Maximum likelihood estimation of treatment effect in clinical trials with multiple follow-up measurements

Author: S. Roberti

Supervisors: Prof. dr. E.A. Cator
Prof. dr. M.J.C. Eijkemans
Dr. N.P.A. Zuithoff

July 2017
Abstract

This thesis compares methods of estimating treatment effect from clinical trials with multiple follow-up measurements, using two different data generating mechanisms. The first such model that has been looked into is the random intercept model where the treatment provides a single sustained effect in the outcome variable, the other one is a random intercept model where the treatment effect is defined as a slope. For both of these models, different approaches such as ANCOVA and change score analysis have been compared in their stability of estimating the treatment effect. Further, the effects of introducing regression to the mean by using a selection criterion have been explored in a simulation study. It was shown that the choice of best estimator depends on specific properties of the context, such as the correlation between measurements, the number of follow-up measurements and the time points at which they are done, and the assumed model itself. It will be shown there are multiple competitive approaches with their own merits, but that the widely used ANCOVA is relatively robust under aforementioned specifics of the trial in the sense that is a relatively safe choice to use this method when the specifics are unknown or unclear.
# Contents

1. Introduction 3

2. Background 4
   - 2.1 Maximum likelihood estimation 4
   - 2.2 Maximum likelihood estimation of misspecified models 6
   - 2.3 Regression theory
     - 2.3.1 The linear model 8
     - 2.3.2 Random effects and mixed models 9
   - 2.4 Problem statement and notation 11

3. Model 1: random intercepts and single sustained treatment effect 13
   - 3.1 The model 13
   - 3.2 Estimating treatment effect
     - 3.2.1 A note on averages 14
     - 3.2.2 Change scores 15
     - 3.2.3 Outcomes with intercept and treatment effect 19
     - 3.2.4 ANCOVA 20
     - 3.2.5 Autoregressive approach 24
     - 3.2.6 Random intercept model 25
     - 3.2.7 ANCOVA with random intercept 27
   - 3.3 Discussion 29

4. Model 2: random intercepts and treatment effect as slope 31
   - 4.1 The model 31
   - 4.2 Estimating the treatment effect
     - 4.2.1 Another note on averages 32
     - 4.2.2 Change scores 33
     - 4.2.3 Using outcomes and ignoring random intercept 35
     - 4.2.4 ANCOVA 36
     - 4.2.5 Autoregressive approach 39
     - 4.2.6 Random intercept model 40
     - 4.2.7 ANCOVA with random intercept 41
   - 4.3 Discussion 42

5. Using an inclusion criterion based on baseline measurements 44
   - 5.1 The effect on the distributions 44
   - 5.2 Simulation strategy 46
   - 5.3 Estimating the treatment effect 47
   - 5.4 Results 47

6. Conclusion and discussion 51

Bibliography 52
Chapter 1

Introduction

Randomised controlled trials are medical experiments to test the effect of a treatment. The participants of the study are randomly divided between groups (treatment and control group or two groups for different treatments), and given a placebo or treatment depending on this allocation. One or more baseline measurements are done before randomisation, followed by the desired number of follow-up measurements at different times after the treatment has been given. The baseline measurement can be used for an inclusion criterion (only including patients with a baseline measurement above or below a certain threshold), or as an outcome to help estimate the treatment effect. In some cases, it can even be ignored completely.

The type of data that comes from this type of trial, namely data where one or more variables are tracked over time, is known as longitudinal data. This is different from some other kinds of experiments because with longitudinal data, measurements on the same person cannot just be assumed to be independent. This means that the widely used linear model cannot be easily applied, because there independent data are assumed.

Fortunately, there are still many ways to analyse longitudinal data and estimate the treatment effect. For example, the method named ‘analysis of covariance’ (ANCOVA) is widely used and it has been studied in great detail. In the case of a single follow-up measurement, it is often suggested to use the ANCOVA. In this thesis we compare different methods of analysis for multiple follow-ups, for different assumed ground truths or data generating mechanisms. One thing we will look out for is if the ANCOVA method holds up well in the cases we study, and how it compares to other widely used estimators. Interestingly, Lord (1967) showed that using ANCOVA may result in a statistically significant treatment effect, while using another valid method on the same problem yields an insignificant effect.

The reason I chose to study this problem is that it is a combination of theoretical, mathematical statistics and more applied biostatistics, which I find fascinating. Also, methodological research is interesting because any results could provide more insight in the methods used in practice. There are more papers about this topic, but most take a different approach for this problem. Twisk & de Vente (2008) [7] describes a similar study, but does not go into the mathematical discussions as much, and Frison & Pocock (1992) [3] describes the use of summary statistics when analysing repeated measures, which involves summarising all follow-up measurements into a single quantity.

We start by giving some background theory and stating the problem in more detail. Then, in Chapters 3 and 4, we assume two data generating mechanisms and compare a number of methods to estimate the treatment effect. In Chapter 5, we discuss the consequences of imposing a selection criterion on the patients and use a simulation study to compare estimators. Finally, in chapter 6 we provide a conclusion and discuss the results. All numerical and simulation procedures throughout this text are done with the R software.
Chapter 2

Background

In this chapter we will outline some necessary probabilistic and statistical background knowledge, that will be used throughout the other chapters. Also, we will give a more detailed problem statement.

2.1 Maximum likelihood estimation

Maximum likelihood estimation is a way of estimating unknown statistical parameters from a set of observations. It has a few optimality properties that make it a widely used method of estimation. First, let us provide the formal definition:

Definition 2.1. Let $x_1, ..., x_n$ be an independent and identically distributed (i.i.d.) sample from a distribution with probability density $p_{\theta_0}$, where $\theta$ is an unknown parameter (vector) that lies in a parameter space $\Theta$. The likelihood function $L(\theta; x_1, ..., x_n)$ is defined by $\theta \mapsto L(\theta; x_1, ..., x_n) = \prod_{i=1}^n p_\theta(x_i)$. Then the maximum likelihood estimator, denoted by $\hat{\theta}$, is given by

$$\hat{\theta} = \arg \max_{\theta \in \Theta} L(\theta; x_1, ..., x_n).$$

(2.1)

Sometimes, when studying the asymptotic properties of the maximum likelihood estimator, the notation $\hat{\theta}_n$ is used to denote its relation to a sample of size $n$.

In a sense, the maximum likelihood estimator can be seen as the element of $\Theta$ that makes observing $x_1, ..., x_n$ the most likely. Instead of maximising the likelihood function, often the log-likelihood function $\ell(\theta; x_1, ..., x_n) := \log(L(\theta; x_1, ..., x_n))$ is maximised. These procedures are equivalent, since log is a monotone function.

One of the properties of the maximum likelihood estimator is that under some conditions it is consistent, meaning that for a large number of observations, $\hat{\theta}$ will be close to $\theta_0$. Another asymptotic property is that (again under some conditions), it is asymptotically normal with a known expression for the asymptotic variance, which is useful to compute (asymptotic) confidence intervals.

The needed regularity conditions may be hard to check in practice when the true data generating mechanism is unknown. They include that the likelihood function should be continuous in $\theta$, that the expectation under the true parameter $\theta_0$ of the log-likelihood exists, and that $\ell(\theta; x_1, ..., x_n)/n$ converges to the expected (under $\theta_0$) marginal log-likelihood. For asymptotic normality, some more conditions are needed, namely that $\theta_0$ is an interior point of $\Theta$, that $\ell$ is twice continuously differentiable in a neighbourhood of $\theta_0$, that the order of integration (expectation) and differentiation can be exchanged, and finally that the matrix $I$ (to be defined
Theorem 2.2. Under regularity conditions and identifiability \((p_0 \neq p_{\theta_0} \text{ for } \theta \neq \theta_0)\),
\[
\sqrt{n} (\hat{\theta} - \theta_0) \sim N \left(0, nI(\theta_0)^{-1}\right).
\] (2.2)

Note that \(A_n \sim A\) denotes convergence in probability, meaning the distribution function of \(A_n\) converges to that of \(A\) as \(n \to \infty\), everywhere where the distribution function of \(A\) is continuous.

The consequence of this theorem is that the asymptotic variance of the \(k\)-th component of the maximum likelihood estimator is equal to \([I(\theta_0)^{-1}]_{k,k}\). The (asymptotic) covariance matrix of \(\hat{\theta}\) is the inverse of the (complete) Fisher information matrix, for which the \(i,j\)-element is evaluated in the true parameter \(\theta_0\).

