Vrije Universiteit Amsterdam

Research Paper

Correlation in Linear Regression

Author:
Yura Perugachi-Diaz
Student nr.: 2566305

Supervisor:
Dr. Bartek Knapik

May 29, 2017

Faculty of Sciences
Research Paper Business Analytics

Abstract

Goal: The aim of this research is to investigate and give an overview of several approaches that can deal with the problem that arises when the independence of errors assumption of the linear regression model (LM) is violated.

Approach: Two approaches, generalized least squares (GLS) and linear mixed effect models (LME), are examined to get an understanding of the basic theory and how they manipulate data to handle dependency of errors. In a practical part the approaches are tested on real and simulated data to see how they perform.

Practice: The GLS is tested by simulating data. Therefore a multiple LM with correlated errors is simulated. After the simulation, the ordinary least squares (OLS) and GLS approach are applied and their results are compared. For the LME model a real life dataset containing longitudinal data is used. By comparing models including and excluding random-effects to capture the dependency of errors, their performance is tested.

Results & conclusion: The GLS showed a big improvement in estimating the unknown $\hat{\beta}$-coefficients of the regression equation compared to the OLS. The GLS obtained a regression equation whereby all $\hat{\beta}$-coefficients were estimated really close to the real values and proved to have a significant influence. For the OLS, two of these coefficients were not estimated close to the real values and showed no significance. Thus, the GLS proved that it was not influenced by the dependent errors. For the LME, incorporating random-effects for repeated measurements from the same person over time, showed a significant difference than not incorporating the dependent errors.

Recommendation: For a future research, the simulated dataset could be extended by adding more explanatory variables. The LME performing the best, can be compared with a LM model including the same fixed-effects to see what difference it makes to include random-effects.

Acknowledgements

As part of the masters program Business Analytics, students have to perform an individual research. During this research, the students examine a problem in the field of business mathematics and computer-science. The aim for the students is to perform a research individually whereby their findings are reported in a research paper in a clear and understandable way. The results are presented to their supervisor(s) and possibly business people and students.

Thanks to my supervisor Bartek Knapik I gained a lot more experience in examining a really mathematical topic while still preserving the enthusiasm and curiosity. I would like to thank him for all his guidance and support throughout the whole journey. I would not know if I would have achieved all of this without him.

Furthermore, I would like to give special thanks to my friends and fellow students for their support and help when I got stuck or needed a motivation boost. This meant a lot to me.

Yura Perugachi-Diaz, May 2017

Introduction

In the statistics a well-known and used statistical model is the linear regression model (LM). The LM is used to examine and predict data by modeling the relationship between the dependent, also called *response*, variable and the independent, also called *explanatory*, variables. The aim of the LM is to find the best statistical relationship between these variables in order to predict the response variable or to examine the relationship between the variables.

Before applying the LM, there are several assumptions the data observations need to satisfy to allow the user to use the LM. One important assumption is the *independence* assumption which is satisfied when the observations are taken on subjects that are not related in any sense. In that case the errors of the data can be assumed to be independent. In case this assumption is violated, the errors exist to be dependent and the quality of statistical inference may not follow from the classical theory.

There are several approaches available that can be used to resolve the dependency of errors. Each of these approaches has different ways of handling dependency of errors. Therefore the main question for this research is:

How can one deal with dependency of errors?

To answer the main question an overview will be given of the theoretical background of two available approaches that are able to handle dependency of errors. To see the impact of these approaches, one will create a practical study. In the practical study the performance of the approaches will be investigated on real and simulated data.

The first part of this research paper will present an overview of the theoretical background of the examined models and approaches. Chapter 1 gives a broad explanation of the LM with corresponding model assumptions and how to measure the goodness of fit of the model. In Chapter 2 the theoretical background of the alternative GLS approach can be found followed by the LME in Chapter 3. The second part of this paper contains examples of a practical study. In Chapter 4 the performance of the GLS is investigated on simulated data. Chapter 5 presents an investigation of the LME on real life data. Finally, Chapter 6 will provide a conclusion and recommendation.

Contents

Part I

Models and Approaches

Chapter 1

Linear Regression Model

1.1 Introduction

In statistics a well-known and used statistical model is the linear regression model (LM). This model is used to model statistical relationship between the independent, also called *explanatory*, variables and one dependent, also called *response*, variable. The response expresses the observations of the data. Because observations, are in most cases of stochastic nature rather than deterministic nature, the LM is aimed to try to find the best possible statistical relationship between the observations. In the model the values of the explanatory variables are known and are used to describe the response variable as good as possible. The explanatory variables have the ability to control, change and manipulate the response variable. Thus they control the environment in which one makes measurements. Therefore the response variable is also called the *dependent* variable because it depends on the explanatory variables.

A real life example to apply a LM could be a case where one examines how study time affects the grade of students. Therefore the amount of study time the student spent, is modeled as the explanatory variable whereby the grades of the student are modeled as the response variable. As one can imagine, the more time the students spent studying, the higher their grades will become. In this way the explanatory variable controls, changes and manipulates the response variable. With this data the LM could produce a best-fitting line, also called *regression line*, through the data observations. The regression line describes how many hours a student should spent studying to obtain a certain grade. With this line one can predict what the grade of a new student will be when spending a certain amount of studying time. In this example the best intercept and slope of the regression line can be found with a simple LM by using the *ordinary least squares*, which will be explained later on in this chapter. A simple LM is a model containing only one explanatory variable. When the LM incorporates multiple explanatory variables, the model is called a multiple LM.

As one can see, the LM can come in handy when it comes to searching for relationships between variables and making predictions. But before one can make use of the model, the data needs to fit several model assumptions. If one or more of these assumptions is/are not satisfied, the quality of statistical inference may not follow from the classical theory. In the upcoming sections an overview of the LM is given.

1.2 The Linear Regression Model

The LM assumes a linear relationship between the response variable y_j and the p-multivector of explanatory variables x_i. To describe the relation between these variables, the best-fitting line or (hyper)plane needs to be found. The best-fitting line or (hyper)plane is also called the *regression line* or *regression equation*. With the *ordinary least squares* method, the regression equation can be found. This method computes the intercept and slope of the regression equation. Before explaining this method let us first take a look at the LM. The model is of the following form:

$$\begin{aligned} y_j &= \beta_1 x_{j1} + \cdots + \beta_p x_{jp} + \epsilon_j \\ &= x_j^T \beta + \epsilon_j \qquad\qquad \text{with } j = 1, \cdots, n \end{aligned}$$

where:

$x_j^T = (x_{j1}, \cdots, x_{jp})$ is a $1 \times p$ vector of explanatory variables for the jth observation.

Note: Explanatory variable x_1^T is a vector only containing the values 1 thus $\beta_1 x_1^T = \beta_1$ a vector of length n only containing the β_1 value.

y_j is the response variable for the jth observations

β is a $p \times 1$ vector of unknown parameters

ϵ_j is the unobserved error which represents the difference between the observed response variable y_j and the predicted value $\hat{y}_i$ which is obtained by the systematic part $x_j^T \beta$

In matrix notation the model is written as follows:

$$\mathbf{y} = X\beta + \epsilon \tag{1.2.1}$$

Where in terms of the n data points:

X is the $n \times p$ matrix, also called the *design* matrix. With jth row equal to the p-multivector of explanatory variables x_j^T.

$\mathbf{y}$ is the $n \times 1$ vector $\mathbf{y} = (y_1, \cdots, y_n)^T$

β is the $p \times 1$ vector $\beta = (\beta_1, \cdots, \beta_p)^T$

ϵ is the $n \times 1$ vector $\epsilon = (\epsilon_1, \cdots, \epsilon_n)^T$

In case $p = 2$ the equation (1.2.1) describes a line through the data observations which expresses the relation between the explanatory variable and response variable. Then, β_1 is interpreted as the intercept, and β_2 as the slope of the line. In other cases the equation describes a plane ($p = 3$) or a hyperplane ($p > 3$).

Example Assume the model: $y = \beta_1 x_1 + \beta_2 x_2 = \beta_1 + \beta_2 x_2$.
In this LM the intercept is given by β_1 and the slope is given by β_2.

Because the β-coefficients are usually unknown, these coefficients need be estimated. A method that can estimate $\hat{\beta}$-coefficients is the *least squares method*.

1.2.1 Least squares method

To describe the linear relationship between the response and explanatory variables, one can fit a regression equation whereby the unknown $\hat{\beta}$-coefficients can be found with the *least squares method*. This method estimates the $\hat{\beta}$-coefficients so one can predict the the response variable $\hat{y}_i$. The model describing the regression equation is of the following matrix form:

$$\hat{y} = X\hat{\beta} \tag{1.2.2}$$

where:

X is the *design* matrix containing the explanatory variables. Each column of the matrix corresponds to an explanatory variable.

$\hat{y}$ is a vector containing the predicted response values

$\hat{\beta}$ are the estimated β-coefficients

The *least squares method* which can find the $\hat{\beta}$-coefficients is the *ordinary least squares*.

Ordinary Least Squares The *ordinary least squares* (OLS) can be applied under the condition that the $p \times p$ matrix X is of full rank, and hence $X^T X$ is invertible. Only then, the *least squares estimator* of $\hat{\beta}$ is in matrix notation:

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T \mathbf{y} \tag{1.2.3}$$

The aim of the OLS is to minimize the sum of squares of the errors by fitting a line through all the observations with model (1.2.2). Therefore the errors are calculated. The errors represent the difference between the observed response values y_i and the predicted response values $\hat{y}_i$. The model with the smallest difference between the observed and predicted response values, contains the smallest value of the sum of the squared errors. The $\hat{\beta}$-coefficient obtaining the the smallest value of the sum of the squared errors, will create the regression equation through the observations. These $\hat{\beta}$-coefficients obtaining the least squares of the errors are thus found with (1.2.3).

Maximum Likelihood Another way of estimating the unknown coefficients β is by making use of the *maximum likelihood estimate* (MLE). In case of dealing with a normal LM 1.2.3, which will be explained later on, where the ϵ_j's are mutually independent and normally distributed, the MLE is identical to the OLS [1].

