) \), i.e. no systematic error, the error terms are uncorrelated and all have the same variance (homoscedasticity)
- \( X \) deterministic

**Remark 2.2.1.** Usually \( p + 1 \leq n \) holds.

---

### The Ordinary Multiple Linear Regression Model

The model

\[
y = X\beta + \varepsilon
\]

is called ordinary (multiple) linear regression model, if the following assumptions hold:

1. \( E[\varepsilon] = 0 \).
2. \( \text{Cov}(\varepsilon) = E[\varepsilon\varepsilon^\top] = \sigma^2 I_n \).
3. The design matrix \( X \) has full column rank, i.e. \( \text{rk}(X) = p + 1 \).

The model is called ordinary normal regression model, if additionally the following assumption holds:

4. \( \varepsilon \sim N(0, \sigma^2 I_n) \).
2.2 The Ordinary Multiple Linear Regression Model

2.2.1 LS-Estimation

Principle of the LS-estimation: minimize the sum of squared errors

\[ LS(\beta) = \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 = \sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon^T \varepsilon \]  

(2.2.2)

with respect to \( \beta \in \mathbb{R}^{p+1} \), i.e.

\[ \hat{\beta} = \arg \min_{\beta} \sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon^T \varepsilon = (y - X\beta)^T (y - X\beta). \]  

(2.2.3)

Alternatively two other approaches are supposable:

(i) Minimize the sum of errors \( SE(\beta) = \sum_{i=1}^{n} (y_i - x_i^T \beta) = \sum_{i=1}^{n} \varepsilon_i \implies \) Problem: positive and negative errors can eliminate each other and thus the solution of the minimization problem is usually not unique.

(ii) Minimize the sum of absolute errors \( AE(\beta) = \sum_{i=1}^{n} |y_i - x_i^T \beta| = \sum_{i=1}^{n} |\varepsilon_i| \implies \) There is no analytical solving method for the computation of the solution and solving methods are more demanding (e.g. simplex-based methods or iteratively re-weighted least squares are used).

Lemma 2.2.2. Let \( B \) be an \( n \times (p+1) \) matrix. Then the matrix \( B^T B \) is symmetric and positive semi-definite. It is positive definite, if \( B \) has full column rank. Then, besides \( B^T B \), also \( BB^T \) is positive semi-definite.

Theorem 2.2.3. The LS-estimator of the unknown parameters \( \beta \) is

\[ \hat{\beta} = (X^T X)^{-1} X^T y, \]

if \( X \) has full column rank \( p+1 \).

Proof. First, we show that \( \hat{\beta} = (X^T X)^{-1} X^T y \) holds.

According to the LS-approach from Equation (2.2.2) one has to minimize the following function

\[ LS(\beta) = \varepsilon^T \varepsilon = (y - X\beta)^T (y - X\beta) = y^T y - 2\beta^T X^T y + \beta^T X^T X \beta. \]

A necessary condition for a minimum is that the gradient is equal to a vector full of zeros. With the derivation rules for vectors and matrices from Proposition A.0.1 (see Appendix A) we obtain

\[ \frac{\partial LS(\beta)}{\partial \beta} = -2X^T y + 2X^T X \beta = 0 \]

\[ \implies X^T X \hat{\beta} = X^T y \]

\[ \iff \hat{\beta} = (X^T X)^{-1} X^T y. \]
A sufficient condition for a minimum is that the Hesse-matrix (the matrix of the second partial derivatives) has to be positive-semidefinite. In our case we obtain
\[ \frac{\partial^2 \text{LS}(\beta)}{\partial \beta \partial \beta^\top} = 2X^\top X \geq 0, \]
as \( X^\top X \) is positive-semidefinite, so \( \hat{\beta} \) is in fact a minimum.

On the basis of the LS-estimator \( \hat{\beta} = (X^\top X)^{-1}X^\top y \) for \( \beta \) we are able to estimate the (conditional) expectation of \( y \) by
\[ \hat{E}[y] = \hat{y} = X\hat{\beta}. \]
Inserting the formula of the LS-estimator yields
\[ \hat{y} = X(X^\top X)^{-1}X^\top y = Hy, \]
where the \( n \times n \) matrix
\[ H = X(X^\top X)^{-1}X^\top \] (2.2.4)
is called prediction-matrix or hat-matrix. The following proposition summarizes its properties.