Assume the context and regularity conditions described earlier in this section, and let \(T(X)\) be an unbiased estimator of \(\theta_0\) (i.e. \(E_{\theta_0}(T(X)) = \theta_0\)). Then \(\text{Cov}_{\theta_0}(T(X)) - I(\theta_0)^{-1}\) is positive semi-definite, meaning that for any column vector \(z\),
\[
\left(z^T \left(\text{Cov}_{\theta_0}(T(X)) - I(\theta_0)^{-1}\right) z - z^T I(\theta_0)^{-1} z\right) \geq 0.
\] (2.5)

In particular, for any component \(T(X)_k\) (which is an estimator for \(\theta_k\)),
\[
\text{Var}_{\theta_0}(T(X)_k) \geq [I(\theta_0)^{-1}]_{k,k}.
\] (2.6)

The last statement follows from taking \(z\) to be the vector of all zeroes and a one in position \(k\). We will not prove this famous result, as it is beyond the scope of this text. However, the consequences are important, and thus we will discuss them.

First, we note that the maximum likelihood estimator for \(\theta\) has a covariance matrix converging to the Fisher information matrix. This means that asymptotically, the maximum likelihood estimator is the most efficient in terms of its variance. However, this only holds asymptotically. It can be the case that for small samples, other estimators will outperform the maximum likelihood estimator. The bound also does not guarantee that no estimator can asymptotically be better than the maximum likelihood estimator. The performance of an estimator is usually described by its mean squared error, which is equal to the sum of the variance and the square of the bias. Since the bound is a statement for unbiased estimators, it leaves room for a biased estimator to have a smaller variance and thus a possibly smaller mean squared error than the (asymptotic) maximum likelihood estimator. Such estimators are unfortunately in many cases very hard to find.
Finally, the bound is given for the full parameter vector $\theta$. Now suppose we have an unbiased estimator $\tilde{\theta}_k$ for the $k$-th component of $\theta$. Expanding this into an unbiased estimator for $\theta$ by taking other unbiased components (such as from the maximum likelihood estimator), Cramér-Rao tells us that the variance of $\tilde{\theta}_k$ is at least equal to $k$-th diagonal element of the inverse Fisher information matrix. This means that even an unbiased estimator for a single component of $\theta$ cannot perform better than the inverse Fisher information. Again, this does not tell us that no unbiased estimator can perform better, but to find such an unbiased estimator is no easy task.

For the discussion in this section we assumed that the true likelihood function is known. However, if one can at best make an educated guess about the underlying data generating mechanism, then there might be a certain degree of misspecification. It is still possible to use a maximum likelihood estimator using this misspecified model, but then it is uncertain if asymptotic normality still holds, and what the asymptotic variance of the estimator is. This will be explored in the next section.

\subsection{Maximum likelihood estimation of misspecified models}

The topic of maximum likelihood estimation of misspecified models has been explored and outlined by White in \cite{8}. In this section, we will highlight some of the most important results and concepts given in the article, staying faithful to the notation used there. We will refer the interested reader to \cite{8} for all the mathematical (measure theoretical) details and proofs. Many of the conditions described in this section will be very similar to the regularity conditions for the maximum likelihood estimator described in the previous section.

First, we will introduce some notation. We assume the random vectors $\{U_t : 1 \leq t \leq n\}$ to have $m$ components, and a common distribution $G$ with density $g$. Further, assume we have a family of densities $\{f(\cdot, \theta) : \theta \in \Theta\}$ for some compact subset $\Theta$ of a $p$-dimensional Euclidean space, such that $f(u, \theta)$ is continuous in $\theta$ for every $u$. If $g = f(\cdot, \theta_0)$ for some $\theta_0 \in \Theta$, then we are in the case of no misspecification and we know (under a few conditions) what the asymptotic behaviour of the maximum likelihood estimator is. However, it is possible that we have chosen an incorrect family of densities.

In the case of misspecification, we are interested in whether the maximum likelihood estimator (also called the quasi-maximum likelihood estimator QMLE when the model might be misspecified) is asymptotically normally distributed. Also, one might wonder if the estimator converges, and to what value. Finally one might be interested in the asymptotic variance of the estimator.

The second question is the one that is first addressed in \cite{8}. First, the quasi-log-likelihood is naturally defined as

$$L_n(U, \theta) = \frac{1}{n} \sum_{t=1}^{n} \log(f(U_t, \theta)),$$

after which the QMLE is defined as

$$\hat{\theta}_n = \arg \max_{\theta \in \Theta} L_n(U, \theta).$$

Both definitions have the same familiar form as in the case of maximum likelihood estimation without misspecification. Intuitively, the maximum likelihood estimator is the parameter that makes the density fit best to the observed data. In the case of no misspecification, it is known that the estimator converges to the true parameter. Roughly speaking, the density $f(\cdot, \hat{\theta}_n)$ "shapes itself" like the true density $g = f(\cdot, \theta_0)$ as $n$ increases. If the model is misspecified,
obviously we can only hope that $f(\cdot, \hat{\theta}_n)$ approaches $g$ as closely as possible. It turns out that indeed, some notion of distance between $f(\cdot, \hat{\theta}_n)$ and $g$ is minimised as $n$ grows to infinity. More specifically, the QMLE converges to the parameter $\theta_*$ that minimises the Kullback-Leibler Information Criterion defined as

$$I(g : f, \theta) = \mathbb{E} \left( \log \left( \frac{g(U)}{f(U, \theta)} \right) \right),$$

where the expectation is taken with respect to the true density $g$.

**Theorem 2.4.** Assume $I(g : f, \theta)$ attains a unique minimum in $\theta_*$ and that $\mathbb{E}(\log(g(U)))$ exists. Also, assume there is a function $m$, that is integrable with respect to the distribution function $G$, such that $|\log(f(u, \theta))| \leq m(u)$ for all $\theta \in \Theta$. Then

$$\hat{\theta}_n \xrightarrow{a.s.} \theta_*.$$  (2.10)

After this, White goes on to show the asymptotic normality (under the required conditions) of the QMLE. In particular, the asymptotic variance is given, which is what we are most interested in (along with the consistency of the estimator) for the context of this text. First some matrices are defined (only where the denoted partial derivatives, expectations and inverses exist):

$$[A_n(\theta)]_{i,j} = \frac{1}{n} \sum_{t=1}^{n} \frac{\partial^2 \log f(U_t, \theta)}{\partial \theta_i \partial \theta_j},$$

$$[B_n(\theta)]_{i,j} = \frac{1}{n} \sum_{t=1}^{n} \frac{\partial \log f(U_t, \theta)}{\partial \theta_i} \frac{\partial \log f(U_t, \theta)}{\partial \theta_j},$$

$$[A(\theta)]_{i,j} = \mathbb{E} \left( \frac{\partial^2 \log f(U_t, \theta)}{\partial \theta_i \partial \theta_j} \right),$$

$$[B(\theta)]_{i,j} = \mathbb{E} \left( \frac{\partial \log f(U_t, \theta)}{\partial \theta_i} \frac{\partial \log f(U_t, \theta)}{\partial \theta_j} \right),$$

$$C_n(\theta) = A_n(\theta)^{-1} B_n(\theta) A_n(\theta)^{-1},$$

$$C(\theta) = A(\theta)^{-1} B(\theta) A(\theta)^{-1}.\quad (2.11)$$

With these definitions, White goes on to show that under some additional assumptions, a familiar form of asymptotic normality holds for the QMLE.

**Theorem 2.5.** Assume the conditions of Theorem 2.4 and further assume:

- The partial derivatives $\partial \log f(u, \theta)/\partial \theta_i$ are continuously differentiable functions of $\theta$ for every $u$.
- The functions $|\partial^2 \log f(u, \theta)/\partial \theta_i \partial \theta_j|$ and $|\partial \log f(u, \theta)/\partial \theta_i \partial \log f(u, \theta)/\partial \theta_j|$ are dominated by functions integrable with respect to $G$ for all $u$ and $\theta$.
- $\theta_*$ is an interior point of $\Theta$.
- $B(\theta_*)$ is nonsingular.
- $A(\theta)$ has constant rank in some open neighbourhood of $\theta_*$.  

Then the asymptotic normality for the QMLE holds in the form

$$\sqrt{n}(\hat{\theta}_n - \theta_*) \rightsquigarrow \mathcal{N}(0, C(\theta_*)),$$  (2.12)

and also $C_n(\hat{\theta}_n) \xrightarrow{a.s.} C(\theta_*)$ elementwise.

In the case of no misspecification\footnote{In fact, White introduces an additional assumption for this that we will not go into here.} so that $g = f(\cdot, \theta_0)$ for a $\theta_0 \in \Theta$, then $\theta_* = \theta_0$ and the identity $A(\theta_0) = -B(\theta_0)$ holds. Note that this identity is the well-known identity that we saw
in the previous section that lets us compute the Fisher information matrix in two ways. When
this holds, \( C(\theta_0) = -A(\theta_0) = B(\theta_0) \) is the inverse of the Fisher information matrix evaluated
in \( \theta_0 \), exactly as we saw before.