1.2.2 Goodness of fit

After having obtained a (multiple) LM that produces a regression line, it is of importance to examine how well the regression line predicts the actual response values. To examine the performance of a model, one can use different computations to interpret the goodness of fit of a model.

R-squared A statistical measure that expresses the percentage of how good the explanatory variables predict the response variable, is the r-squared (R^2). The R^2 works as follows, take the mean of the actual response values and compute the distance of the actual response values to the mean. Next, take the regression equation with the estimated response values and compute the distance of the estimated response values to the mean. Comparing these distances and subtracting this of 1, expresses how well the regression equation predicts the actual response values. The R^2 is mathematically written as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)} \tag{1.2.4}$$

Example Assuming the student data explained in the introduction, a regression line is fitted through all the observations and one obtains an high $R^2 = 0.876$. This thus means that 87.6% of the predicted response values can be explained by the explanatory variable. Which means that predicting the grade of the students can be explained for 87.6% with explanatory variable, the studying time.

Standard error Another statistical measure to see how good a multiple LM performs, is the standard error. The standard error is a measure that computes the range where one can expect the errors to fall in. The smaller the standard errors, the better one can rely on the estimated $\hat{\beta}$-coefficients. The standard errors can be computed with the C matrix which is a symmetric matrix with on the diagonal, the variance of the estimated $\hat{\beta}$-coefficients, also written as: $diag(C_{11}, \cdots, C_{ii})$. This matrix is computed as follows:

$$\begin{aligned} C &= \hat{\sigma^2}(X^T X)^{-1} \\ &\text{with:} \\ \hat{\sigma}^2 &= \mathbf{y}^T(\mathbf{I} - H)\mathbf{y}(n-p-1)^{-1} \\ &= \mathbf{y}^T(\mathbf{I} - X(X^T X)^{-1}X^T)\mathbf{y}(n-p-1)^{-1} \end{aligned}$$

where $(n-p-1)$ represents the *degrees of freedom*, n is the number of observations and p are the number of variables participating in the model [2]. To compute the standard error, one simply needs to take the square roots of the diagonal values of the C matrix.

$$SE_i = \sqrt{C_{ii}}, \quad i = 1, 2, \cdots \tag{1.2.5}$$

t- and p-values The t-values are obtained by dividing the $\hat{\beta}$-coefficient with the *standard errors* of (1.2.5) and the H_0 hypothesis can be tested and is rejected if:

$$|T_i| = \frac{|\hat{\beta}_i|}{SE_i} \geq t_{(n-p-1);1-\frac{\alpha}{2}} \quad i = 1, 2, \cdots$$

As stated in Bijma [2] the $t_{(n-p-1);1-\frac{\alpha}{2}}$ is the $(1-\frac{\alpha}{2})$-quantile of the t-distribution with $n-p-1$ degrees of freedom. With the corresponding p-value one can test the already mentioned null hypothesis:

$$H_0 : \hat{\beta}_i = 0 \quad \text{vs.} \quad H_1 : \hat{\beta}_i \neq 0 \quad \text{for} \quad i = 1, 2, \cdots$$

This test tests if a $\hat{\beta}$-coefficient has (no-)effect on the model. If a $\hat{\beta}$-coefficient has effect on the model $\hat{\beta}_i \neq 0$, this can be noticed by a low p-value < 0.05. If the p-value < 0.05, the null hypothesis is rejected. This means that the $\hat{\beta}$-coefficient is meaningful to the model because changes in the values of the corresponding explanatory variable are related with changes in the response variable [3].

1.2.3 Normal linear model

Model (1.2.1) is called a *normal linear model* if:

> The errors ϵ_j are mutually independent and are from the normal distribution with $\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$, where $\mathbf{I}$ is an $n \times n$ identity matrix. Then the response variable y_j is an independent normal random variable with mean [4]:
> $\mathbb{E}[y_j] = \mathbb{E}[x_j^T\beta + \epsilon_j] = x_j^T\beta$ and variances: $\mathbb{V}ar[y_j] = \mathbb{V}ar[x_j^T\beta + \epsilon_j] = \sigma^2$.

For the examination of the relationship between the response and predictors in linear regression, one is assuming a normal linear model with errors mutually independent and from the normal distribution.

1.2.4 The assumptions

Before making use of the LM, the data needs to satisfy several assumption. From the point of view of this research paper the most important assumptions are described bellow. For the remaining assumptions it can be useful to look into [4, 5].

Linearity One needs to verify if there is linearity in the parameters. This means that the there needs to be a linear relation between the response and explanatory variables. Thus the y of the model needs to be the result of a linear combination of the explanatory variables plus the error terms [6].

> **Example** Assume a linear model: $y = \beta_1 x_1 + \beta_2 x_2 + \epsilon$. As one can see, there is a linear relationship between y and X and this model would create a straight line.
>
> If the parameters are not linear, the model could for example look like this: $y = \beta_1 x_1 + x_2^{\beta_2} + \epsilon$. As one can see, the parameter β_2 is expressed in such a way that it is not a linear combination anymore. If data consist of such a form, the result would be that the y of the model is obtained by a non-linear combination of the explanatory variables and errors.

Note: Even when an explanatory variable is non-linear, for example a polynomial, the model is still linear in its parameters. This is because one is assuming a linearity in the unknown $\hat{\beta}$-coefficients. This case can be seen in the practical part of this research paper.

independence One of the most important assumption is the *independence* assumption, whereby the errors need to be independent and identically. This assumption is divided into three sub assumptions that needs to be satisfied before applying the LM. Every sub-assumption will be provided with a real life example.

1. Expected value of the error term is equal to zero: This assumption states that the expectation of the errors is equal to zero $\mathbb{E}[\epsilon_i] = 0$. Which means that the error of a subject cannot be predicted from the knowledge of the error for another subject [7].

> **Example** In case the expected value of the error term is equal to zero: Then one knows the time a particular student spent studying and their corresponding grade (thus given one observation). Then this will say nothing about the grade of another student spending the same amount of time studying being bellow or above the mean for the studying time of all students.

Note: From the point of view of this research paper, one will assume that this sub-assumption is satisfied.

2. Homoscedasticity: This assumption states that the error terms of the model need to have a constant variance $\mathbb{V}ar[\epsilon_i] = \sigma^2 \quad \forall i$. Which means that the distribution of the errors stay constant along with the explanatory variable(s) [4]. This is also called *homoscedasticity*.

In case the errors have a constant variance, the corresponding variance-covariance matrix, also known as *covariance* matrix, for the errors would be a diagonal matrix with on the diagonal constant variances. This is also written as: $diag(\sigma^2 \cdots \sigma^2)$. In case the errors appear to have non-constant variance this would correspond to a covariance matrix with unequal variances on the diagonal, also written as: $diag(\sigma_1^2 \cdots \sigma_n^2)$.

If the errors do not have a constant variance along with the explanatory variables, this is known as *heteroscedasticity* and the assumption is violated.

> **Example** *Homoscedasticity* can be seen in a case where one wants to predict the grocery spending of persons, based on their income. People with low/high incomes are spending certain low/high amounts of money on their groceries. In this case the variation in spendings are similar and there is no violation of the assumption.
> In case people with high income spend as little as people with a much lower income, the variation for these persons will be higher and one will observe non-constant variance of the errors. Thus one is then dealing with *heteroscedasticity*.

One way of dealing with heteroscedasticity is by using the *generalized least squares* (GLS). In section 2.2 a broad explanation of this approach is given. Another way of dealing with non-constant variances is with the *linear mixed effect models* (LME), which can be found in chapter 3.

3. No-correlation: Also known as the independence of errors. For this assumption one assumes that there is no-correlation between errors, which means that the errors are independent of each other.

If the errors show no-correlation, the corresponding covariance matrix is a matrix with all the off-diagonal values being equal to zero $\mathbb{E}[\epsilon_i, \epsilon_j] = 0, \forall i \neq j$. If the errors

are correlated the matrix will be a covariance matrix with one or more off-diagonal values being unequal to zero $\mathbb{E}[\epsilon_i, \epsilon_j] \neq 0, \forall i \neq j$ [5].

> **Example** On a school participating students where observed for an experiment. For the experiment the students all took the same mathematics test. They were studying all by themselves and they were not influenced by there fellow students. In this case there would be *no-correlation of errors*. Thus when investigating these data observations, the corresponding covariance matrix only contained off-diagonal values equal zero. Therefore one concluded that there was no-correlation of errors.
> In a similar experiment on another school, some of the observed students where working together for the same test and they helped each other while preparing for the test. When examining these data observations, the corresponding covariance matrix contained off-diagonal values unequal to zero for the students working together. Therefore one concluded that the errors were correlated and the LM assumption was not satisfied.

One way of dealing with correlated errors is by using the GLS. Another way of dealing with correlated errors is with the LME. In the upcoming chapters a theoretical overview will be given.

In the LM one assumes that expected value of the error term is equal to zero, homoscedasticity and no-correlation of errors. Violating this assumption does not mean that there is no model possible for the data, it only means that the linear regression model is inappropriate.

1.2.5 Gauss-Markov theorem

The *Gauss-Markov* theorem (G-M) says that if the errors in a linear regression model are uncorrelated, have expectation zero and have equal variances, then the *best linear unbiased estimator* (BLUE) of the coefficients can be found with the OLS estimator. As stated in Sen and Srivastava [5] the G-M conditions for the errors in a linear regression model are the following:

$$\begin{aligned} \mathbb{E}[\epsilon_i] &= 0 \\ \mathbb{V}ar[\epsilon_i] &= \sigma^2 < \infty \\ \mathbb{C}ov[\epsilon_i, \epsilon_j] &= 0, \qquad \forall i \neq j \end{aligned}$$

In matrix notation these conditions are:

$$\begin{aligned} \mathbb{E}[\epsilon] &= 0, \\ \mathbb{E}\left[\epsilon\epsilon^T\right] &= \sigma^2 I \end{aligned}$$

As one can see, these conditions are the same as the *independence* assumption in subsection 1.2.4.