**Proposition 2.2.4.** The hat-matrix \( H = (h_{ij})_{1 \leq i,j \leq n} \) has the following properties:

(i) \( H \) is symmetric.

(ii) \( H \) is idempotent (Definition: a quadratic matrix \( A \) is idempotent, if \( AA = A^2 = A \) holds).

(iii) \( rk(H) = tr(H) = p + 1 \). Here, \( tr(\cdot) \) denotes the trace of a matrix.

(iv) \( 0 \leq h_{ii} \leq 1, \forall i = 1, \ldots, n \).

(v) the matrix \( I_n - H \) is also symmetric and idempotent with \( rk(I_n - H) = n - p - 1 \).

**Proof.** → see exercises.

**Remark 2.2.5.** One can even show that \( \frac{1}{n} \leq h_{ii} \leq \frac{1}{r}, \forall i = 1, \ldots, n \), where \( r \) denotes the number of rows in \( X \) that are identical, see for example [Hoaglin and Welsch(1978)]. Hence, if all rows are distinct, one has \( \frac{1}{n} \leq h_{ii} \leq \frac{1}{r}, \forall i = 1, \ldots, n \).

Next, we derive an estimator for \( \sigma^2 \). From a heuristic perspective, as \( E[\varepsilon_i] = 0 \), it seems plausible to use the empirical variance as an estimate: \( \hat{\sigma}^2 = \hat{Var}(\varepsilon_i) = \frac{1}{n} \varepsilon^\top \varepsilon \).
Problem: the vector \( \varepsilon = y - X\beta \) of the true residuals is unknown (because the true coefficient vector \( \beta \) is unknown as well). Solution: we use
the vector $\hat{e} = y - X\hat{\beta}$ of estimated residuals. This estimate is also obtained by the following strategy.

It seems likely to estimate the variance $\sigma^2$ using maximum-likelihood (ML) estimation technique. Under the assumption of normally distributed error terms, i.e. $\varepsilon \sim N(0, \sigma^2 I_n)$, we get $y \sim N(X\beta, \sigma^2 I_n)$ (using Proposition D.0.1 from Appendix D) and obtain the likelihood function

$$L(\beta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp \left( -\frac{1}{2\sigma^2} (y - X\beta)^\top (y - X\beta) \right). \quad (2.2.5)$$

Taking the logarithm yields the log-likelihood

$$l(\beta, \sigma^2) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} (y - X\beta)^\top (y - X\beta). \quad (2.2.6)$$

**Theorem 2.2.6.** The ML-estimator of the unknown parameter $\sigma^2$ is

$$\hat{\sigma}^2_{ML} = \frac{\hat{\varepsilon}^\top \hat{\varepsilon}}{n}, \text{ with } \hat{e} = y - X\hat{\beta}.$$

**Proof.** For the computation of the ML-estimator for $\sigma^2$, we have to maximize the likelihood or the log-likelihood from Equations (2.2.5) and (2.2.6), respectively. Setting the partial derivative of the log-likelihood (2.2.6) with respect to $\sigma^2$ equal to zero yields

$$\frac{\partial l(\beta, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} (y - X\beta)^\top (y - X\beta) = 0.$$

Inserting the LS-estimator $\hat{\beta}$ for $\beta$ into the last equation yields

$$-\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} (y - X\hat{\beta})^\top (y - X\hat{\beta})$$

$$= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} (y - \hat{y})^\top (y - \hat{y}) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \hat{\varepsilon}^\top \hat{\varepsilon} = 0$$

and $\sigma^2 \neq 0$, we hence obtain $\hat{\sigma}^2_{ML} = \frac{\hat{\varepsilon}^\top \hat{\varepsilon}}{n}$. $\square$

Yet, note that this estimator for $\sigma^2$ is only rarely used, because it is biased. This is shown in the following proposition.