To compute the asymptotic variance, one would like to compute the matrix \( C \) in the pa-
rameter \( \theta_* \). However, the true density and thus \( \theta_* \) are unknown, and thus it is not feasible
to compute this matrix. Luckily, the above theorem offers a way out: for large \( n \), we can
simply estimate the variance matrix by \( C_n(\hat{\theta}_n) \). This only involves averages of derivatives of
the known densities \( f \), and can thus be computed without any knowledge of the true \( g \).

2.3 Regression theory

In this section, we will briefly highlight some of the concepts and theory used in regression
analysis. For proofs and more detailed descriptions of the underlying theory we refer to texts
such as Seber & Lee (2003) or lecture notes such as Cator (2012). We will assume
knowledge of basic properties of the multivariate normal distribution, which can be found in
many introductory texts on statistics and probability. We start by describing the widely used
linear model.

2.3.1 The linear model

Let \( Y \) be a random variable that is observed \( n \) times, for observations \( Y_1, ..., Y_n \), and for
each observation there some explanatory variables (fixed constants) \( x_{i,1}, ..., x_{i,r} \). Then we can
model

\[
Y_i = \beta_1 x_{i,1} + ... + \beta_r x_{i,r} + \epsilon_i \tag{2.13}
\]

for parameters \( \beta_1, ..., \beta_r \), where \( \epsilon_i \) follows a normal distribution with mean zero and variance
\( \sigma^2 \). Often, the first explanatory variable is set equal to one, in which case \( \beta_1 \) is also called the
intercept. The above is a system of \( n \) equations, which can be written in matrix form as

\[
\begin{pmatrix}
Y_1 \\
Y_2 \\
... \\
Y_n
\end{pmatrix} =
\begin{pmatrix}
x_{1,1} & x_{1,2} & ... & x_{1,r} \\
x_{2,1} & x_{2,2} & ... & x_{2,r} \\
... & ... & ... & ... \\
x_{n,1} & x_{n,2} & ... & x_{n,r}
\end{pmatrix}
\begin{pmatrix}
\beta_1 \\
\beta_2 \\
... \\
\beta_r
\end{pmatrix} +
\begin{pmatrix}
\epsilon_1 \\
\epsilon_2 \\
... \\
\epsilon_n
\end{pmatrix},
\]

or more shortly as \( Y = X\beta + \epsilon \). The matrix \( X \) is also referred to as the design matrix or input
matrix. Further, note that \( \epsilon \sim N(0, \sigma^2 I) \). This can be used to apply a maximum likelihood
procedure to estimate the unknown parameter \( \beta \). The likelihood function is maximised precisely
when \( \|Y - X\beta\|_2 \) is minimised. This can in turn be solved to find as the maximum likelihood
estimator

\[
\hat{\beta} = (X^T X)^{-1} X^T Y \tag{2.15}
\]

which is also the ordinary least squares estimator.

Let \( p = \text{dim} \{X\beta : \beta \in \mathbb{R}^r \} \), and define the residual sum of squares \( \text{SS}_{\text{res}} = \|Y - X\hat{\beta}\|_2 \).
Then it can be shown that

\[
\frac{\text{SS}_{\text{res}}}{\sigma^2} \sim \chi^2(n - p) \tag{2.16}
\]

where \( n \) and \( \sigma \) are as used above. Thus \( \text{SS}_{\text{res}}/(n - p) \) can be used as an unbiased estimator
for \( \sigma^2 \). For large \( n \), this is approximately equal to the average of the squared residuals \( Y_i - X_i\hat{\beta} \).

Furthermore, assuming the \( X \)-matrix is deterministic (meaning the only randomness is in
\( Y \)), it can be shown that

\[
\text{Var}(\hat{\beta}) = \sigma^2 (X^T X)^{-1} \tag{2.17}
\]
Note that this result only requires uncorrelated errors with equal variance $\sigma^2$. Under the assumption of normality, along with some other assumptions beyond the scope of this text (see the references mentioned above), it turns out that

$$\frac{\hat{\beta}_i - \beta_i}{\text{SE}(\hat{\beta}_i)} \sim t(n - p).$$  \hspace{1cm} (2.18)

There is also a more general version of this statement involving contrasts $c^T \hat{\beta}$ for $c \in \mathbb{R}^r$. The statement above is used to construct confidence intervals and for testing purposes, mostly to test whether or not certain variables are statistically significant.

### 2.3.2 Random Effects and Mixed Models

Not for all types of data it can be assumed that errors are uncorrelated. In the case of clinical trials, it might not be unreasonable to think that doing multiple measurements on the same individual produces correlated measurements. One way to expand the linear model to also accommodate correlated errors is to use a random effects model, also called a mixed model. These kinds of models are used when dealing with multilevel or clustered data, where it is assumed that measurements within each specific cluster are not independent. One specific example is longitudinal data, the subject of this text. Each vector of measurements (in this case every patient has their own vector of measurements) is independent from the other vectors, but the measurements within a single vector form one cluster of correlated measurements.

For the linear model, we assumed the parameters $\beta_1, \ldots, \beta_k$ to be fixed. Now, however, we will allow (some of) them to vary for each cluster. This means that two individuals with the exact same properties (explanatory variables) can have different parameters, for example each patient could have their own intercept (that is unknown and can only be estimated). In practice, it is assumed that the random parameters follow some distribution with their own set of parameters.

Intuitively, this means that there is some natural underlying distribution of parameters (intercepts, say) for the population, from which a parameter is sampled for each individual. In practice, any choice could be made, however one that is often used is a normal distribution. This means that instead of a fixed parameter $\beta_i$ we assume $\beta_{1,i} \sim N(\beta_1, \nu_1^2)$ where $\beta_{1,i}$ is the first parameter for the $i$-th individual. Instead of $\beta_{1,i}$, we could also write this as the sum of a fixed parameter $\beta_1$ and a random effect $\gamma_{1,i}$ that has normal distribution with mean zero and variance $\nu_1^2$. Thus for a model with one random effect, we would have

$$Y_{i,k} = \beta_{1,i}x_{i,1} + \beta_{2,i}x_{i,2} + \ldots + \beta_{r,i}x_{i,r} + \epsilon_{i,k}$$

$$= (\beta_1 + \gamma_{1,i})x_{i,1} + \beta_{2,i}x_{i,2} + \ldots + \beta_{r,i}x_{i,r} + \epsilon_{i,k}$$

$$= \beta_{1,i} + \beta_{2,i}x_{i,2} + \ldots + \beta_{r,i}x_{i,r} + \epsilon_{i,k} + \gamma_{1,i}x_{i,1}. \hspace{1cm} (2.19)$$

This means the expected outcome is the same linear combination of input variables as we saw in the linear model, but the variance properties are different. In the simple case of one random effect we have

$$\text{Cov}(Y_{i,k}, Y_{i,l}) = \text{Cov}(\epsilon_{i,k} + \gamma_{i,1}x_{i,1}, \epsilon_{i,l} + \gamma_{i,1}x_{i,1})$$

$$= \text{Cov}(\epsilon_{i,k}, \epsilon_{i,l}) + \text{Cov}(\gamma_{i,1}x_{i,1}, \gamma_{i,1}x_{i,1})$$

$$= x_{i,1}^2\nu_1^2 + 1_{(k=l)}\sigma^2, \hspace{1cm} (2.20)$$

which is easily generalised to multiple random effects. In the case of multiple random effects, these effects can even be assumed to be correlated with the other random effects. However, residuals and random effects are assumed to be independent.

Parameters are usually estimated through a (numerical) maximum likelihood or restricted maximum likelihood (REML) approach. The exact details are beyond the scope of this text,
but the difference between the two boils down to one of the approaches leading to biased estimates of the variance parameters, while for the other approach models are less easily compared. For more details, we refer to Diggle, Heagerty, Liang and Zeger (2002) [2]. In this text, we will only use maximum likelihood estimation. Now we move on to some concrete examples.

For these examples, we will assume a variable of interest (for example blood pressure) is followed throughout time for each individual. Further, we will assume that the behaviour of the variable is given by an intercept and slope parameter. For clarity, we will not use a treatment effect variable.