Thus for example, if the error term is a function of the explanatory variables, which means that the expectation of the error term is not zero anymore $\mathbb{E}[\epsilon_i] \neq 0$, the OLS estimator is biased and therefore is not BLUE anymore.

Note: For the G-M conditions, the errors do not need to be normal and identically distributed.

1.3 Summary

Given data consisting of explanatory variables and a response variable. Whereby the relationship between these variables is linear and the errors are independent, the LM (1.2.1) is a very often used statistical model to predict and estimate the response values. This model fits a regression line through the actual observations. The regression line is a line which best describes the relation between the explanatory variables and response variable. To obtain the line, one needs to estimate the unknown $\hat{\beta}$-coefficients, representing the intercept and slope(s) of the line. If one has data, satisfying the *normal linear model* and the corresponding model assumptions, one can use the *ordinary least squares method* (1.2.3) to estimate the $\hat{\beta}$-coefficients and model the regression line. To see how good the model predicts the actual response values, one can use the R^2, standard errors or p-values to interpret the performance of the model.

As already mentioned, one can use the normal linear model if the data satisfies several assumptions. The most important assumption is the *independence* assumption. The independence assumption states that the error terms are independent and identically distributed. The mean of the errors is equal to zero $\mathbb{E}[\epsilon_i, \epsilon_j] = 0, \quad \forall i \neq j$, the variance is equal to a constant $\mathbb{V}ar[\epsilon_j] = \sigma^2$ which is also known as homoscedasticity and the covariance is equal to zero $\mathbb{E}[\epsilon_i, \epsilon_j] = 0 \quad \text{for} \quad \forall i \neq j$, also known as no-correlation of errors.

The independence of errors means that the errors are uncorrelated. Looking at the variance and covariance of these errors, one obtains a corresponding variance-covariance matrix whereby the diagonal of the matrix corresponds to the variance $\mathbb{V}ar[\epsilon_j] = \sigma^2$ and the non-diagonal values correspond to the covariance of the errors $\mathbb{E}[\epsilon_i, \epsilon_j] = 0 \quad \text{for} \quad \forall i \neq j$.

If the data does not satisfies the independence assumption this does not mean that there is no model possible for the data containing correlated errors, it only means that the normal linear model is inappropriate. Therefore other methods and approaches, which can handle correlated errors, need to step in.

Chapter 2

Correlated Errors

2.1 Introduction

In the previous chapter one has seen that the normal linear model assumes a model of form $\mathbf{y} = X\beta + \epsilon$ with homoscedasticity of errors, thus a constant variance $\mathbb{V}ar[\epsilon_j] = \sigma^2\mathbf{I}$. This means that the errors are having the same distance from the regression line across all the values of the explanatory variables so they are having the same distribution. In this chapter one will assume that the *homoscedasticity* of errors does not hold and that the errors are instead *heteroscedasticity*. This means that the errors are having non-constant variances. One will assume that the variance of the errors will then be of the following form:

$$\mathbb{E}\left[\epsilon\epsilon^T\right] = \sigma^2\Omega = \Sigma \tag{2.1.1}$$

The corresponding matrix for this model will be a variance-covariance matrix, from now on noted as *covariance matrix*, of the form: $\Sigma = \sigma^2\Omega$. The diagonal of the covariance matrix Σ does not have to consist of a constant σ^2 but can also take other values: $diag(\sigma_1^2 \cdots \sigma_n^2)$. If however, the entries on the diagonal of Σ are constant and σ^2, then one can write $\Sigma = \sigma^2\Omega$. The off-diagonal values of this matrix do not always have to be equal to zero everywhere. Which means that one can also have *correlation of errors*. Thus both sub-assumptions of the LM, *homoscedasticity* and *no-correlation* of errors, can be violated and captured in this matrix.

If one is dealing with heteroscedasticity the *ordinary least squares* (OLS) is not appropriate anymore. This is because the OLS tries to minimize the squares of errors and when the homoscedasticity of errors is present, the OLS gives equal weights to the observations finding the best $\hat{\beta}$-coefficients. But when heteroscedasticity of errors is present the OLS will give improper weights to the observation because the errors are biased which will result in the OLS not being the *best linear unbiased estimator* (BLUE) anymore. Thus, it will then not find the best $\hat{\beta}$-coefficients. In the upcoming sections the alternative approach, GLS, of dealing with heteroscedasticity and also correlated errors will be given and explained.

2.2 Generalized Least Squares

In the linear regression model:

$$\mathbf{y} = X\beta + \epsilon \tag{2.2.1}$$

The β-coefficient is a vector with unknown regression parameters. Under the Gauss-Markov conditions of section 1.2.5, the BLUE of the β-coefficients can be found with the OLS.

Now let us assume that the first G-M condition, the expected value of the error term is equal to zero $\mathbb{E}[\epsilon_i] = 0$, holds and that the second condition is not given by $\mathbb{E}\left[\epsilon\epsilon^T\right] = \sigma^2\mathbf{I}$ but is given by (2.1.1) with a known symmetric, positive definite correlation matrix Ω of order n [5]. This is under the assumption that the errors are normally distributed with a mean of zero and non-constant variance. Then the covariance matrix is:$\Sigma = \sigma^2\Omega$ [8]. With errors satisfying: $\epsilon \sim N(0, \Sigma)$.

Then the G-M condition $\mathbb{E}\left[\epsilon\epsilon^T\right] = \sigma^2\mathbf{I}$ is not satisfied and the OLS is not BLUE anymore.

Under these new assumptions, one will show that the best way to estimate the β-coefficients is with the GLS estimator which can be seen as a transformation of the linear regression model whereby the coefficients are obtained with a transformed version of the OLS. Which looks as follows:

$$\beta_{GLS} = \left(X^T\Sigma^{-1}X\right)^{-1}X^T\Sigma^{-1}\mathbf{y} \tag{2.2.2}$$

Here Σ is a positive definite covariance matrix containing (non-)constant variances on the diagonal and one or more covariances not being equal to zero on the off-diagonals.

Since Σ is positive definite, one can find a matrix such that Σ can be written as:

$$\Sigma = \Xi^T\Xi$$

where $\Xi^T\Xi$ is non-negative because it is a sum of squares and it is positive definite for all matrices Ξ. [5].

To obtain a model that satisfies the second G-M condition, homoscedasticity and no-correlation of errors, one can transform the linear regression model (2.2.1) by pre-multiplying both sides off the equation with Ξ^{-1}:

$$\Xi^{-1}\mathbf{y} = \Xi^{-1}X\beta + \Xi^{-1}\epsilon$$

Knowing all of the above and that the variance will be equal to (2.1.1), the mean and covariance of the errors will become:

$$\begin{aligned}
\mathbb{E}\left[\Xi^{-1}\epsilon\right] &= 0 \\
\mathbb{C}ov\left[\Xi^{-1}\epsilon\right] &= \Xi^{-1}\mathbb{C}ov\left[\epsilon\right](\Xi^T)^{-1} \\
&= \Xi^{-1}\mathbb{V}ar\left[\epsilon\right](\Xi^T)^{-1} \\
&= \sigma^2\Xi^{-1}\Sigma(\Xi^T)^{-1} \\
&= \sigma^2\Xi^{-1}(\Xi\Xi^T)(\Xi^T)^{-1} \\
&= \sigma^2\mathbf{I}
\end{aligned}$$

Now it becomes clear that transforming the data by multiplying both sides of the linear regression model with Ξ^{-1}, a model is obtained that satisfies the G-M conditions. Accordingly, one can then conclude that the OLS estimator is the BLUE estimator again for the β-coefficients of the transformed model. Thus:

$$\begin{aligned}\beta_{GLS} &= \left(X^T(\Xi^T)^{-1}\Xi^{-1}X\right)^{-1} X^T(\Xi^T)^{-1}\Xi^{-1}\mathbf{y} \\ &= \left(X^T\Sigma^{-1}X\right)^{-1} X^T\Sigma^{-1}\mathbf{y}\end{aligned} \tag{2.2.3}$$

2.3 Estimated Generalized Least Squares

In most natural cases the covariance matrix Σ is not known, therefore this matrix has to be estimated. To do so, take the model:

$$\mathbf{y_t} = X\beta + \epsilon_t \tag{2.3.1}$$

Where X is an $n \times p$ matrix with $t = 1, \cdots N$ and with $\mathbb{E}[\epsilon] = 0$, $\mathbb{C}ov[\epsilon] = \Sigma$.
Now let's apply the following:

$N^{-1}\sum_{t=1}^{N} \mathbf{y_t} = \bar{\mathbf{y}}$

$N^{-1}\sum_{t}^{N} \epsilon_t = \bar{\epsilon}$

By taking the mean of all the observations of the response variable $\mathbf{y_t}$ and error terms ϵ_t, this gives the following transformation to the model (2.3.1):

$$\begin{aligned}\mathbf{y_t} &= X\beta + \epsilon_t \\ N^{-1}\sum_{t=1}^{N}\mathbf{y_t} &= X\beta + N^{-1}\sum_{t}^{N}\epsilon_t \\ \bar{\mathbf{y}} &= X\beta + \bar{\epsilon}\end{aligned}$$

Then the unbiased estimator of Σ will be given by the following equation:

$$\hat{\Sigma} = (N-1)^{-1}\sum_{t=1}^{N}\left(\mathbf{y_t} - \bar{\mathbf{y}}\right)\left(\mathbf{y_t} - \bar{\mathbf{y}}\right)^T$$

Knowing that $\hat{\Sigma}$ is an estimate of Σ, the *estimated generalized least squares* (EGLS) of β is then given by:

$$\beta_{EGLS} = \left(X^T\hat{\Sigma}^{-1}X\right)^{-1} X^T\hat{\Sigma}^{-1}\bar{\mathbf{y}} \tag{2.3.2}$$

In Sen and Srivistava [5] there is a broad explanation about the estimation of $\hat{\Sigma}$.