**Proposition 2.2.7.** For the ML-estimator $\hat{\sigma}^2_{ML}$ of $\sigma^2$ it holds that

$$E[\hat{\sigma}^2_{ML}] = \frac{n - p - 1}{n} \sigma^2.$$

**Proof.**

$$E[\hat{\varepsilon}^\top \hat{\varepsilon}] = E[(y - X\hat{\beta})^\top (y - X\hat{\beta})]$$

$$= E[(y - X(X^\top X)^{-1}X^\top y)^\top (y - X(X^\top X)^{-1}X^\top y)]$$
\[
E[(y - Hy)^\top(y - Hy)] = E[y^\top(I_n - H)^\top(I_n - H)y] = E[y^\top(I_n - H)y] 
\]
\[
= tr((I_n - H)\sigma^2 I_n) + \beta^\top X^\top(I_n - X(X^\top X)^{-1}X^\top)X\beta 
\]
\[
= \sigma^2(n - p - 1) + \beta^\top X^\top X\beta - \beta^\top X^\top(X^\top X)^{-1}X^\top X\beta 
\]
\[
= \sigma^2(n - p - 1). 
\]

In (*) calculation rule 6 from Theorem D.0.2 for expectation vectors and covariance matrices has been used (see Appendix D).

Hence, immediately an unbiased estimator \( \hat{\sigma}^2 \) for \( \sigma^2 \) can be constructed by:

\[
\hat{\sigma}^2 = \frac{\hat{\varepsilon}^\top \hat{\varepsilon}}{n - p - 1}. 
\]

(2.2.7)

There also exists an alternative representation for this estimator, which is shown next.

**Proposition 2.2.8.** The adjusted estimator from (2.2.7) of the unknown parameter \( \sigma^2 \) can also be written as

\[
\hat{\sigma}^2 = \frac{\hat{\varepsilon}^\top \hat{\varepsilon}}{n - p - 1} = \frac{y^\top y - \hat{\beta}^\top X^\top y}{n - p - 1}, 
\]

with \( \hat{\varepsilon} = y - X\hat{\beta} \).

**Proof.**

\[
\hat{\varepsilon}^\top \hat{\varepsilon} = (y - X\hat{\beta})^\top(y - X\hat{\beta}) 
\]
\[
= (y - X(X^\top X)^{-1}X^\top y)^\top(y - X(X^\top X)^{-1}X^\top y) 
\]
\[
= y^\top(I_n - X(X^\top X)^{-1}X^\top)y 
\]
\[
= y^\top(I_n - X(X^\top X)^{-1}X^\top)y, 
\]

Remark 2.2.9. It can be shown that the estimator (2.2.7) maximizes the marginal likelihood

\[
L(\sigma^2) = \int L(\beta, \sigma^2) d\beta 
\]

and is thus called a restricted maximum-likelihood (REML) estimator.
Proposition 2.2.10. The LS-estimator $\hat{\beta} = (X^T X)^{-1} X^T y$ is equivalent to the ML-estimator based on maximization of the log-likelihood \(2.2.6\).

Proof. This follows immediately, because maximization of the term $-\frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta)$ with respect to $\beta$ is equivalent to minimization of $(y - X\hat{\beta})^T(y - X\beta)$, which is exactly the objective function in the LS-criterion \(2.2.3\).

Nevertheless, for the sake of completeness, we compute the ML-estimator for $\beta$ by maximizing the likelihood or the log-likelihood from Equations \(2.2.5\) and \(2.2.6\), respectively. Setting the partial derivative of the log-likelihood \(2.2.6\) with respect to $\beta$ equal to zero yields

$$\frac{\partial l(\beta, \sigma^2)}{\partial \beta} = -\frac{1}{\sigma^2}(X^T y - X^T X \beta) = 0.$$ 

Similar to the proof of Theorem \(2.2.3\) it follows:

$$-X^T y + X^T X \beta \overset{\perp}{=} 0 \quad \Leftrightarrow \quad X^T X \hat{\beta} = X^T y \quad \Leftrightarrow \quad \hat{\beta} = (X^T X)^{-1} X^T y.$$ 