The first example we will look into is the random intercept model. In this case, every individual has the same slope, but start at different levels. An example can be found in the left panel of Figure 2.1. The model can be written as

\[ Y_{i,k} = \mu_i + \alpha t_{i,k} + \epsilon_{i,k}, \]  

(2.21)

where \( \mu_i \sim N(\mu, \tau^2) \) is independent of the errors \( \epsilon_{i,k} \sim N(0, \sigma^2) \), which are all independent from each other. The \( t_{i,k} \) are the time points for which measurements are done, and the slope \( \alpha \) is the same for each individual. We find

\[ \text{Cov}(Y_{i,k}, Y_{i,l}) = \tau^2 + 1_{\{k=l\}} \sigma^2, \]  

(2.22)

meaning that different measurements on the same individual are correlated with a correlation coefficient equal to \( \tau^2/\sigma^2 \). The consequence is that all measurements are equally strongly correlated, no matter the time separation between them. A large correlation is equivalent with \( \tau \) being large compared to \( \sigma \), which means most of the variation in the outcomes is caused by the random intercept. One example where the correlations are not all equal is the model with both a random intercept and a random slope.

A realisation of the random slope and intercept model is displayed in the right panel of Figure 2.1. We can write this model as

\[ Y_{i,k} = \mu_i + \alpha_i t_{i,k} + \epsilon_{i,k}, \]  

(2.23)

where the random effects \( \mu_i \sim N(\mu, \tau^2) \) and \( \alpha_i \sim N(\alpha, \nu^2) \) have \( \text{Cov}(\mu_i, \alpha_i) = \kappa \). Further, the i.i.d. errors \( \epsilon_{i,k} \sim N(0, \sigma^2) \) and the random effects are assumed to be independent. This leads to

\[ \text{Cov}(Y_{i,k}, Y_{i,l}) = \text{Cov}(\mu_i + \alpha_i t_{i,k} + \epsilon_{i,k}, \mu_i + \alpha_i t_{i,l} + \epsilon_{i,l}) \]
\[ = \tau^2 + \tau^2 \cdot \text{Cov}(t_{i,k}, t_{i,l}) + 1_{\{k=l\}} \sigma^2 + (t_{i,k} + t_{i,l}) \kappa. \]  

(2.24)
In particular, we see that both the variances and correlations change between different measurements and pairs of measurements, respectively.

2.4 Problem statement and notation

In this section, we will state the problem that we will explore in this text, as well as introduce the notation that will be used throughout. The problem studied in this text concerns longitudinal data, obtained from a clinical trial. This means that for each patient included in the trial, some variable of interest is measured throughout time. Such a variable could for example be blood pressure, an outcome linked to a questionnaire about back pain the patient fills in at multiple times, or the concentration of myoglobin in certain muscle cells.

The type of clinical trial studied is one with a single treatment, so that participating patients are split into a treatment and control group. Further, we that assume the variable of interest has a nice distribution, meaning the variable is continuous and normal distributions can be used to analyse the obtained data, without any transformations. More specifically, this is so we can use linear models and random effects models with Gaussian errors. This choice also makes some integrals we will come across tractable, which is helpful.

In the context of the clinical trials, we are interested in estimating the effect the treatment has on the outcome variable of interest. Since we are dealing with longitudinal data, this will be a behaviour throughout time. Thus the type of model that will be used to analyse the data will depend on the type of behaviour that is expected to occur. For example, it could be assumed that the treatment shifts the outcome variable by a certain, fixed amount. Then the treatment effect could be defined as the size of this sustained effect.

Another example is when the outcome variable is assumed to develop linearly throughout time, with the treatment changing the slope. In this case, a natural definition of the treatment effect would be the difference in slopes. These are two examples where the interpretation of the effects are very different, illustrating that in studying the time-dependent nature of the variable of interest, there is certainly no universal definition of the treatment effect. The treatment effect that is studied should be defined when designing the trial, and that is an area beyond the scope of this text.

In this text, we will analyse two different treatment effects, namely the two examples just given. In both cases, we will assume there is no development of the outcome variable in the control group, meaning that the placebo given to the patients in the control group has no effect at all. The placebo effect is an interesting phenomenon, and one might question our assumption of the placebo having no effect. However, it simplifies the computations and interpretation of the treatment effect, which is why we have chosen to take this approach.

After defining the treatment effect, there are many different ways to estimate the effect. The approach that is chosen will depend on the believed data generating mechanism, which is unknown and can be very complex in practice. The mixed model approach is a widely used approach when analysing clinical trials, as well as the ANCOVA (mixed model) approach, where post-treatment measurement are regressed on a pre-treatment measurement and treatment indicator.

Different approaches lead to different estimators, and in general these will all have different variances. In this text, we will focus on comparing the variances of a few unbiased estimators of the treatment effect, for the two different ground truth models. Some approaches used are naive, and would not be used in practice. These methods, ignoring the correlation between outcomes, were still included to see what happens under this kind of misspecification. The other methods accommodate correlated measurements and random effects, and could be used in practice when the assumed model seems reasonable. Note that all our approaches are max-
imum likelihood approaches for different assumed likelihood functions.

We would like to note that in this text we can only describe and highlight a very small fraction of all possibilities. The two chosen models are only two of the many thinkable examples, and similarly the few approaches we have chosen to compare are only a few of many. We will only look at some examples of maximum likelihood estimation, and left out other methods like generalised estimating equations, restricted maximum likelihood, and other options that accommodate correlated measurements. For a treatment of these methods in the case of longitudinal data, we refer to Diggle et al. (2012).

Now we will describe the notation that will be used throughout this work. We have $n$ patients, each with one baseline and $m$ follow-up measurements for the outcome variable of interest. From now on, we will use blood pressure as the outcome variable of interest as a concrete example. The baseline and follow-up blood pressures will be denoted as $Y_{i,0}$ and $Y_{i,1},...,Y_{i,m}$ respectively. Further, each patient has an indicator variable $\delta_i$ denoting whether the patient is in the treatment ($\delta_i = 1$) or control ($\delta_i = 0$) group. Thus each patient contributes the vector $(\delta_i, Y_{i,0}, Y_{i,1},...,Y_{i,m})$ to the data. Write $n_{\text{trt}} = \sum_{i=1}^{n} \delta_i$ as the number of treated patients, and let $q = n_{\text{trt}}/n$ be the fraction of all patients that are allocated to the treatment group. Finally, when using averages we will use the notation $\overline{A}$ as the average of all components of the vector $A$. 
CHAPTER 3

MODEL 1: RANDOM INTERCEP'TS AND SINGLE SUSTAINED TREATMENT EFFECT

In this chapter, we will define and analyse the first model. It is a relatively simple model, both conceptually and to analyse. We start by defining the model and commenting on the distribution of some averages of random variables. Then we will go on to use a regular linear model to compute some estimators and their distributions that are mostly interesting from a theoretical view, but not very useful in practice. After this, we move on to some more complicated methods involving random effects that are best analysed using numerical methods. Finally, we will compare the results found in all of the analyses.

3.1 The model

For the first model, we choose the random intercept model with the treatment effect defined as a fixed change. More specifically, we take

\[ \mu_i \sim N(\mu_0, \tau^2), \]
\[ Y_{i,0} | \mu_i \sim N(\mu_i, \sigma^2), \]
\[ Y_{i,k} | \mu_i \sim N(\mu_i + \beta \delta_i, \sigma^2) \quad \text{for} \quad k = 1, \ldots, m. \]  

(3.1)

This means that patient \( i \) has an underlying (characteristic) blood pressure \( \mu_i \), with a between-patient variation of \( \tau^2 \). This \( \tau \) is a measure of the spread in characteristic blood pressures between different individuals in the population. Further, the treatment effect is assumed to start somewhere between the baseline and first follow-up measurement, and then to be fixed over time, with an effect size equal to \( \beta \). Finally, \( \sigma^2 \) is the variance in measuring a blood pressure, which is equal to the sum of the within-patient variation and the measurement error. This model is also called the ‘random intercept’ model, since each patient has a random intercept. Note that the different measurements within a certain patient are correlated through having the same underlying blood pressure \( \mu_i \). The model is illustrated in Figure 3.1.

We can write \( Y_{i,0} = \mu_i + \epsilon_0 \), \( Y_{i,k} = \mu_i + \beta \delta_i + \epsilon_{i,k} \) where \( \epsilon_0 \) and \( \epsilon_{i,k} \) are independent and both normally distributed with mean zero and variance \( \sigma^2 \). From this, we see that

\[ \text{Var}(Y_{i,k}) = \text{Var}(\mu_i + \epsilon_{i,k}) = \sigma^2 + \tau^2 \quad \text{for} \quad k = 0, \ldots, m \]  

(3.2)
Figure 3.1: Illustration of the first model, where the follow-up measurements are done after \( t = 1 \)

and

\[
\text{Cov}(Y_{i,0}, Y_{i,k}) = \text{Cov}(\mu_i + \epsilon_{i,0}, \mu_i + \beta \delta_i + \epsilon_{i,k}) = \text{Var}(\mu_i) = \tau^2
\]

\[
\text{Cov}(Y_{i,j}, Y_{i,k}) = \text{Cov}(\mu_i + \beta \delta_i + \epsilon_{i,j}, \mu_i + \beta \delta_i + \epsilon_{i,k}) = \text{Var}(\mu_i) = \tau^2
\]

for \( k = 1, \ldots, m, j \neq k \). Note that outcomes from different patients are independent from each other.