2.4 Summary

Sometimes it happens that the Gauss-Markov condition where one assumes *homoscedasticity* and *no-correlation* of errors, does not hold. Then there is *heteroscedasticity* and/or *correlation* of errors. Which means that variance of errors vary along the explanatory variable(s) and/or the errors show dependency between observations. If one or both violations occur, the *ordinary least squares* estimator is not the *best linear unbiased estimator* for the β-coefficients.

The corresponding covariance matrix $\sigma^2\Omega = \Sigma$ can take the following forms:

- Diagonal values $diag(\sigma_1^2, \cdots \sigma_n^2)$ for heteroscedasticity.
- Off-diagonal values unequal to zero $\mathbb{C}ov\left[\epsilon_i, \epsilon_j\right] \neq 0, \forall i \neq j$ for correlated errors

To resolve the violation of the G-M condition(s), one can transform the linear regression model to a model whereby the G-M conditions do hold. This can be done by pre-multiplying both sides of the linear regression model (2.2.1) with Ξ^{-1} (a symmetric matrix).

Applying this, results in and estimator that can estimate the $\hat{\beta}$-coefficients which is also known as the GLS estimator (2.2.3). The GLS estimator can be seen as applying an OLS estimator to a linear transformation of the data.

Because in most natural cases the covariance matrix Σ is not known, this matrix can be estimated with $\hat{\Sigma}$ being an estimator of Σ. Then the $\hat{\beta}$-coefficients can be estimated with the *estimated generalized least squares* (2.3.2).

Chapter 3

Linear Mixed Effect Models

3.1 Introduction

As seen in the previous chapter, the *(estimated) generalized least squares* approach can be applied to resolve the problem that arises when errors have *heteroscedasticity* (non-constant variance) and/or have *correlation* (dependence) of errors. The (estimated) GLS handles this problem by transforming the data.

Another way of handling heteroscedasticity and/or correlation of errors is with the *linear mixed effect models* (LME). The LME model offers another way out. These models incorporate random-effects in the *linear regression model*. By incorporating random-effects, the model deals with responses coming from the same subject by adding random-effects for each subject or by adding random-effects on data taken in time. Applying this model can help with the dependency of errors. Example cases whereby a LME model can resolve the heteroscedasticity and/or correlation of errors are the following:

- Taking the example case in Chapter 1 about predicting the grade of a student (subject) by hand of their studying time. It could be the case that some students are studying together which will influence their grade. Students working together can be observed in the covariance matrix Σ of the errors. Then Σ will indicate correlations in the off-diagonal for students working together, which implies that the errors are correlated. The LME model can resolve this problem by incorporating random-effects for subjects. These effects take into account the heteroscedasticity of errors by creating different regression lines for every subject.

- Taking the example case about predicting the amount of grocery spending of a person (subject) based on their income, can be hypothetically speaking, data with non-constant variance. Because people with low income would have a small variance in errors while people with high income can have a different variance. It could also be the case that the data is as well correlated. This happens when these people, measured nearby each other, are influenced by their neighbours when shopping groceries in the same store. By incorporating random-effects the model uses this effect to create different regression lines which takes into account the variance per person thus the heteroscedasticity and/or correlation of errors.

In the upcoming sections an explanation will be given about the LME and how they incorporate random-effects which helps by resolving *heteroscedasticity* and/or *correlation* of errors.

3.2 Linear Mixed Effect Models

To recap Chapter 1, the normal linear model is of the following form:

$$\begin{aligned} y_j &= \beta_1 x_{j1} + \cdots + \beta_p x_{jp} + \epsilon_j \\ \epsilon_j &\sim \mathcal{N}\left(\mathbf{0}, \sigma^2\right) \end{aligned}$$

In matrix notation the model is written as follows:

$$\begin{aligned} \mathbf{y} &= X\beta + \epsilon \\ \epsilon &\sim \mathbf{N}\left(\mathbf{0}, \sigma^2\mathbf{I}\right) \end{aligned}$$

While the normal linear regression model has only one random-effect, namely ϵ_j, the *linear mixed effect model* is able to include additional random-effect terms [9] and is of the following form:

$$y_{ij} = \beta_1 x_{1ij} + \cdots + \beta_p x_{pij} + b_{i1} z_{1ij} + \cdots + b_{iq} z_{qij} + \epsilon_{ij} \tag{3.2.1}$$

With:

$$\begin{aligned} b_{ik} &\sim \mathbf{N}\left(0, \psi_k^2\right), & \mathbb{C}ov\left[b_k, b_{k^T}\right] &= \psi_{kk^T} \\ \epsilon_{ij} &\sim \mathbf{N}\left(0, \sigma^2\lambda_{ijj}\right), & \mathbb{C}ov\left[\epsilon_{ij}, \epsilon_{ij}T\right] &= \sigma^2\lambda_{ijj^T} \end{aligned}$$

Where, as cited in Fox [9]:

> y_{ij} is the response variable for the jth of n_i observations in the ith of M groups or clusters.
>
> $\beta_1, \cdots, \beta_p$ are the fixed effects coefficients which are unknown parameters.
>
> $x_{1ij}, \cdots, x_{pij}$ are the fixed design variables for observation j in group i.
>
> $b_{i1}, \cdots, b_{iq}$ are the random effect coefficients for group i which are multivariate normally distributed.
>
> $z_{ij}, \cdots, z_{qij}$ are the random effect design variables.
>
> ψ_k^2 are variances. ψ_{kk^T} are the covariances among the random effects. These are assumed to be constant across groups.
>
> ϵ_{ij} is the error for observation j in group i. The errors for group i are multivariate normal distributed where $\sigma^2\lambda_{ijj^T}$ represents the covariance between errors in group i

In matrix notation the model is written as follows:

$$\mathbf{y}_i = \mathbf{X}_i\beta + \mathbf{Z}_i\mathbf{b}_i + \epsilon_i \tag{3.2.2}$$

With:

$$\mathbf{b}_i \sim \mathbf{N}_q(\mathbf{0}, \mathbf{\Psi})$$
$$\epsilon_i \sim \mathbf{N}_{ni}\left(0, \sigma^2\mathbf{\Lambda}_i\right)$$

Where $\mathbf{b}_i$ and ϵ_i are independent and in terms of n data points and given in Fox [9]:

$\mathbf{y}_i$ is the $n_i \times 1$ response vector for observations in the ith group.

$\mathbf{X}_i$ is the $n_i \times p$ design matrix for the fixed effect for observations in group i.

β is the $p \times 1$ parameter vector of fixed-effect coefficients.

$\mathbf{Z}_i$ is the $n_i \times q$ design matrix dor random effects for observations in group i.

$\mathbf{b}_i$ is the $q \times 1$ vector of random-effect coefficients for group i.

ϵ_i is the $n_i \times 1$ vector of errors for the observations in group i.

$\mathbf{\Psi}$ is the $q \times q$ covariance matrix for the random effects.

$\sigma^2\mathbf{\Lambda}_i$ is the $n_i \times n_i$ covariance matrix for the errors in group i.

Since the model is Gaussian, the maximum likelihood is used for the estimation of the parameters. For testing the goodness of fit of the parameters, the likelihood and likelihood ratio tests are used in testing [1, 9].

3.2.1 Random-effects

In the LME the fixed-effects are mixed with the random-effects. The fixed-effects are the parameters of the statistical model which are expected to have a predictable influence on the data [6]. The random-effects are unobserved random variables which are an unpredictable factor, having a random influence on the data [1, 10]. Because random-effects are dependent variables and create correlation between observations, one obtains [11]:

$$\begin{aligned}
\mathbb{C}ov\left[\mathbf{y}_i\right] &= \mathbb{C}ov\left[\mathbf{X}_i\beta_i + \mathbf{Z}_i\mathbf{b}_i + \epsilon_i\right] \\
&= \mathbb{C}ov\left[\mathbf{X}_i\beta_i\right] + \mathbb{C}ov\left[\mathbf{Z}_i\mathbf{b}_i\right] + \mathbb{C}ov\left[\epsilon_i\right] \\
&= \mathbf{Z}_i\mathbb{C}ov\left[\mathbf{b}_i\right]\mathbf{Z}_i^T + \sigma^2\mathbf{\Lambda}_i \\
&= \sigma^2\mathbf{Z}_i\mathbf{\Psi}\mathbf{Z}_i^T + \sigma^2\mathbf{\Lambda}_i \\
&= \sigma^2(\mathbf{Z}_i\mathbf{\Psi}\mathbf{Z}_i^T + \mathbf{\Lambda}_i)
\end{aligned}$$

While for the *standard* fixed-effect model we have independence between the observations:

$$\begin{aligned}
\mathbb{C}ov\left[\mathbf{y}_i\right] &= \mathbb{C}ov\left[\mathbf{X}_i\beta_i + \epsilon_i\right] \\
&= \mathbb{C}ov\left[\mathbf{X}_i\beta_i\right] + \mathbb{C}ov\left[\epsilon_i\right] \\
&= \sigma^2\mathbf{\Lambda}_i
\end{aligned}$$

3.2.2 Different designs

The LME model can be used in different situations and can handle different kind of data structures. The following most important structures of data as point of view for this research paper, are described bellow. With for each different type of data structure an example case.

Blocked design In experiments *blocks* are variables that are not of interest in an experiment, but can have an effect on the measurements.

> **Example** In an experiment the effect of anti-depressants helping by the treatment of depression, is measured. The used anti-depressants in the experiment are delivered by three different fabrics. Thus we test three different anti-depressants. In a block design one could include a random-effect for each of the three anti-depressants. The randomness could then express the different blends of the anti-depressants per fabric. This is not of interest for the experiment but could have an effect on the medication.

Repeated measurements If measurements are repeatedly taken on an individual, one is dealing with repeated measurements.

> **Example** For an examination off the weekly growth of babies, babies are measured once a week to keep up with their weekly growth. Sometimes it happens that a baby is accidentally measured twice. Because these repeated measurements are included in the dataset we speak of *repeated measurements*, whereby this response is coming for the same subject.

Longitudinal design The data is called longitudinal data when repeated measurements, coming from the same subject are taken over time.

> **Example** For the prediction of grocery spending of households, the incomes are keeping record of for a couple of years. Therefore one sees in the dataset that the income of the households are coming from the same subject (household) over time. Thus one is dealing *longitudinal data* with repeated measurements coming from the same subject over time.