Parameter estimators in the multiple linear regression model

<table>
<thead>
<tr>
<th>Estimator for $\beta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>In the ordinary linear model the estimator $\hat{\beta} = (X^T X)^{-1} X^T y$ minimizes the LS-criterion $LS(\beta) = \sum_{i=1}^{n} (y_i - x_i^T \beta)^2$. Under the assumption of normally distributed error terms the LS-estimator is equivalent to the ML-estimator for $\beta$.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Estimator for $\sigma^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>The estimate $\hat{\sigma}^2 = \frac{1}{n-p-1} \hat{e}^T \hat{e}$ is unbiased and can be characterized as REML-estimator for $\sigma^2$.</td>
</tr>
</tbody>
</table>
Proposition 2.2.11. For the LS-estimator $\hat{\beta} = (X^\top X)^{-1}X^\top y$ and the REML-estimator $\hat{\sigma}^2 = \frac{1}{n-p-1}\hat{\varepsilon}^\top \hat{\varepsilon}$ the following properties hold:

(i) $E[\hat{\beta}] = \beta, \text{Cov}(\hat{\beta}) = \sigma^2(X^\top X)^{-1}$.

(ii) $E[\hat{\sigma}^2] = \sigma^2$

Proof. (i) The LS-estimator can be represented as

$$\hat{\beta} = (X^\top X)^{-1}X^\top y = (X^\top X)^{-1}X^\top (X\beta + \varepsilon) = \beta + (X^\top X)^{-1}X^\top \varepsilon$$

(and also $\hat{\beta} - \beta = (X^\top X)^{-1}X^\top \varepsilon$ holds).

Now, we are able to derive the vector of expectations and the covariance matrix of the LS-estimator:

$$E[\hat{\beta}] = \beta + (X^\top X)^{-1}X^\top E[\varepsilon] = \beta$$

and

$$\text{Cov}(\hat{\beta}) = E[(\hat{\beta} - E[\hat{\beta}])(\hat{\beta} - E[\hat{\beta}])^\top] = E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)^\top] = E[(X^\top X)^{-1}X^\top \varepsilon \varepsilon^\top X(X^\top X)^{-1}] = (X^\top X)^{-1}X^\top E[\varepsilon \varepsilon^\top]X(X^\top X)^{-1} = \sigma^2(X^\top X)^{-1} = \sigma^2(X^\top X)^{-1}.$$

Altogether we obtain

$$\hat{\beta} \sim (\beta, \sigma^2(X^\top X)^{-1}).$$

(ii) The proof follows directly from Proposition 2.2.7. An alternative proof is based on a linear representation of $\hat{\varepsilon} = y - X\hat{\beta}$ with respect to $\varepsilon$. We have

$$\hat{\varepsilon} = y - X\hat{\beta} = y - X(X^\top X)^{-1}X^\top y = (I_n - X(X^\top X)^{-1}X^\top)y = (I_n - X(X^\top X)^{-1}X^\top)(X\beta + \varepsilon) = X\beta - X(I_n - X(X^\top X)^{-1}X^\top)\varepsilon ((I_n - X(X^\top X)^{-1}X^\top)\varepsilon = 0) = (I_n - X(X^\top X)^{-1}X^\top)\varepsilon \quad \text{(linear function of } \varepsilon) = M\varepsilon,$$

where $M := (I_n - X(X^\top X)^{-1}X^\top)$ is a (deterministic) symmetric and idempotent matrix, see Proposition 2.2.4 (v). Hence, we can write
\[ \hat{\epsilon}^\top \hat{\epsilon} = \epsilon^\top M^\top M \epsilon = \epsilon^\top M \epsilon, \]

and obtain a quadratic form in \( \epsilon \), with other words a scalar. With the help of the trace operator \( tr \) we obtain

\[
E[\hat{\epsilon}^\top \hat{\epsilon}] = E[\epsilon^\top M \epsilon] \\
= E[tr(\epsilon^\top M \epsilon)] \quad (\epsilon^\top M \epsilon \text{ is a scalar!}) \\
= E[tr(M \epsilon \epsilon^\top)] \quad (\text{use } tr(AB) = tr(BA)) \\
= tr(M \epsilon \epsilon^\top) \\
= tr(M \sigma^2 I_n) \\
= \sigma^2 tr(M) \\
= \sigma^2 tr(I_n - X(X^\top X)^{-1} X^\top) \\
= \sigma^2 [tr(I_n) - tr(X(X^\top X)^{-1} X^\top)] \quad (\text{use } tr(A + B) = tr(B) + tr(A)) \\
= \sigma^2 (n - p - 1)
\]

and hence

\[
E[\hat{\sigma}^2] = E \left[ \frac{\hat{\epsilon}^\top \hat{\epsilon}}{n-p-1} \right] = \sigma^2.
\]

\( \square \)
In this chapter we present a short summary with some fundamental properties from Linear Algebra, Analysis and Stochastic, see e.g. Pruscha (2000).