### 3.2 Estimating treatment effect

To estimate the treatment effect, we can use different methods. First we will give some expressions for certain averages of outcomes, and then explore a few approaches using linear models. After this, we will move on to methods that better accommodate random effects.

#### 3.2.1 A note on averages

In the following sections, we will come across a number of averages of certain variables. From our chosen model, it is easy to compute the distribution of these averages and write for example the average baseline measurement for treated and untreated subjects as

\[
\bar{Y}^{\text{trt}}_0 = \mu_0 + \frac{\sqrt{\sigma^2 + \tau^2}}{\sqrt{n_{\text{trt}}}} Z^{\text{trt}}_0
\]

and

\[
\bar{Y}^{-\text{trt}}_0 = \mu_0 + \frac{\sqrt{\sigma^2 + \tau^2}}{\sqrt{n - n_{\text{trt}}}} Z^{-\text{trt}}_0
\]

respectively. Since the averages are taken over different individuals, the random variables \( Z^{\text{trt}}_0 \) and \( Z^{-\text{trt}}_0 \) are independent. Further, from (3.2) it is clear that the \( Z_0 \)-variables have a standard normal distribution. Note that this holds for any \( n \) by the choice of model (3.1), but it still holds asymptotically for a non-normal with the same mean and variance by the central limit theorem.
(as long as the distribution has a finite mean and variance). Finally, we will write
\[ Y_{\text{trt}}^{trt} = \frac{1}{n_{\text{trt}}} \sum_{i=1}^{m} \sum_{k=1}^{m} Y_{i,k} \]
\[ = \frac{1}{n_{\text{trt}}} \sum_{i=1}^{m} (\mu_i + \beta + \epsilon_{i,k}) \]
\[ = \beta + \frac{1}{n_{\text{trt}}} \sum_{i=1}^{m} \mu_i + \frac{1}{n_{\text{trt}}} \sum_{i=1}^{m} \epsilon_{i,k} \]
\[ = \mu_0 + \beta + \frac{\sqrt{\sigma^2/m + \tau^2}}{\sqrt{n_{\text{trt}}}} \]
\[ Z_{\text{trt}}^{trt} \] \quad (3.6)
and similarly
\[ \overline{Y}_{\text{trt}}^{trt} = \mu_0 + \frac{\sqrt{\sigma^2/m + \tau^2}}{\sqrt{n - n_{\text{trt}}}} \]
\[ Z_{\text{trt}}^{trt} \] \quad (3.7)
Here, the \( Z_{\text{trt}}^{trt} \)-variables again have a standard normal distribution, which follows from the chosen distributions of \( \mu_i \) and \( \epsilon_{i,k} \). However, the latter two \( Z \)-variables are not independent from the two defined above by the simple fact that two combinations of them concern the same individuals, for which all measurements depend on \( \mu_i \). From (3.6) and (3.7) it is not difficult to compute
\[ \text{Cov}(Z_{\text{trt}}^{trt}, Z_{\text{trt}}^{trt}) = \text{Cov}(Z_{\text{trt}}^{trt}, Z_{\text{trt}}^{trt}) = \frac{\tau^2}{\sqrt{\sigma^2/m + \tau^2}} \] \quad (3.8)
Similar computations can be done for averages of the first \( m \) measurements (every measurement apart from the final follow-up measurement). We find
\[ \overline{Y}_{\text{trt}}^{trt} = \mu_0 + m - 1 m \beta + \frac{\sqrt{\sigma^2/m + \tau^2}}{\sqrt{n_{\text{trt}}}} Z_{0,...,m-1}^{trt} \] \quad (3.9)
\[ \overline{Y}_{\text{trt}}^{trt} = \mu_0 + \frac{\sqrt{\sigma^2/m + \tau^2}}{\sqrt{n - n_{\text{trt}}}} Z_{0,...,m-1}^{trt} \] \quad (3.10)
Another average we will come across later is the average of the squares of all baseline measurements \( Y_{0}^{2} \). Using the moments of the normal distribution we find
\[ \overline{Y}_{0}^{2} = \sigma^2 + \tau^2 + \mu_0^2 + \sqrt{2(\sigma^2 + \tau^2)(\sigma^2 + \tau^2 + 2\mu_0^2)} Z_{0}^{eq} \] \quad (3.11)
where once again \( Z_{0}^{eq} \) has a standard normal distribution. Note that \( Z_{0}^{eq} \) is not independent from the previous \( Z_{0} \)-variables, however it is not necessary to know the specific covariance for our applications. A similar result holds for the square of the first \( m - 1 \) measurements:
\[ \overline{Y}_{0,...,m-1}^{2} = \sigma^2 + \tau^2 + \mu_0^2 + \frac{m - 1 m}{\sqrt{n}} \beta + \frac{\sqrt{\tau^2}}{\sqrt{n}} Z_{0,...,m-1}^{eq} \] \quad (3.12)
where we will not use the constant \( \kappa \) further, and \( Z_{0,...,m-1}^{eq} \) has a standard normal distribution. The correlation between this and the other \( Z \)-variables will also not be used.

3.2.2 Change scores

Since we are dealing with random effects, it seems like it may be desirable to get rid of the variance associated with the random effect through taking differences. More specifically, we can define
\[ \Delta_{i,k} := Y_{i,k} - Y_{i,0} \sim N(\beta \delta_i, 2\sigma^2) \] \quad (3.13)
for \( k = 1, ..., m \). By doing this, there is a tradeoff between increasing the variance by taking differences and decreasing it by losing the variance of the random effect. Note that \( \Delta_{i,k} \) and
\(\Delta_{i,j}\) are not independent for \(j \neq k\), even though the dependence on \(\mu_i\) is now gone! Using the notation from before, we have
\[
\Delta_{i,k} = \beta \delta_i + \epsilon_{i,k} - \epsilon_{i,0},
\]
so that
\[
\text{Cov}(\Delta_{i,k}, \Delta_{i,j}) = \text{Cov}(\beta \delta_i + \epsilon_{i,k} - \epsilon_{i,0}, \beta \delta_i + \epsilon_{i,j} - \epsilon_{i,0}) = \text{Var}(\epsilon_{i,0}) = \sigma^2.
\]

By taking differences we lost the random effect, which is a good thing if we want to apply a regular linear model. However, the above illustration of a non-zero covariance between measurements indicates that the independence assumption of the linear model is violated. This means that the linear model using change scores as we have defined them above may not be very applicable in practice. However, there is a way to transform the change score data in such a way that the linear model can be applied without violating any assumptions.

First, note that nonzero correlations between measurements only happen within the same individual. Focusing on one individual, with changes scores \(\Delta_i = (\Delta_{i,1}, ..., \Delta_{i,m})^T\), we have, according to the above,
\[
\Delta_i = \beta \delta_i \mathbf{1} + \epsilon_i,
\]
where we let \(\mathbf{1}\) denote the column vector of \(m\) ones, and \(I\) is the \(m \times m\) identity matrix. If we write
\[
T_0 = I + \frac{1}{m} \left( \frac{1}{\sqrt{m+1}} - 1 \right) \mathbf{1} \mathbf{1}^T,
\]
then it can be shown that \(T_0 \epsilon_i \sim N(0, \sigma^2 I)\) by noting that \(T_0(I + \mathbf{1} \mathbf{1}^T)T_0 = I\). Finally let \(T\) be the block-diagonal matrix with \(n\) equal \(m \times m\) blocks \(T_0\).

If one were to analyse change scores, there is a choice of whether or not to include an intercept alongside the treatment effect. Knowing the true model, we know no intercept is necessary, so we will first look into that case.

**Without intercept**

In this case we would like to assume
\[
\Delta_{i,k} = \beta' \delta_i + \eta'_{i,k},
\]
for \(k = 1, ..., m\), where \(\eta'_{i,k} \sim N(0, \sigma'^2)\) are all independent. We have used \(\eta\) instead of \(\epsilon\) to avoid confusion with the real underlying residuals of the outcomes \(\epsilon\). Note that we have used \(\beta'\) and \(\sigma'\) instead of \(\beta\) and \(\sigma\) to differentiate from the true parameters which are denoted by the latter set. However, in the future we will just use the latter notation as doing so will not lead to confusion.