3.3 Measuring goodness of fit

Incorporating the random- and fixed-effects gives a lot of possibilities specifying a model. There is no standard guideline which fixed- and random-effects one should or should not include in the model. Therefore it is of importance to explore the data carefully to see which variables could be of interest. Next, it is a question of trying; adding and removing fixed- and random-effects to see if this improves the model. In R there is a function called anova() which tests nested models.

Nested models are models whereby a smaller model contains the same variables as a bigger model (also known as *full-model*). The difference is that the full-model contains additional variables than the smaller model. With the function anova() in R, one can test if models show a significant difference.

Function anova() This function tests the null hypothesis that the full-model with extra effects, adds an explanatory value over the smaller-model. With the corresponding p-values one can test this null hypothesis. If the model has a significant influence the p-value will be smaller than 0.05. Then H_0 will be rejected and H_1 will be accepted. This means that the extra fixed- and/or random-effects in the larger model are related with changes in the response variable.

3.4 Summary

When the independence assumption of homoscedasticity and/or correlation of errors is violated, the quality of statistical inference may not follow from the classical theory. An alternative approach that deals with non-independence of errors is the linear mixed effect model. This model deals with non-independence of errors by incorporating random-effects besides the fixed-effects.

In this chapter one has seen that the LME can handle different kinds of data structure like the blocked design, repeated measurements and longitudinal design by incorporating random-effects besides the fixed-effects. Random-effects are unobserved random variables which are an unpredictable factor having a random influence on the data. By incorporating this, a random intercept and potential random slopes can be added to the model. These random intercept and slopes have the ability to aid the fixed-effect by capturing an unpredictable influence on data. This allows the model to model errors containing heteroscedasticity and/or are correlated.

Because there is a wide variety of choices when it comes to including fixed- and random-effects to the LME. It is recommended to try including and excluding different fixed- and random-effects. To see if these changes improve the model, one can make use of the function anova() in R. This function tests if the full-model adds a clarifying value over the smaller-model. If the models has a significant influence, the p-value is smaller than 0.05. This means that certain effects in the full-model have a significant influence on the response variable.

Part II

Examples

Chapter 4

Correlated Errors

4.1 Introduction

In Chapter 2 of this research paper a theoretical overview of the *generalized least squares* approach is given. As theoretically shown, this approach is applied for the estimation of the unknown $\hat{\beta}$-coefficients when the *independence* assumption of the *(multiple) linear regression model* is violated. To see the effect of this approach, we simulated data in R(studio). The generated data contains a correlated covariance matrix with violation of the assumptions *homoscedasticity* and *correlation* of errors. We will use both the *ordinary least squares* and GLS to compute the $\hat{\beta}$-coefficients and compare these results.

4.2 Roadmap

For the regression equation the explanatory variables will be determined as well as fixed β-coefficients and the corresponding response variable $\mathbf{y}$ of our function will be computed. After computing these values we will simulate errors from the multivariate normal distribution with heteroscedasticity and correlation of errors $\epsilon \sim \mathcal{N}(0, \Sigma)$ and add these to our function. Next, we will calculate the final response value $\mathbf{y}$ and 'forget' the β-coefficients and errors. We will then compute the 'unknown' $\hat{\beta}$-coefficients with the following two approaches:

1. Ordinary least squares (ols) (1.2.3)

2. Generalized least squares (gls) with a known, simulated, Σ matrix (2.2.2)

By simulating the data, applying and comparing both approaches, an insightful view can be given about the effect of both approaches. By hand of the accuracy and concerning information we can see what the effect is when the *independence* assumption of the model is violated and what the power of the GLS approach is with reference to the OLS.

Polynomial For the research the following regression equation is created:

$$\mathbf{y} = \beta_1 + \beta_2 x_{linear} + \beta_3 x_{quadratic}$$

This equation contains a linear trend x_{linear} and a polynomial $x_{polynomial}$ explanatory variable. A polynomial regression is considered to be a special case of multiple LM, because the mean of the response variable $\mathbf{y}$ is linear in the unknown $\hat{\beta}$-coefficients.

Note: A polynomial regression was used because this function obtained the most interesting results compared to a simple LM.

4.3 Simulating data

To create a multiple LM of the form: $y = X\beta + \epsilon$ with correlated errors $\epsilon \sim \mathcal{N}(0, \Sigma)$, where $\Sigma = \sigma^2\Omega$, we can simulate data on the following way.

To start with, we fix the following three β-coefficients values:

$$\beta_1 = -10, \quad \beta_2 = 2, \quad \beta_2 = 5$$

Next, one creates three explanatory variables x of the following form:

$$x_1 = (1, 1, 1, \cdots, 1), \quad x_2 = (0, 1, 2, \cdots, n), \quad x_3 = (0, 1, 2^2, \cdots, n^2)$$

This is simulated for $n = 50$ observations. The model looks as follows:

$$f = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$
$$f(x) = -10 \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} + 2 \begin{pmatrix} 0 \\ \vdots \\ n \end{pmatrix} + 5 \begin{pmatrix} 0 \\ \vdots \\ n^2 \end{pmatrix}$$

Where $\beta_1 x_1$ corresponds to the intercept of the regression equation and $\beta_2 x_2 + \beta_3 x_3$ corresponds to the slope of the regression equation. To obtain this function $f(x)$, the code bellow can be used.

```
> fx <- b1*x1 + b2*x2 + b3*x3
```

Next, the covariance matrix Σ is simulated under the assumption that Σ is a known symmetric, positive definite matrix. Therefore we create a symmetric matrix with random values, drawn from the uniform distribution $\mathrm{U}(-5, 5)$. This matrix is then multiplied with its transpose form [5]. The result is a positive definite matrix wherefore we can simulate data from the multivariate normal distribution $\mathcal{N}(\mu, \sigma^2\Omega) = \mathcal{N}(0, \Sigma)$ to obtain the errors for our model:

```
> A <- matrix(runif(n*n,-5,5),n,n)
> Sigma <- t(A)%*%A
> mu <- rep(0,n)
> errors <- mvrnorm(1, mu, Sigma)
```

Having obtained the errors drawn from the multivariate normal distribution, it is time to compute our response variable y.

```
> y <- fx + residuals
```

Now we have obtained all the variables needed for the multiple linear regression model:

$$y = X\beta + \epsilon$$

Note: the the entire code to simulate this data can be found in Appendix A.1.

4.4 Fitting the data

In the following steps we assume that we only know the response variable y and explanatory variables x_2 and x_3. Thus forget the rest. With this information we are able to predict the response variable y by estimating the unknown $\hat{\beta}$-coefficients with the OLS and GLS.

4.4.1 Ordinary least squares

First we will use the OLS approach to estimate the unknown $\hat{\beta}$-coefficients. This can be done with the lm() function in R. The lm() function in R models a (multiple) LM and estimates the unknown coefficients with the OLS approach. To apply a multiple LM on our data, the following code can be used:

```
> fit.1 <- lm(y ~ x2 + x3, data=data)
```

After applying the model, we are able to obtain a summary of the fit of the model.

```
> summary(fit.1)

Call:
lm(formula = y ~ x2 + x3, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-36.988 -14.102   0.645  14.309  39.433

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.6214     7.9357   0.078    0.938
x2           -3.6650     3.6701  -0.999    0.323
x3            5.5580     0.3549  15.660   <2e-16 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05
   .    0.1         1

Residual standard error: 19.46 on 47 degrees of freedom
Multiple R-squared:  0.9861,    Adjusted R-squared:  0.9855
F-statistic:  1666 on 2 and 47 DF,  p-value: < 2.2e-16
```

Let us start examining the summary of our fitted model and explain the most important results with regard to the research.

Formula The first line shows the model formula which we wished to model. In this case y modeled as a function of x_2 and x_3.

Coefficients In the lines of the *Coefficients*, the *Estimate* column denotes the estimated $\hat{\beta}$-coefficients computed with the OLS.
As we can see the regression equation is given with an intercept of: $\hat{\beta}_1 = 0.621$ and for the first variable a slope of: $\hat{\beta}_2 = -3.665$ and for the second variable a slope of: $\hat{\beta}_3 = 5.558$. Our obtained regression hyperplane would thus become:

$$\hat{y} = 0.621 + -3.665x_2 + 5.558x_3$$

In the same lines the *standard error*, *t-value* and *p-value* can be found.

standard error The standard error (SE) is a measure that tells you how much

the $\hat{\beta}$-coefficient can vary from the estimation. This means that a low SE leads to more precision in the model. As we can see the SE for $\hat{\beta}_3 = 0.355$, which looks like a good estimation because the estimated value is 5.558. The SE of $\hat{\beta}_1 = 7.936$, is almost 13 times as high as its estimated value of 0.6214. This indicates that the estimate of the $\hat{\beta}_3$-coefficient has a big variance. The SE for $\hat{\beta}_2 = 3.670$ which is also really high compared to its estimated value of -3.665. Therefore it could be the case that these coefficients are not meaningful to include in the model.

t- and p-values With the standard errors, the corresponding t- and p-values can be found as we have seen in section 1.2.2. The p-value tests the following null hypothesis:

$$H_0 : \hat{\beta}_i = 0 \quad \text{vs.} \quad H_1 : \hat{\beta}_i \neq 0 \quad \text{for} \quad i = 1, 2, 3$$

When a $\hat{\beta}$-coefficient has effect on the model $\hat{\beta}_i \neq 0$ and the null hypothesis should be rejected. The null hypothesis is rejected when the p-value is smaller than 0.05. This means that this coefficient is meaningful to our model because changes in the corresponding explanatory variable values are related with changes in the response variable.

As we can see in the summary $\hat{\beta}_3$ is statistically significant with a p-value $= 2 * 10^{-16}$ which is smaller than 0.05. But $\hat{\beta}_1$ and $\hat{\beta}_2$ are not significant with p-values of respectively 0.938 and 0.323 which are both greater than 0.05. Therefore the associated null hypothesis is not rejected. This implies that including the corresponding explanatory variables x_1 and x_2 in the model, has no significant influence on the response variable [3]. Therefore it could be removed out of the model.