Proposition A.0.1. Let $x \in \mathbb{R}^n$, $b \in \mathbb{R}^n$ and $a \in \mathbb{R}^m$ be vectors of dimension $n$ and $m$, respectively. Then the following derivation rules hold:

$\frac{d}{dx} (Ax) = A$ \hspace{1cm} [A $m \times n$ - matrix]

$\frac{d}{dx} (x^T A) = A$ \hspace{1cm} [A $n \times m$ - matrix]

$\frac{d}{dx} (x^T Ax) = 2Ax$ \hspace{1cm} [A symmetric $n \times n$ - matrix]

$\frac{d^2}{dx^2} (x^T Ax) = 2A$ \hspace{1cm} [A symmetric $n \times n$ - matrix]

$\frac{d}{dx} ( (Ax - a)^T (Ax - a) ) = 2A^T (Ax - a)$ \hspace{1cm} [A $m \times n$ - matrix].

$\frac{d}{dx} (b^T x) = b$.

Proposition A.0.2. Let $A$ be a $n \times n$ and $Q$ be a $n \times m$ matrix. Then:

1. If $A$ is positive semi-definite, then also $Q^T AQ$ is positive semi-definite.
2. If $A$ is positive definite and $Q$ has full column rank, then also $Q^T AQ$ is positive definite.
In this chapter we present a short summary of the most important properties of estimation functions as well as some concepts of estimation theory. A general extensive introduction into the basic concepts of inductive statistics can be found in [Fahrmeir et al. (2007)] or [Mosler and Schmid (2005)].

B.1 Some one-dimensional distributions

**Definition B.1.1 (Gamma-distribution).** A continuous, non-negative random variable $X$ is called gamma-distributed with parameters $a > 0$ and $b > 0$, abbreviated by the notation $X \sim G(a,b)$, if it has a density function of the following form

$$f(x) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp(-bx), \quad x > 0.$$ 

**Lemma B.1.2.** Let $X \sim G(a,b)$ be a continuous, non-negative random variable. Then its expectation and variance are given by:

- $E[X] = \frac{a}{b}$
- $\text{Var}(X) = \frac{a}{b^2}$

**Definition B.1.3 (χ²-distribution).** A continuous, non-negative random variable $X$ with density

$$f(x) = \frac{1}{2^\frac{n}{2} \Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} \exp\left(-\frac{1}{2}x\right), \quad x > 0,$$

is called $\chi^2$-distributed with $n$ degrees of freedom, abbreviated by the notation $X \sim \chi^2_n$.

**Lemma B.1.4.** Let $X \sim \chi^2_n$ be a continuous, non-negative random variable. Then its expectation and variance are given by:
• $E[X] = n$
• $Var(X) = 2n$

Remark B.1.5. The $\chi^2$-distribution is a special gamma-distribution with $a = n/2$ and $b = 1/2$.

Lemma B.1.6. Let $X_1, \ldots, X_n$ be independent and identically standard normally distributed, then

$$Y_n = \sum_{i=1}^{n} X_i^2$$

is $\chi^2$-distributed with $n$ degrees of freedom.

Definition B.1.7 (t-distribution). A continuous random variable $X$ with density

$$f(x) = \frac{\Gamma(n+1)/2}{\sqrt{n\pi} \Gamma(n/2)(1+x^2/n)^{(n+1)/2}}$$

is called t-distributed with $n$ degrees of freedom, abbreviated by the notation $X \sim t_n$.

Lemma B.1.8. Let $X \sim t_n$ be a continuous, non-negative random variable. Then its expectation and variance are given by:

• $E[X] = n$, $n > 1$
• $Var(X) = n/(n-2)$, $n > 2$.

The $t_1$-distribution is also called Cauchy-distribution. If $X_1, \ldots, X_n$ are iid with $X_i \sim N(\mu, \sigma^2)$, it follows that

$$\frac{\bar{X} - \mu}{S} \sqrt{n} \sim t_{n-1},$$

with

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 \quad \text{and} \quad \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Definition B.1.9 (F-distribution). Let $X_1$ and $X_2$ be independent random variables with $\chi^2_n$- and $\chi^2_m$-distributions, respectively. Then the random variable

$$F = \frac{X_1/n}{X_2/m}$$

is called F-distributed with $n$ and $m$ degrees of freedom, abbreviated with the notation $F \sim F_{n,m}$.