As we have shown above, the change scores within an individual are correlated. To be accurate, it should be noted that instead, \(\eta_i\) has covariance matrix \(\sigma^2(I + \mathbf{1} \mathbf{1}^T)\). Multiplying by \(T_0\), we obtain the linear model (in vector notation, instead of componentwise)
\[
T_0 \Delta_i = \beta \delta_i T_0 \mathbf{1} + T_0 \epsilon_i = \frac{\beta}{\sqrt{m+1}} \delta_i \mathbf{1} + \eta'_i,
\]
with \(\eta'_i = T_0 \eta_i \sim N(0, \sigma'^2 I)\). The second equality can be computed from the definition of \(T_0\). This can be written as a system for all individuals as
\[
T \Delta = X \beta + \eta,
\]
with \(\eta \sim N(0, \sigma^2 I)\) and where \(\Delta\) is the vector containing all \(\Delta_{i,k}\). This means that we can now use the familiar linear model on this system. By taking the treatment group as the first \(n_{\text{trt}}\) patients, we can w.l.o.g. write the \(X\)-matrix as
\[
X = (1, 1, ..., 1, 0, ..., 0)^T / \sqrt{m+1},
\]

(3.14)
where the vector is a single column with \( n_{\text{trt}} m \) ones followed by \((n - n_{\text{trt}}) m\) ones. Then from regression analysis theory we know that the maximum likelihood estimator is given by

\[
\hat{\beta} = (X^TX)^{-1}X^T\Delta = (m + 1)(n_{\text{trt}} m)^{-1} \sum_{i=1}^{n_{\text{trt}}} \sum_{k=1}^{m} \frac{\Delta_{i,k}}{m + 1} = \Delta_{\text{trt}}. \tag{3.21}
\]

As a linear combination of normally distributed random variables, this estimator has a normal distribution. Its mean is equal to \( \beta \) by the above, while the variance can be computed through

\[
\text{Var}(\Delta_{\text{trt}}) = \text{Cov}(\Delta_{\text{trt}}, \Delta_{\text{trt}}) = \frac{1}{n_{\text{trt}}^2 m^2} \sum_{i=1}^{n_{\text{trt}}} \sum_{k=1}^{m} \left[ (m - 1)\sigma^2 + 2\sigma^2 \right] = \frac{n_{\text{trt}} m (m + 1) \sigma^2}{n_{\text{trt}}^2 m^2} = \frac{(1 - \rho)(m + 1)(\sigma^2 + \tau^2)}{n_{\text{trt}} m}. \tag{3.22}
\]

Now we will look into the variance reported by statistical software, which produces a covariance matrix

\[
\hat{\sigma}^2 (X^TX)^{-1}. \tag{3.23}
\]

The parameter \( \sigma^2 \) will be estimated by the average of the squared residuals (or more accurately this average multiplied by a factor tending to 1 as \( n \to \infty \)), with residuals in this case taking the form

\[
[T\Delta]_{i,k} - [T\Delta]_{i,k} = \beta \delta_i / \sqrt{m + 1} + \eta_{i,k} - \hat{\beta} \delta_i / \sqrt{m + 1} \approx \eta_{i,k}, \tag{3.24}
\]

where we used the expected value for \( \hat{\beta} \) to obtain a first order approximation. Asymptotically, the average of the squared residuals will be equal to the expected value \( \text{E}(\eta_{i,k}^2) = \sigma^2 \). Thus the reported variance will be approximately equal to

\[
\text{Var}(\hat{\beta}) = \frac{(m + 1) \sigma^2}{n_{\text{trt}} m} = \frac{(1 - \rho)(m + 1)(\sigma^2 + \tau^2)}{n_{\text{trt}} m}. \tag{3.25}
\]

Thus, the reported variance is close to the real variance of the estimator, getting closer as \( n \) increases.

**With intercept**

Instead of leaving out the intercept when analysing change scores, we can also choose to leave it in. This intercept would mean an overall shift between baseline and follow-up for any patient, regardless of whether or not they received a treatment. The treatment parameter then can be interpreted as an additional effect the treatment has on the outcome. Intuitively, as we do not need such a parameter, estimating another parameter on top of the necessary ones should not
improve the estimate for the other parameters. In the same notation as above, we would now like to model
\[ \Delta_{i,k} = \mu + \beta \delta_i + \epsilon_{i,k}, \] again for \( k = 1, \ldots, m \) and independent errors \( \epsilon_{i,k} \sim N(0, \sigma^2) \). However, we have seen that the independence assumption is violated. As input matrix for the equation above we have
\[
X = \begin{pmatrix}
1 & 1 & \cdots & 1 \\
1 & 1 & \cdots & 1 \\
\vdots & \vdots & \ddots & \vdots \\
1 & 0 & \cdots & 1 \\
\end{pmatrix}
\begin{array}{c}
\text{ntrt}\text{m rows} \\
\text{(n - nttr)m rows} \\
\end{array}
\] (3.27)
so that we obtain the equation
\[ T \Delta = TX \theta + \eta, \] (3.28)
where \( \eta \) satisfies the assumptions of the linear model, and \( \theta \) is the parameter vector \((1, \beta)^T\). It can be shown that \( TX = X/\sqrt{m+1} \), from which we can find
\[
(TX)^T X = \frac{m}{m+1} \begin{pmatrix} n & nttr \\ nttr & nttr \end{pmatrix}
\] (3.29)
\[
((TX)^T X)^{-1} = \frac{m+1}{nttr m (n - nttr)} \begin{pmatrix} nttr & -nttr \\ -nttr & n \end{pmatrix}
\] (3.30)
\[
((TX)^T X)^{-1} (TX)^T T \Delta = \begin{pmatrix} \Delta_{\text{trt}} \\ \Delta_{\text{trt}} - \Delta^{\text{trt}} \end{pmatrix}
\] (3.31)
Now note that as seen above, \( \Delta^{\text{trt}} \sim N(\beta, (m+1)\sigma^2/(nttr m)) \), and similarly \( \Delta_{\text{trt}} \sim N(0, (m+1)\sigma^2/((n - nttr)m)) \), so that for the maximum likelihood estimator of the treatment effect we have
\[ \hat{\beta} \sim N \left( \beta, \frac{(1-\rho)(m+1)(\sigma^2 + \tau^2)}{nttr m} \frac{n}{n - nttr} \right). \] (3.32)
Comparing the variance of this result to (3.22), we see that this approach yields more variability than not including the intercept: the variance of the estimator is a factor of \( n/(n - nttr) \) larger than the variance when not including an intercept. In practice, this would mean a variance about twice as large, due to the fact that the treatment and control groups will contain roughly the same number of patients. The observation that not including an intercept yields a more stable estimator is of course not unexpected, since the chosen model specifies no need for an intercept. The estimated standard error \( \hat{\sigma} \) will be the same as above since we estimate the intercept by something tending to zero, which results in (approximately) the same residuals and thus the same estimator as above. However, the second diagonal element of \((TX)^T X)^{-1}\) is equal to
\[ \frac{n(m+1)}{nttr m (n - nttr)} \]
which leads to an estimated variance of
\[ \hat{\text{Var}}(\hat{\beta}) = \frac{(1-\rho)(m+1)(\sigma^2 + \tau^2)}{nttr m} \frac{n}{n - nttr}, \] (3.33)
meaning this variance is also estimated consistently.
The downside of using change scores as outcomes is the doubling of the residual variance due to taking differences. If one believes the variance of the random effect to be reasonably small, it may possibly be beneficial to ignore the random effect and model the outcomes with an intercept and treatment effect. We will study this approach in the following section.

3.2.3 Outcomes with intercept and treatment effect

In this section, we assume

\[ Y_{i,k} = \mu + \beta \delta_i \mathbb{1}_{\{k > 0\}} + \epsilon_{i,k} \]  

(3.34)

for \( k = 0, \ldots, m \) and uncorrelated errors \( \epsilon_{i,k} \sim N(0, \sigma^2) \). Note that this independence assumption is again violated in practice. Also, the random intercept is deliberately ignored. These inaccuracies cause this model to not be the best or most realistic choice, but we look into it for the simple reason that we can easily obtain analytic expressions to compare with other analytic expressions of variances. For a description of this model where the random intercept is not ignored, see the section on the random intercept approach later in this chapter.