4.4.2 Generalized least squares

Because the data does not full-fill the independence of error assumption of the multiple LM, we know that the GLS can be applied to deal with the dependence of errors and obtain results whereby we can make trustworthy statistical inference. The formula of section 2.2 will be used to compute the unknown $\hat{\beta}$-coefficients with a known covariance matrix Σ.

$$\beta_{GLS} = \left(X^T \Sigma^{-1} X\right)^{-1} X^T \Sigma^{-1} \mathbf{y}$$

Implementing the formula for the β_{GLS} in R results in the following code for the model:

```
> b_gls <- solve(t(X)%*%solve(Sigma)%*%X)%*%(t(X)%*%solve(Sigma)%*%y)
```

Next, a function is written to extract the most important information of the GLS from the point of view of this research and gives a summary of the fit of the GLS. This function can be found in Appendix A.1 and can than be called with the following code:

```
> summary.gls(X,y,n,b_gls)
              Estimate  Std. Error   t value     Pr(>|t|)
(Intercept) -10.053061 0.115596110 -86.96712 1.400414e-53
x2            1.966976 0.061717408  31.87068 1.763008e-33
x3            5.004854 0.006351191 788.01826 1.659277e-98
```

Estimate As we can see the regression equation which is obtained with the GLS, has an intercept of: $\hat{\beta}_1 = -10.053$, for the first variable a slope of: $\hat{\beta}_2 = 1.967$ and for the second variable a slope of: $\hat{\beta}_3 = 5.005$. Our obtained regression hyperplane would be of the following form:

$$\hat{y} = -10.053 + 1.967x_2 + 5.005x_3$$

Standard error The SE of the $\hat{\beta}$ coefficients are really small. For example the SE of $\hat{\beta}_1 = 0.116$ which indicates a small variance compared to its estimated value of -10.053. The low SE's could indicate a precise model. Therefore, let us take a look a the corresponding t- and p-values.

t- and p-values As we can see, all $\hat{\beta}$-coefficients are statistically significant with for all a p-value smaller than 0.05. Therefore, the associated null hypothesis are rejected. This implies that including all the corresponding explanatory variables x_1, x_2 and x_3 in the model, has a significant influence on the response variable.
Note: The entire R-code for the function and the simulation of the GLS approach can be found in Appendix A.1.

4.5 Results

After having obtained all the results we can see that the GLS approach obtained much better estimates of the $\hat{\beta}$-coefficients than the OLS approach. Summarizing the estimate $\hat{\beta}$-coefficients with the OLS and GLS with reference to their real values and p-values gives the following results:

Coefficients	Real values	Ols	Gls	Ols: p-value	Gls: p-value
$\hat{\beta}_1$	-10	0.621	-10.053	0.938	0.000
$\hat{\beta}_2$	2	-3.665	1.967	0.323	0.000
$\hat{\beta}_3$	5	5.558	5.004	0.000	0.000

Table 4.1: Summary β-coefficients computed with OLS and GLS

As we can see in Table 4.1, the GLS estimates the $\hat{\beta}$-coefficients much better than the OLS. With the GLS, all the $\hat{\beta}$-coefficients are estimated really close to the real parameter values and are all three statistically significant. This implies that including all the corresponding explanatory variables does have a significant influence on the response variable. The OLS only estimates the $\hat{\beta}_3$-coefficients close to the real parameter value. But $\hat{\beta}_1$ and $\hat{\beta}_3$ are not even close to the real values and show both to have no statistically significant influence on the response variable.

Concluding, after having examined the difference of the OLS and GLS we obtained the following results:

- When testing the null hypothesis of the $\hat{\beta}$-coefficients, $\hat{\beta}_1$ and $\hat{\beta}_2$ showed no significance when estimated with the OLS. Also, the estimated values were not even close to the real values. E.g. The $\hat{\beta}_1 = 0.621$ while the real value is: -10
- For the GLS all the $\hat{\beta}$-coefficients were statistically significant and were estimated really close to the real parameter values. E.g. The $\hat{\beta}_1 = -10.053$ which lays close to the real value: -10.
- From the examination we can conclude that the GLS shows not to be influenced by the heteroscedasticity and correlation of errors. Whereas the OLS shows a big influence when estimating the $\hat{\beta}$-coefficients under these circumstances.

Chapter 5

Linear Mixed Effect Models

5.1 Introduction

In the following example we will use a *longitudinal* study which means that the used dataset contains repeated measurements of the same subject taken over time. We will use the *linear mixed effect model* to examine the dataset. To use the LME, the package 'lme4' in R will be used. The data can be found in the 'faraway' package. The following R-code can be used to find the model and data:

```
library(lme4)
library(faraway)
data(psid)
```

The 'psid' data is a data frame containing 6 variables with each 1661 observations. The measurements are taken from 1968 until 1990 and are taken from The Panel Study of Income Dynamics (PSID). The data is a representative sample of U.S. individuals. The study is conducted at the Survey Research Center, Institute for Social Research, University of Michigan and is still continuing and described in [12].

5.2 Examining PSID data

Before applying the LME, let us first explore the data by using some *descriptive* statistics.

Descriptive statistics are used to get a better insight of the data dealing with. The user can produce (simple) summaries about the measurements and can create (simple) figures to express the summary. Descriptive statistics is examining the given data without making predictions. It is helpful for obtaining clarifying insights of the data.

To summarize the data one can use the following code:

```
> summary(psid)
```

As can be found in the summary:

- The persons are aged from 25-39.
- The minimum years of education is 3 the highest years of education is 16.

- There are 732 females and 929 males in the dataset.
- The annual income in dollars differs from 3-180,000.
- The people were keeping track off for 11 years from 1968 until 1990. Where the *median* = 78 years and the *mean* = 78.61 years.
- Every person has a personal ID number and there are 85 heads of households.

Having summarized the data from the summary, let us take a look at some possible relationships expressed in figures. For example it could be possible that sex has a relation with the income of a person. Or that the degree of education results in a higher income. Therefore we could make a boxplot and scatterplot to express the data in figures and see if these variables are of interest.

```
boxplot(psid$income ~ psid$sex, ylab="Income", xlab ="Sex")
```

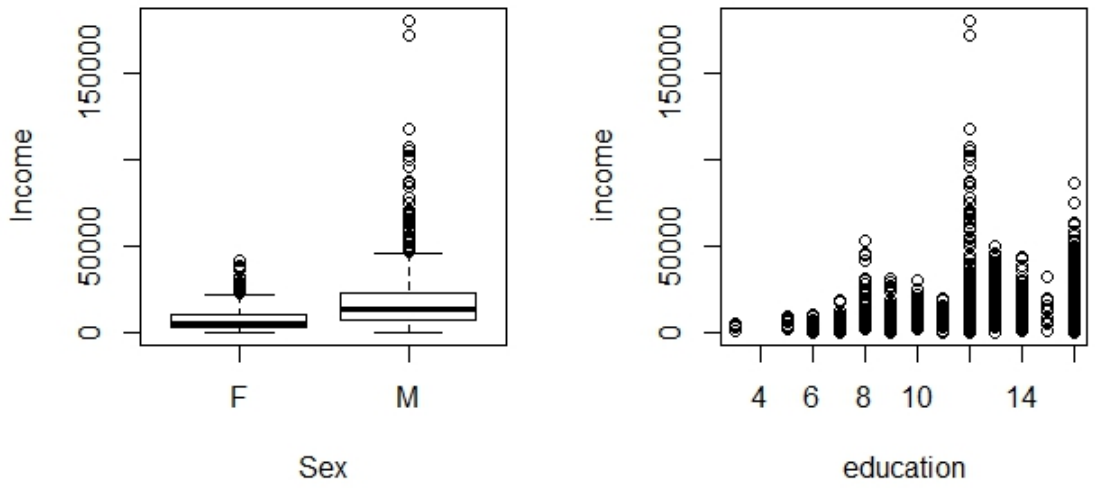

Figure 5.1: *The figure on the left is a boxplot of income against sex, on the right a scatterplot of income against education.*

As can be seen in the boxplot, females are earning less than males because the boxplot for females has a smaller range than that for males. Also, the mean of income lays lower for females than for males. In the scatterplot we notice that a higher education could have an influence on the income of a person. For example, a person with 12 years of education seems to have a higher income that people with 11 years of education or lower.

Next, we assume that the income of the households changes over time. If we want to predict the income of the households, it is of importance to examine which of the given explanatory variables could help to predict the income of a person (subject). Before specifying our LME, let us first recap the model as is described in section 3.2:

$$\mathbf{y}_i = \mathbf{X}_i\beta + \mathbf{Z}_i\mathbf{b}_i + \epsilon_i \tag{5.2.1}$$

With:

$$\mathbf{b}_i \sim \mathbf{N}_q(\mathbf{0}, \mathbf{\Psi})$$
$$\epsilon_i \sim \mathbf{N}_{ni}\left(0, \sigma^2 \Lambda_i\right)$$

Now as we know we are dealing with longitudinal data. Therefore it becomes clear that there will be variance within each subject from year to year. For now, we will consider that the error for the fixed-effects is uncorrelated and homogeneous. Then the corresponding covariance matrix for the fixed-effects is: $\Lambda_i = I$, thus the identity matrix. This implies that for the fixed-effects model we do not have correlations between the observations.

Next, it is finally time to specify our LME model. Therefore we need to consider which variables will be used in the model to predict the income of a person. In Table 5.1 a description of the chosen variables for the model can be found.

Effect:	**Variable:**	**Subscription:**
Fixed	Year	The income of a subject changes over time.
Fixed	Sex	The income differs per sex.
Fixed	Education	The income differs over the level of education.
Fixed	Age	The income differs per age category.
Random	Person	
Random	Year	

Table 5.1: Choice of variables for the LME model

To provide a good interpretation of the results of the LME we will center the year factor by creating the new factor 'cyear' which is the year minus the median of the year.

```
> psid$cyear <- psid$year-median(psid$year)
```

Also, we will take the logarithm of the income because taking the logarithm takes care of the distribution of the income. This means that the distribution of income will make sure that changes in small values will result in more separated values, while changes in big incomes with, for example, the same amount of change will result in a relative small separation.