Definition B.1.10 (Log-normal distribution). A continuous, non-negative random variable $X$ is called logarithmically normally distributed, $X \sim LN(\mu, \sigma^2)$, if the transformed variable $Y = \log(X)$ is following a normal distribution, $Y \sim N(\mu, \sigma^2)$. The density of $X$ is given by
B.2 Some Important Properties of Estimation Functions

The expectation and the variance yield

\[ E[X] = \exp(\mu + \sigma^2/2) \]
\[ Var(X) = \exp(2\mu + \sigma^2) \cdot (\exp(\sigma^2) - 1). \]

B.2 Some Important Properties of Estimation Functions

**Definition B.2.1.** An estimation function or statistic for a parameter \( \theta \) in the population is a function \( \hat{\theta} = g(X_1, \ldots, X_n) \) of the sample variables \( X_1, \ldots, X_n \). The numerical value \( g(x_1, \ldots, x_n) \) obtained from the realisations \( x_1, \ldots, x_n \) is called estimate.

**Definition B.2.2.** An estimation function \( \hat{\theta} = g(X_1, \ldots, X_n) \) is called unbiased for \( \theta \), if

\[ E_{\theta}[\hat{\theta}] = \theta. \]

It is called asymptotically unbiased for \( \theta \), if

\[ \lim_{n \to \infty} E_{\theta}[\hat{\theta}] = \theta. \]

The bias is defined by

\[ Bias_{\theta}(\hat{\theta}) = E_{\theta}[\hat{\theta}] - \theta. \]

**Definition B.2.3.** The mean squared error is defined by

\[ MSE_{\theta}(\hat{\theta}) = E[(\hat{\theta} - \theta)^2] \]

and can be expressed in the following form

\[ MSE_{\theta}(\hat{\theta}) = Var(\hat{\theta}) + Bias_{\theta}(\hat{\theta})^2. \]

**Definition B.2.4.** An estimation function \( \hat{\theta} \) is called MSE-consistent or simply consistent, if

\[ MSE_{\theta}(\hat{\theta}) \xrightarrow{n \to \infty} 0. \]

**Definition B.2.5.** An estimation function \( \hat{\theta} \) is called weak consistent, if for any \( \varepsilon > 0 \)

\[ \lim_{n \to \infty} P(|\hat{\theta} - \theta| < \varepsilon) = 1 \]

or

\[ \lim_{n \to \infty} P(|\hat{\theta} - \theta| \geq \varepsilon) = 0, \]

respectively.
Central Limiting Value Theorems

For dealing with asymptotic statistical methods one needs definitions, concepts and results for the convergence of sequences of random variables. An extensive overview of the different convergence definitions can be found in [Pruscha (2000)]. There, also references for the corresponding proofs are given. In the following we only present those definitions, which are important for this course.

**Definition C.0.1.** Let the sequence of $p$-dimensional random vectors $X_n, n \geq 1,$ and $X_0$ be given with distribution functions $F_n(x), x \in \mathbb{R}^p,$ and $F_0(x), x \in \mathbb{R}^p,$ respectively. The sequence $X_n, n \geq 1,$ converges in distribution to $X_0,$ if

$$\lim_{n \to \infty} F_n(x) = F_0(x) \quad \forall x \in C_0,$$

with $C_0 \subset \mathbb{R}^p$ denoting the set of all points where $F_0(x)$ is continuous. In this case one also uses the notation $X_n \xrightarrow{d} X_0.$
D

Probability Theory

This chapter contains a short summary of some important results from stochastic and probability theory.

**Proposition D.0.1.** Uni- and multivariate normally distributed random variables can be transformed in the following way:

- **Univariate case:**
  \[ x \sim N(\mu, \sigma^2) \iff ax \sim N(a\mu, a^2\sigma^2) \quad (a \text{ is a constant}). \]

- **Multivariate case:**
  \[ x \sim N_p(\mu, \Sigma) \iff Ax \sim N_q(A\mu, A\Sigma A^\top), \]
  where \( A \) is a deterministic \((q \times p)\)-dimensional matrix.