We have

\[
X = \begin{pmatrix}
1 & 1 & \cdots & 1 \\
0 & 0 & \cdots & 0 \\
1 & 1 & \cdots & 1
\end{pmatrix}_{n \times (m+1)}
\]

(3.35)

\[
X^T X = \begin{pmatrix}
n(m+1) & n_{trt} \times m \\
n_{trt} \times m & n_{trt} \times m
\end{pmatrix}
\]

(3.36)

\[
(X^T X)^{-1} = \frac{1}{n_{trt} \times m (n - n_{trt} m + n)} \begin{pmatrix}
n_{trt} \times m & -n_{trt} \times m \\
-n_{trt} \times m & n (m+1)
\end{pmatrix}
\]

(3.37)

\[
X^T Y = \begin{pmatrix}
n(m+1) Y_{0,\ldots,m} \\
n_{trt} \times m Y_{1,\ldots,m}
\end{pmatrix}
\]

(3.38)

from which we see

\[
\hat{\beta} = [(X^T X)^{-1} X^T Y)_2 = \frac{n(m+1)}{(n - n_{trt} m + n)} \left( Y_{1,\ldots,m} - Y_{0,\ldots,m} \right).
\]

(3.40)

We find (using the derived distributions of the averages)

\[
\hat{\beta} \sim N \left( \beta, \frac{(m+1)(\sigma^2 + \tau^2)}{qnm((1-q)m+1)} \left[ 1 + \frac{(1-q)m^2 - 1}{(1-q)m+1} \rho \right] \right)
\]

(3.41)

This estimator has smaller variance than the change score estimator if and only if

\[
\rho < \frac{(1-q)m+1}{(2-q)m+2}.
\]

(3.42)
For $q = 1/2$, we note that since $1/3 < (m + 2)/(3m + 4)$ for all $m \geq 1$, in this case using the outcomes is always preferable to using change scores for $\rho < 1/3$.

A further calculation shows that the estimator for the intercept has expectation $\mu_0$. This means that the residuals will be approximately equal to

$$Y_{i,k} - \hat{Y}_{i,k} = Y_{i,k} - \mu_0 - \beta\delta_i 1\{k > 0\} = (\mu_i - \mu_0) + \epsilon_{i,k},$$

for which the square has expectation $\sigma^2 + \tau^2$. Multiplying this by the second diagonal element of $(X^TX)^{-1}$ we find that the variance of $\hat{\beta}$ will approximately be estimated by statistical software as

$$\text{Var}(\hat{\beta}) = \frac{m + 1}{qnm((1 - q)m + 1)}.$$

Thus the true variance of the estimator is underestimated by a factor

$$\left(1 + \frac{(1 - q)m^2 - 1}{(1 - q)m + 1} \rho\right)^{-1}.$$

### 3.2.4 ANCOVA

For the ANCOVA, the follow-up measurements are regressed on the treatment indicator and baseline measurement. We make the same unrealistic assumptions as before in assuming

$$Y_{i,k} = \mu + \beta\delta_i + \alpha Y_{i,0} + \epsilon_{i,k},$$

for $k = 1,\ldots,m$, with independent $\epsilon_{i,k} \sim N(0, \sigma^2)$. Later in this chapter, we will use the ANCOVA without ignoring the random intercept. For now, the $X$-matrix takes the form

$$X = \begin{pmatrix}
1 & 1 & Y_{1,0} \\
\vdots & \vdots & \vdots \\
1 & 1 & Y_{n_{trt},0}
\end{pmatrix}
\begin{pmatrix}
1 \ 1 \ Y_{1,0} \\
\vdots \ \vdots \ \vdots \\
1 \ 1 \ Y_{n_{trt},0} \\
1 \ 0 \ Y_{n_{trt},+1,0} \\
\vdots \ \vdots \\
1 \ 0 \ Y_{n,0}
\end{pmatrix}
\begin{pmatrix}
m \text{ rows}
\end{pmatrix}$$

This gives us

$$X^TX = \begin{pmatrix}
n_{trt} & n_{trt} & n_{trt} Y_{0} \\
n_{trt} & n_{trt} & n_{trt} Y_{0} \\
n_{trt} Y_{0} & n_{trt} Y_{0} & n_{trt} Y_{0} Y_{0}^{-1}
\end{pmatrix}
\begin{pmatrix}
1 & q & q Y_{0} Y_{0}^{-1}
q & q & q Y_{0} Y_{0}^{-1}
q Y_{0} Y_{0}^{-1} & q Y_{0} Y_{0}^{-1} & Y_{0} Y_{0}^{-1}
\end{pmatrix}.$$

$$= nm \begin{pmatrix}
1 & q & q Y_{0} Y_{0}^{-1} + (1 - q) Y_{0} Y_{0}^{-1}
q & q & q Y_{0} Y_{0}^{-1}
q Y_{0} Y_{0}^{-1} & q Y_{0} Y_{0}^{-1} & Y_{0} Y_{0}^{-1}
\end{pmatrix}.$$
where \( q = n_{trt}/n \). By the results of Section 3.2.1, we can write the matrix in the equation above as

\[
\begin{pmatrix}
1 & q & \mu_0 \\
q & q & q\mu_0 \\
\mu_0 & q\mu_0 & \sigma^2 + \tau^2 + \mu_0^2
\end{pmatrix}
\begin{pmatrix}\epsilon \\0 \end{pmatrix}
+ \epsilon
\begin{pmatrix}
0 & 0 & \sqrt{q}Z_{trt}^{0.5} + \sqrt{1-q}Z_0^{0.5} \\
0 & 0 & \sqrt{q}Z_{trt}^{0.5} + \sqrt{1-q}Z_0^{0.5} \\
\sqrt{q}Z_{trt}^{0.5} + \sqrt{1-q}Z_0^{0.5} & \sqrt{q}Z_{trt}^{0.5} + \sqrt{1-q}Z_0^{0.5} & \sqrt{2}\sigma^2 + \tau^2 + 2\mu_0^2Z_0^2
\end{pmatrix}
\begin{pmatrix}\epsilon \\0 \end{pmatrix}
\]

where \( \epsilon = \sqrt{\sigma^2 + \tau^2}/\sqrt{n} \). The next object we need is \( X^TY \), which we can compute to be

\[
\begin{pmatrix}
qY_{1,...,m} + (1-q)Y_{trt}^{0.5} \\
qY_{1,...,m} \\
Y_{01,...,m}
\end{pmatrix}
\]

Here \( Y_{01,...,m} \) is the average of the products of baseline and follow-up measurements. In the same vein as in Section 3.2.1, we will write

\[
Y_{01,...,m} = \mu_0^2 + \tau^2 + q\beta\mu_0 + \frac{C}{\sqrt{n}}Z_x.
\]

However, it will turn out that this term vanishes later and as such we will not put any effort into computing \( C \) or the covariance between \( Z_x \) and the other \( Z \)-variables. We will now write

\[
X^TY = \begin{pmatrix}
\mu_0 + q\beta \\
q\mu_0 + q\beta \\
\mu_0^2 + \tau^2 + q\beta\mu_0
\end{pmatrix}
+ \frac{1}{\sqrt{n}} \begin{pmatrix}
\sqrt{\sigma^2/m + \tau^2} (\sqrt{q}Z_{1,...,m}^{0.5} + \sqrt{1-q}Z_1^{0.5}) \\
\sqrt{\sigma^2/m + \tau^2} \sqrt{q}Z_{1,...,m}^{0.5} \\
CZ_x
\end{pmatrix}
\]

Note that we have used an equality sign above, although it is strictly speaking an approximate equality. However, we will now move to an asymptotic context and let \( n \to \infty \). Thus from now on we will use equality signs. Using a linear model, the maximum likelihood estimator for \( \beta \) will be

\[
\hat{\beta} = [X^TX]^{-1}X^TY = [(B_0 + \epsilon A)^{-1}(D_0 + D_1)/\sqrt{n}]_{2}. \tag{3.53}
\]

For small \( \epsilon \), we have \( (B_0 + \epsilon A)^{-1} \approx B_0^{-1} - \epsilon B_0^{-1}AB_0^{-1} \). If we ignore terms of order smaller than \( n^{-1/2} \), we obtain

\[
\beta = [B_0^{-1}D_0]_2 - \epsilon [B_0^{-1}AB_0^{-1}D_0]_2 + [B_0^{-1}D_1]_2/\sqrt{n} \tag{3.54}
\]

as \( n \to \infty \). Note that the vector \( B_0^{-1}D_0 \) is the first order term of the estimator of all three parameters, and not difficult to compute. We have

\[
B_0^{-1} = \begin{pmatrix}
\frac{\mu_0^2}{\sigma^2 + \tau^2} + \frac{1}{1-q} & -\frac{1}{1-q} & -\frac{\mu_0}{\sigma^2 + \tau^2} \\
-\frac{1}{1-q} & \frac{1}{1-q} & 0 \\
-\frac{\mu_0}{\sigma^2 + \tau^2} & 0 & \frac{1}{\sigma^2 + \tau^2}
\end{pmatrix} \tag{3.55}
\]

which leads to

\[
B_0^{-1}D_0 = \begin{pmatrix}
\mu_0 \frac{\sigma^2}{\sigma^2 + \tau^2} \\
\beta \\
\beta \frac{\sigma^2}{\sigma^2 + \tau^2}
\end{pmatrix} = \begin{pmatrix}
\mu_0 (1-\rho) \\
\beta \\
\beta \rho
\end{pmatrix} \tag{3.56}
\]
where \( \rho = \frac{\tau^2}{\sigma^2 + \tau^2} = \text{Cov}(Y_{i,l}, Y_{i,k})/\sqrt{\text{Var}(Y_{i,l})\text{Var}(Y_{i,k})} = \text{Corr}(Y_{i,l}, Y_{i,k}) \) is the correlation between measurements on the same individual. An interesting conclusion we can draw from this is that the estimator for the intercept is not unbiased, but is in fact biased toward zero depending on the correlation between measurements. Fortunately, the estimator for the treatment effect turns out to be unbiased. Now we will go on to compute the first order behaviour of the MLE \( \hat{\beta} \).