Example (*Income of a person*)
The income of a person A with little income (\$1,000 per month), increases with an amount of \$1,000 and the income of a person B with a high income (\$10,000 per month), also increases with \$1,000. When examining the 'absolute change' effect of the total income for both A and B this will be equal, namely: $\$1,000 - \$1,100 = \$100$ and $\$10,000 - \$11,000 = \$100$. If one looks at the 'relative change' per person, one will see that the amount of money hs a big influence on person A than that of person B. Thus for A: $\frac{\$1,100-\$1,000}{\$1,000} = 0.5 = 50\%$ against that of B $\frac{\$11,000-\$10,000}{\$10,000} = 0.09 = 9\%$
By taking the logarithm of the income, this will make sure that we are looking at the 'relative change' instead of the 'absolute change'.

Having specified the data transformations and the variables that can be used in the LME, let us continue to the examples. In the upcoming subsections we will try two different LME by including random-effects.

Note: In Appendix A.2 the entire R-code for the LME examination can be found.

5.2.1 Example 1

In this example, we will make a LME with the specified fixed-effects of Table 5.1. We will also take into account a random intercept $\mathbf{b}_1$ for every person. The LME model will than expect that there can be changes in the income per person which means that a person is allowed to shift in income. The model will look as follows:

$$\begin{aligned}\log(\text{income}) = \beta_1 + \beta_2 \text{year}_i + \beta_3 \text{sex}_j + \beta_4 \text{age}_j + \beta_5 \text{education}_j \\ + \mathbf{b}_1 + \epsilon_{ij}\end{aligned} \tag{5.2.2}$$

In this model we estimate the parameters for the fixed-effects intercept and slope just like in a linear regression model. But because we model an LME, the model will also include an estimation for a random intercept for every subject ('person'). Next, we can estimate and summarize this model (5.2.2) with the following codes:

```
> model.1 <- lmer(log(income) ~ cyear + sex + age + educ
              + (1|person), data = psid)
> summary(model.1)
```

The summary of this model gives the following information. Note, that from the point of view of this research paper only the most important parts are discussed.

```
Formula: log(income) ~ cyear + sex + age + educ + (1 | person)
   Data: psid

REML criterion at convergence: 3966.1

Scaled residuals:
    Min      1Q  Median      3Q     Max
-9.4412 -0.2171  0.1226  0.4574  2.4111

Random effects:
 Groups   Name        Variance Std.Dev.
 person   (Intercept) 0.2889   0.5375
 Residual             0.5561   0.7457
Number of obs: 1661, groups:  person, 85

Fixed effects:
            Estimate Std. Error t value
(Intercept) 6.803007   0.558193  12.188
cyear       0.069685   0.002959  23.551
sexM        1.115594   0.122960   9.073
age         0.007971   0.013887   0.574
educ        0.105444   0.022060   4.780

Correlation of Fixed Effects:
      (Intr) cyear  sexM   age
cyear  0.015
sexM  -0.104  0.027
age   -0.874 -0.012 -0.025
educ  -0.598 -0.023  0.009  0.167
```

In R the formula corresponding to model (5.2.2) is written in the first line.

The next important part is the *Fixed effects* part. Just like in the LM, the **Estimate** column tells the user what the estimated $\hat{\beta}$-coefficients are. Please note that these coefficients are representing percentages because of taking the logarithm of income. Different is the column *Fixed effects* which takes into account that the income per person can vary. Now let us summarize the **Fixed effects** output of the summary [1]:

- (intercept): The intercept for the fixed-effects is $\hat{\beta}_1 = 6.803$.
- cyear: The slope for cyear is $\hat{\beta}_2 = 0.0697$. This means that income increases around 6.97% a year.
- SexM: The income of persons is determined as follows: income $= e^{\beta_1} e^{\beta_2 \text{sex}_j} \cdots$. the variable sex_j can take the value 1 or 0 which indicate respectively 'male' or 'female'. The slope for males will be $e^{1.116*1} = 3.05$ and for females $e^{1.116*0} = 1$. Thus the income for males are expected to be 3.05 times higher than for females.
- Age: The slope for age is $\hat{\beta}_4 = 0.00797$. This means that for every added year to a persons life, income increases with 0.80%.
- Educ: The slope for education is $\hat{\beta}_5 = 0.1054$. This means that for every extra year of education, the income increases with 10.54%.

Now let us look at the **Random effects**. We see that for the random-effects the standard deviation for the random intercept is 0.5375 with no correlation because we only added one random-effect. The corresponding covariance matrix Ψ for the random effect only contains one value, namely $\hat{\Psi}_1 = 0.5375$.

5.2.2 Example 2

In the first example we assumed a random intercept by taking into account that the income per person can change. Next, we will try to estimate a hyperplane with a random intercept and a random slope and see how that changes our model containing the same fixed-effects as in Table 5.1. The difference will lay in including the random slope.

Thus we will now assume that, besides taking into account the variation per income per person, we will also take into account the variation from year to year per person. The model will then be of the following form:

$$\begin{aligned} \log(\text{income}) = \beta_1 + \beta_2 \text{year}_i + \beta_3 \text{sex}_j + \beta_4 \text{age}_j + \beta_5 \text{education}_j \\ + \mathbf{b}_1 + \mathbf{b}_2 \text{year}_i + \epsilon_{ij} \end{aligned} \tag{5.2.3}$$

Having specified our model, we can now estimate and summarize model (5.2.3) with the following R-codes:

```
> model.2 <- lmer(log(income) ~ cyear + sex + age + educ +
                              (cyear|person), data=psid)
> summary(model.2)
```

The summary of this model gives a lot of information. Again, only the most important parts form the point of view of the research paper are discussed.

```
Formula: log(income) ~ cyear + sex + age + educ + (cyear | person)
   Data: psid

REML criterion at convergence: 3817.4

Scaled residuals:
     Min       1Q   Median       3Q      Max
-10.2140  -0.1974   0.0751   0.4062   2.8440

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 person   (Intercept) 0.281234 0.53032
          cyear       0.002494 0.04994  0.19
 Residual             0.467687 0.68388
Number of obs: 1661, groups:  person, 85

Fixed effects:
            Estimate Std. Error t value
(Intercept) 6.669167   0.542660  12.290
cyear       0.071101   0.006191  11.485
sexM        1.191649   0.119665   9.958
age         0.010658   0.013506   0.789
educ        0.103722   0.021409   4.845

Correlation of Fixed Effects:
      (Intr) cyear  sexM   age
cyear  0.027
sexM  -0.105  0.024
age   -0.874 -0.008 -0.025
educ  -0.597 -0.011  0.010  0.167
```

The **Estimate** column of the *Fixed effects* part of the summary, tells the user something about the effect on income per variable. The following part summarizes the effect of each fixed-effect variable [1]:

- (intercept): The intercept of the hyperplane for the fixed-effects is $\hat{\beta}_1 = 6.6691$.
- cyear: The slope for cyear is $\hat{\beta}_1 = 0.0711$. This means that income increases around 7.11% a year.
- SexM: The slope for males is $e^{1.191*1} = 2.72$ and for females $e^{1.191*0} = 1$. Thus the income for males are expected to be 2.72 times higher than the income for females.
- Age: The slope for age is $\hat{\beta}_4 = 0.0107$. This means that for every added year to a persons life, income increases with 1.07%.
- Educ: The slope for education is $\hat{\beta}_5 = 0.1037$. This means that for every extra year of education, the income increases with 10.37%

For the random-effects we see that the standard deviation for the intercept is 0.531 and the slope has a standard deviation of 0.049. Also these coefficients have a correlation of 0.19.

The corresponding covariance matrix for the random-effects Ψ would become: $\hat{\Psi}_1 = 0.530$, $\hat{\Psi}_2 = 0.0499$ and $\hat{\Psi}_{12} = 0.19 \times 0.530 \times 0.0499 = 0.005$. In matrix notation this will take the following form:

$$\Psi = \begin{pmatrix} \hat{\Psi}_1^2 & \hat{\Psi}_{12} \\ \hat{\Psi}_{12} & \hat{\Psi}_2^2 \end{pmatrix} = \begin{pmatrix} 0.530^2 & 0.005 \\ 0.005 & 0.0499^2 \end{pmatrix}$$

5.2.3 Comparing models

Having fitted model (5.2.2) and (5.2.3) it is time to see how good these models actually perform. Therefore we use the function anova() in R.

This function tests if the nested models are significant or not. Thus in our case it would test if the random-effect for year is statistically significant or not.

$$H_0 : \mathbf{b}_2 = 0 \quad \text{vs.} \quad H_1 : \mathbf{b}_2 \neq 0$$

We obtained the following result with the function anova():

```
> anova(model.2, model.1)
refitting model(s) with ML (instead of REML)
Data: psid
Models:
model.1: log(income) ~ cyear + sex + age + educ + (1 | person)
model.2: log(income) ~ cyear + sex + age + educ + (cyear | person)
        Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
model.1  7 3951.5 3989.4 -1968.8   3937.5
model.2  9 3808.1 3856.8 -1895.0   3790.1 147.43      2  < 2.2e-16 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05
   .    0.1         1
```

As we can see in the output, the full-model ('model.2') shows a p-value $= 2.2 * 10^{-16}$ which is smaller than 0.05. This shows us that including the random-effect for year to year variation per person improves the performance model. Thus the model assuming time dependence is better than the model not including this effect. Conclusion, the income of people in the study are affected by the year to year variance.

5.3 Results

Having tried two different models to test the effect of the LME model, we have seen that it does make sense to include random-effects. As seen above, adding an extra random-effect that includes year to year variation and variance per person, does have a significant effect on the model. Therefore it seems like adding random-effects makes the LME model predict the income of persons better. But as mentioned in Faraway [1], using random-effects can also make the model tricky. Because in our case, adding random-effects that includes year to year variation and variance per person, also creates more complicated models, complicated algebraic calculations and makes it more difficult to understand and work with [1]. Thus modeling the dependency of errors, results in better predictions but can be tricky and must be handled carefully.