**Theorem D.0.2.** Let \( \mathbf{x} \) and \( \mathbf{y} \) be random vectors, \( \mathbf{A}, \mathbf{B}, \mathbf{a}, \mathbf{b} \) matrices and vectors, respectively, of suitable dimension, and let \( E[\mathbf{x}] = \mu \) and \( \text{Cov}(\mathbf{x}) = \Sigma \).

Then:

1. \( E[\mathbf{x} + \mathbf{y}] = E[\mathbf{x}] + E[\mathbf{y}] \).
2. \( E[A\mathbf{x} + \mathbf{b}] = A \cdot E[\mathbf{x}] + \mathbf{b} \).
3. \( \text{Cov}(\mathbf{x}) = E[\mathbf{x}\mathbf{x}^\top] - E[\mathbf{x}]E[\mathbf{x}]^\top \).
4. \( \text{Var}(a^\top\mathbf{x}) = a^\top \text{Cov}(\mathbf{x})a = \sum_{i=1}^p \sum_{j=1}^p a_i^2 \sigma_{ij} \).
5. \( \text{Cov}(A\mathbf{x} + \mathbf{b}) = A \text{Cov}(\mathbf{x}) A^\top \).
6. \( E[\mathbf{x}\mathbf{x}^\top] = \text{tr}(A\Sigma) + \mu^\top A\mu \).

**D.1 The Multivariate Normal Distribution**

**Definition D.1.1.** A \( p \)-dimensional random vector \( \mathbf{x} = (x_1, \ldots, x_p)^\top \) is called multivariate normally distributed, if \( \mathbf{x} \) has the density function

\[
 f(\mathbf{x}) = (2\pi)^{-\frac{p}{2}}|\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \mu)^\top \Sigma^{-1}(\mathbf{x} - \mu)\right\}, \quad \text{(D.1.1)}
\]

with \( \mu \in \mathbb{R}^p \) and positive semi-definite \( p \times p \) matrix \( \Sigma \).
Theorem D.1.2. The expectation and the covariance matrix of a multivariate normally distributed vector $\mathbf{x}$ with density function [D.1.1] are given by $E[\mathbf{x}] = \mu$ and $\text{Cov}(\mathbf{x}) = \Sigma$. Hence, we use the notation $\mathbf{x} \sim N_p(\mu, \Sigma)$, similarly to the univariate case. Often the index $p$ is suppressed, if the dimension is clear by context. For $\mu = 0$ and $\Sigma = I$ the distribution is called (multivariate) standard normal distribution.

Theorem D.1.3. Let $\mathbf{X} \sim N_n(\mu, \Sigma)$ be multivariate normally distributed. Now regard a partition of $\mathbf{X}$ into two sub-vectors $\mathbf{Y} = (X_1, \ldots, X_r) \top$ and $\mathbf{Z} = (X_{r+1}, \ldots, X_n) \top$, i.e.

$$\mathbf{X} = \begin{pmatrix} \mathbf{Y} \\ \mathbf{Z} \end{pmatrix}, \quad \mu = \begin{pmatrix} \mu_Y \\ \mu_Z \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_Y & \Sigma_{YZ} \\ \Sigma_{ZY} & \Sigma_Z \end{pmatrix}.$$ 

Then the sub-vector $\mathbf{Y}$ is again $r$-dimensional normally distributed with $\mathbf{Y} \sim N_r(\mu_Y, \Sigma_Y)$. Besides, the conditional distribution of $\mathbf{Y}$ given $\mathbf{Z}$ is also a multivariate normal distribution with expectation

$$\mu_{Y|Z} = \mu_Y + \Sigma_{YZ} \cdot \Sigma_Z^{-1}(\mathbf{Z} - \mu_Z)$$

and covariance matrix

$$\Sigma_{Y|Z} = \Sigma_Y - \Sigma_{YZ} \Sigma_Z^{-1} \Sigma_{ZY}.$$ 

Furthermore, for normally distributed random variables the situations of independency and of being uncorrelated are equivalent: $\mathbf{Y}$ and $\mathbf{Z}$ are independent if and only if $\mathbf{Y}$ and $\mathbf{Z}$ are uncorrelated, i.e. $\Sigma_{YZ} = \Sigma_{ZY} = 0$. For non-normal random variables this equivalence does not hold in general. Here, from independency merely follows uncorrelatedness.

D.1.1 The Singular Normal Distribution

Definition D.1.4. Let $\mathbf{x} \sim N_p(\mu, \Sigma)$. The distribution of $\mathbf{x}$ is called singular, if $\text{rk}(\Sigma) < p$ holds. In such cases, the distribution is often characterized by the precision matrix $\mathbf{P}$ (with $\text{rk}(\mathbf{P}) < p$) instead of the covariance matrix (see e.g. the field of spatial statistics for examples). The random vector $\mathbf{x}$ then has a density function

$$f(\mathbf{x}) \propto \exp \left[ -\frac{1}{2}(\mathbf{x} - \mu) \top \mathbf{P}(\mathbf{x} - \mu) \right],$$

which is only defined up to proportionality.
Theorem D.1.5. Let $\mathbf{x}$ be a $p$-dimensional random vector, which is singularly distributed, i.e. $\mathbf{x} \sim N_p(\mu, \Sigma)$ with $\text{rk}(\Sigma) = r < p$. Let $(GB)$ be an orthogonal matrix, in which the rows of the $p \times r$ matrix $G$ form a basis of the column space of $\Sigma$ and the rows of $B$ a basis of the null-space of $\Sigma$. Consider the transformation

$$
\begin{pmatrix}
y_1 \\
y_2
\end{pmatrix} = (GB)^\top \mathbf{x} =
\begin{pmatrix}
G^\top \mathbf{x} \\
B^\top \mathbf{x}
\end{pmatrix}.
$$

Then $y_1$ is the stochastic part of $\mathbf{x}$ and not singular with

$$y_1 \sim N_r(G^\top \mu, G^\top \Sigma G),$$

$y_2$ is the deterministic part of $\mathbf{x}$ with

$$E[y_2] = B^\top \mu \quad \text{and} \quad \text{Var}(y_2) = 0.$$

The density of the stochastic part $y_1 = G^\top \mathbf{x}$ has the form

$$f(y_1) = \frac{1}{(2\pi)^{\frac{r}{2}}(\prod_{i=1}^{r}(\lambda_i))^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(y_1 - G^\top \mu)^\top (G^\top \Sigma G)^{-1}(y_1 - G^\top \mu)\right\},$$

where $\lambda_i$ denote the corresponding $r$ different non-zero eigen values and $\Sigma^{-}$ is a generalized inverse of $\Sigma$.

D.1.2 Distributions of Quadratic Forms

Distributions of quadratic forms of normally distributed random vectors play an important role for testing linear hypotheses, compare Section ??.

Theorem D.1.6 (Distributions of quadratic forms).

(i) Let $\mathbf{x} \sim N_p(\mu, \Sigma)$ with $\Sigma > 0$. Then:

$$y = (\mathbf{x} - \mu)^\top \Sigma^{-1}(\mathbf{x} - \mu) \sim \chi^2_p.$$

(ii) Let $\mathbf{x} \sim N_p(0, I_p)$, $B$ an $n \times p$ $(n \leq p)$ matrix and $R$ a symmetric, idempotent $p \times p$ matrix with $\text{rk}(R) = r$. Then:

- $\mathbf{x}^\top R \mathbf{x} \sim \chi^2_r$,

- From $BR = 0$ follows that the quadratic form $\mathbf{x}^\top R \mathbf{x}$ and the linear form $B \mathbf{x}$ are independent.

(iii) Let $X_1, \ldots, X_n$ be independent random variables with $X_i \sim N(\mu, \sigma^2)$ and

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2.$$ 

Then:
• $\frac{n-1}{\sigma^2} S^2 \sim \chi^2_{n-1}$.

• $S^2$ and $\bar{X}$ are independent.

(iv) Let $x \sim N_n(0, I_n)$, $R$ and $S$ be symmetric and idempotent $n \times n$ matrices with $rk(R) = r$ and $rk(S) = s$ and $RS = 0$. Then:

• $x^\top Rx$ and $x^\top Sx$ are independent.

• $\frac{s}{r} \frac{x^\top Rx}{x^\top Sx} \sim F_{r,s}$.
References