Part III

Conclusion & Recommendation

Chapter 6

Conclusion & Recommendation

6.1 Introduction

For this research the aim was to examine how to deal with dependency of errors, a problem that arises when the *independence* assumption of the *linear regression model* is violated. Therefore we deepened into the literature of the linear regression model and investigated alternative approaches that are able to produce a regression line by dealing with dependency of errors. From the point of view of this research paper, the violation of the following sub-assumptions are examined:

- The errors of the observations have non-constant variance, also known as *heteroscedasticity*. Then the corresponding covariance matrix will have non-constant variances on the diagonal, thus: $diag(\sigma_1^2 \cdots \sigma_n^2)$.
- The observations are dependent of each other, which results in *correlated* errors. In the corresponding covariance matrix this is observed by having one or more off-diagonal values unequal to zero.

When one or both of these sub-assumptions is/are violated, then the *generalized least squares* offers a solution by transforming the LM. The *linear mixed effect model* handles the dependency of errors by including random-effects. To investigate how these approaches performed practically, we simulated data and examined a dataset from the 'faraway' package in R. In the upcoming sections a conclusion about the investigation is given, followed by a recommendation for further research.

6.2 Conclusion

During the research we have tried to find an answer on the main question:

How can one deal with dependency of errors?

To answer the main question we gave an overview of the LM to understand the basics and assumptions of the model. Next, there was an examination of the theory of the GLS and LME, two alternative approaches that can handle the violation of the independence assumption. To investigate the performance of these approaches practically, we made use of real and simulated data.

For the GLS we simulated a multiple LM with self-determined explanatory variables, β-coefficients and errors. The errors where simulated from the multivariate normal distribution with a covariance matrix Σ containing heteroscedasticity and correlation of errors. After the simulation we applied the OLS and GLS to investigate how good they estimated the three $\hat{\beta}$-coefficients. The GLS obtained good estimates for all three $\hat{\beta}$-coefficients. According to their p-values, all the coefficients were statistically significant. This implies that including all the corresponding explanatory variables in the multiple LM model, has a significant influence on the response variable. For the OLS only $\hat{\beta}_3 = 0.000$ was smaller than 0.05 thus only $\hat{\beta}_3$ was statically significant. This implies that only the corresponding x_3 should be included in the model because this is the only variable having a significant influence on the response variable. Thus we can conclude that the model, obtained with the GLS, performs better than the model obtained with the OLS.

The performance of the LME was tested on a longitudinal study. The data observations contained repeated measurements of the same subject over time. First, we fitted a small model with fixed-effects and added a random-effect. This random-effect took the changes for the response value per subject into account by adding a random intercept per person. Next, we fitted a full-model containing the same fixed-effects and random intercept as in the smaller model, but also a random slope which took the time variation per subject into account. Finally, we tested if the full-model showed a significant difference with the function anova() in R. The results showed that the full-model had a significant influence. This implies that the extra random-effect is related with changes in the response variable. Concluding, the full-model performed better when it took the time variance per subject into account. But please note that including random-effects can be tricky and need to be handled carefully.

We saw, that it is still possible to produce regression equations and make predictions, while having dependency of errors. By applying an GLS approach, the data is transformed by multiplying the LM model with a symmetric matrix. The results of this approach showed big differences in estimating the $\hat{\beta}$-coefficient of the LM than the OLS. The GLS was applied on data containing errors with a non-constant variance and correlation. The LME was applied on data containing errors that contained repeated measurements coming from the same subject. The LME showed that taking into account the time variance per subject, which causes dependency of errors, improved the performance of the model.

6.3 Recommendation

The results from the practical part of the research gave useful insights about the performance of alternative approaches that handle heteroscedasticity and/or correlation of errors. The results of the approaches showed improvements and differences when applying them on data, violating the independence assumption. However, to extend the investigation and obtain more insights about the performance of the approaches, the following points are recommended for further research:

- For the simulation of data more explanatory variables can be added. One of these variables can be absolutely random to create a more realistic environment.
- Also, the GLS could be applied on real data with non-constant variance and correlated errors to investigate the performance. Apply the OLS and compare the results of both approaches.
- For the LME more random-effects can be added to the model to investigate if the model can be improved.
- Also, modeling an LM and an LME containing the same fixed-effects (but for the LME with added random-effects), could be applied and compared by looking at the estimated $\hat{\beta}$-coefficients. See what changes this creates in the performance of the models.
- Use another dataset, for example, with repeated measurements only. Then use a similar approach by making a smaller and bigger model and compare these models with the function anova() to see how random-effects influence the prediction of the response variable.
- Investigate the performance of other alternative approaches that can be applied when the homoscedasticity and/or no-correlation of errors assumption is violated.

Bibliography

[1] J.J. Faraway. *Extending the Linear Model with R; generalized linear, mixed effects and nonparametric regression models*. Chapman & Hall/CRC.

[2] F. Bijma & M.C.M de Gunst. *Statistical Data Analysis*. Department of Mathematics, Faculty of Sciences, VU University Amsterdam.

[3] P.H. Franses T.Kloek & H.K. van Dijk C. Heij, P. de Boer. *Econometric Methods with Applications in Business and Economics*. Oxford University Press, 2004.

[4] A.C. Davidson. *Statistical Models*. Cambridge University Press, 2003.

[5] A. Sen & M. Srivastava. *Regression Analysis: Theory, methods, and applications*. Springer-Verlag New York Inc.

[6] B. Winter. Linear models and linear mixed effects models in r with linguistic applications.

[7] H.J. Seltman. Experimental design and analysis, June 2015. URL http://www.stat.cmu.edu/~hseltman/309/Book/Book.pdf.

[8] T.M. Palmer & J.A.C Sterne. Meta-analysis in stata: An updated collection from the stata journal. URL http://cphs.huph.edu.vn/uploads/tainguyen/sachvabaocao/Meta-AnalysisinStata.pdf#page=169.

[9] J. Fox. Linear mixed models: Appendix to an r and s-plus companion to applied regression.

[10] D.M. Bates. lme4: Mixed-effects modeling with R. Springer.

[11] B. Knapik. Advanced methodology: Lecture 6.

[12] M. S. Hill. *The Panel Study of Income Dynamics: A User's Guide*. Sage Publications, Newbury Park, CA, 1992.

Appendices

Appendix A

R-codes

A.1 Generalized least squares

Function:
1. Computing the summary with the GLS:

```
summary.gls <- function(X,y,n,b_gls) {
  Sigma_inv = solve(Sigma)
  X_Sigma = Sigma_inv%*%X
  y_Sigma = Sigma_inv%*%y
  H <- X_Sigma%*%solve(t(X_Sigma)%*%X_Sigma)%*%t(X_Sigma)
  SS_e <- t(y_Sigma)%*%(diag(n)-H)%*%y_Sigma
  MS_e <- SS_e/(n-(2+1))
  sigma.2 <- as.numeric(MS_e)
  C <- sigma.2*solve(t(X)%*%Sigma_inv%*%X)
  se_b1 <- sqrt(C[1,1])
  se_b2 <- sqrt(C[2,2])
  se_b3 <- sqrt(C[3,3])
  std.errors <- c(se_b1,se_b2,se_b3)
  gls_t.stat1 <- b_gls[1,1]/se_b1
  gls_t.stat2 <- b_gls[2,1]/se_b2
  gls_t.stat3 <- b_gls[3,1]/se_b3
  gls_t.values <- c(gls_t.stat1 , gls_t.stat2 , gls_t.stat3)
  gls_p.value1 <- 2* pt(-abs(gls_t.stat1),df=n-3) # wrong df previously in all 3 pvals
  gls_p.value2 <- 2* pt(-abs(gls_t.stat2),df=n-3)
  gls_p.value3 <- 2* pt(-abs(gls_t.stat3),df=n-3)
  gls_p.values <- c(gls_p.value1 , gls_p.value2 , gls_p.value3)
  matrix.a <- cbind(c(b_gls),c(std.errors),c(gls_t.values),c(gls_p.values))
  rownames(matrix.a) <- c("(Intercept)","x2","x3")
  colnames(matrix.a) <- c("Estimate", "Std.␣Error", "t␣value", "Pr(>|t|)")
  return(matrix.a)
}
```

Packages:

```
library(MASS)
```

Simulation of the data:

```
set.seed(524234)

n = 50
x1 <- rep(1,n)
x2 <- seq(0,10,length=n)
x3 <- x2^2
X <- matrix(c(x1, x2, x3), nrow=n)
b1 = -10
b2 = 2
```

```
b3 = 5
fx <- b1*x1 + b2*x2 + b3*x3

#Computing positive semi-definite covariance matrix
A <- matrix(runif(n*n,-5,5),n,n)
Sigma <- t(A)%*%A

mu <- rep(0,n)
errors <- mvrnorm(1, mu, Sigma)

#COMPUTING Y
y <- fx + errors
data<- data.frame(y=y, x=X)

#LINEAR MODEL FIT
fit.1 <- lm(y ~ x2 + x3, data=data)
summary(fit.1)

#FIT GLS WITH KNOWN OMEGA
b_gls <- solve(t(X)%*%solve(Sigma)%*%X)%*%(t(X)%*%solve(Sigma)%*%y)
summary.gls(X,y,n,b_gls)
```

A.2 Linear mixed effect models

Packages:

```
library(lme4)
library(faraway)
```

Examining data:

```
data(psid)
summary(psid)

education <- psid$educ
income <- psid$income

par(mfrow=c(1,2))
boxplot(psid$income ~ psid$sex, ylab="Income", xlab ="Sex")
plot(education, income)
```

Making & comparing models:

```
psid$cyear <- psid$year-median(psid$year)

model.1 <- lmer(log(income) ~ cyear + sex + age + educ + (1|person), data=psid)
summary(model.1)

model.2 <- lmer(log(income) ~ cyear + sex + age + educ + (cyear|person), data=psid)
summary(model.2)

anova(model.2, model.1)
```