For the term \([B_0^{-1} A B_0^{-1} D_0]_2\), we only need the second row of \( B_0^{-1} A B_0^{-1} \), for which we multiply the second row of \( B_0^{-1} \) with \( A B_0^{-1} \). Note that due to the zero in the third position of the former, only the first two rows of \( A B_0^{-1} \) need to be computed. Using this, we find the second row of \( B_0^{-1} A B_0^{-1} \) to be equal to

\[
\frac{1}{\sigma^2 + \tau^2} \left( Z_{0 \text{trt}}^{\frac{1}{\sqrt{q}}} - \frac{Z_{0 \text{trt}}^{\frac{1}{\sqrt{1-q}}}}{\sqrt{1-q}} \right) \left( -\mu_0, 0, 1 \right),
\]

from which we compute

\[
[B_0^{-1} A B_0^{-1} D_0]_2 = \rho \left( Z_{0 \text{trt}}^{\frac{1}{\sqrt{q}}} - \frac{Z_{0 \text{trt}}^{\frac{1}{\sqrt{1-q}}}}{\sqrt{1-q}} \right).
\]

The other term is a straightforward calculation which results in

\[
[B_0^{-1} D_1]_2 = \sqrt{\sigma^2/m + \tau^2} \left( Z_{1 \text{trt}, m}^{\frac{1}{\sqrt{q}}} - \frac{Z_{1 \text{trt}, m}^{\frac{1}{\sqrt{1-q}}}}{\sqrt{1-q}} \right).
\]

In conclusion, for the MLE of the treatment effect we find

\[
\hat{\beta} = \beta + \sqrt{\frac{1}{n}} \sqrt{\frac{\sigma^2}{m + \tau^2}} \left( Z_{1 \text{trt}, m}^{\frac{1}{\sqrt{q}}} - \frac{Z_{1 \text{trt}, m}^{\frac{1}{\sqrt{1-q}}}}{\sqrt{1-q}} \right) - \sqrt{\frac{1}{n}} \sqrt{\frac{\tau^2}{\sigma^2 + \tau^2}} \left( Z_0^{\frac{1}{\sqrt{q}}} - \frac{Z_0^{\frac{1}{\sqrt{1-q}}}}{\sqrt{1-q}} \right).
\]

Now note that we have shown before that

\[
\begin{pmatrix}
Z_{0 \text{trt}}^{\frac{1}{\sqrt{q}}}
Z_{1 \text{trt}, m}^{\frac{1}{\sqrt{q}}}
Z_{0}^{\frac{1}{\sqrt{q}}}
Z_{0 \text{trt}}^{\frac{1}{\sqrt{1-q}}}
Z_{1 \text{trt}, m}^{\frac{1}{\sqrt{1-q}}}
Z_{0}^{\frac{1}{\sqrt{1-q}}}
\end{pmatrix} \sim \mathcal{N}
\begin{pmatrix}
0 \\
0 \\
0 \\
0 \\
0 \\
0
\end{pmatrix}
\begin{pmatrix}
1 & 0 & \gamma & 0 \\
0 & 1 & 0 & \gamma \\
\gamma & 0 & 1 & 0 \\
0 & \gamma & 0 & 1
\end{pmatrix}
\]

where

\[
\gamma := \text{Cov}(Z_0^{\text{trt}}, Z_{1 \text{trt}, m}) = \text{Cov}(Z_0^{-\text{trt}}, Z_{1 \text{trt}, m}) = \frac{\tau^2}{\sqrt{\sigma^2/m + \tau^2}} \frac{1}{\sqrt{1-q}} \frac{1}{\sqrt{1-q}}
\]

Using this, we finally compute

\[
\hat{\beta} \sim \mathcal{N}
\begin{pmatrix}
\beta \\
\sigma^2 + \tau^2
\end{pmatrix}
\begin{pmatrix}
\left[ 1 + (m - 1)\rho - m\rho^2 \right]^{-1} \\
\left[ 1 + (m - 1)\rho - m\rho^2 \right]^{-1/2}
\end{pmatrix}
\]

Dividing the variance of the ANCOVA estimator by that of the change score estimator (without intercept) we obtain the ratio

\[
1 + m\rho
1 - q m + 1
\]

The ANCOVA provides a more stable estimator than the change scores approach precisely when the ratio above is smaller than 1. This is the case if and only if

\[
\rho < \frac{(1 - q)(m + 1) - 1}{m} = 1 - q \frac{q}{m} < 1 - q.
\]
This means that for correlations at least $1 - q$ it is always a better idea to use the changes from baseline as outcomes. If one prefers to still add an intercept, the ratio becomes at most 1, meaning that the ANCOVA is always at least as stable as the analysis of changes when including an intercept. The advantage gained by using the ANCOVA increases when $\rho$ decreases. Comparing the variance of the ANCOVA estimator and that of the outcomes estimator \[3.41\], we obtain the ratio

\[
\frac{(1 - q)m + 1}{(1 - q)(m + 1)} \frac{1 + (m - 1)\rho - m\rho^2}{1 + \frac{(1 - q)m^2 - 1}{(1 - q)m + 1}\rho},
\]

where a ratio smaller than 1 means the outcomes estimator is more stable than the ANCOVA estimator. For general $q$, there is no short expression for the interval of $\rho$ where the inequality holds true. However, for $q = 1/2$ we can obtain (assuming $\rho \geq 0$)

\[
\rho < \frac{m^2 + m - 1 + \sqrt{2m^4 + 8m^2 + 11m^2 + 6m + 1}}{m(m + 2)^2}.
\]

The fist few values of the above ratio can be found in Table 3.1. Note that for large $m$, \[3.67\] tends to 0 as $1/m$. This means that in the hypothetical situation where we have an extreme number of follow-up visits for each patient, ANCOVA will always be better than just regressing the outcomes on treatment. However, as is clear from Table 3.1, this is not the case for small $m$ and $\rho$.

The reported variance of $\hat{\beta}$ will be equal to $\hat{\sigma}^2[(X^TX)^{-1}]_{2,2} \approx \hat{\sigma}^2[B_0^{-1}]_{2,2}/nm$. The residuals take the form

\[
Y_{i,k} - \hat{Y}_{i,k} = \mu_i + \beta\delta_i + \epsilon_{i,k} - \mu_0(1 - \rho) - \rho Y_{i,0}
\]

\[
= \mu_i + \beta\delta_i + \epsilon_{i,k} - \mu_0(1 - \rho) - \rho\mu_i - \rho\epsilon_{i,0}
\]

\[
= (1 - \rho)[\mu_i - \mu_0] + \epsilon_{i,k} - \rho\epsilon_{i,0},
\]

after which taking the expectation of the square yields

\[
\hat{\sigma}^2 \approx \sigma^2 + (1 - \rho)^2\tau^2 + \rho^2\sigma^2 = \sigma^2 + (1 - \rho)^2\tau^2,
\]

which gets more accurate as $n \to \infty$. Here we expanded $(1 - \rho)^2$ and used the identity $(\sigma^2 + \tau^2)\rho^2 = \tau^2\rho$. Thus, the estimated variance of $\hat{\beta}$ is (asymptotically)

\[
\hat{\text{Var}}(\hat{\beta}) = \frac{\sigma^2}{nmq(1 - q)} \left[ 1 + \frac{(1 - \rho)^2\tau^2}{\sigma^2} \right]
\]

\[
= \frac{(1 + \rho)\sigma^2}{nmq(1 - q)}
\]

\[
= \frac{(1 - \rho^2)(\sigma^2 + \tau^2)}{nmq(1 - q)}.
\]

Thus the variance is underestimated by a factor

\[
\frac{1 + \rho}{1 + m\rho},
\]

independent of $q$.}

Table 3.1: Correlation thresholds after which the ANCOVA estimator is more stable than \[3.41\] for some values of $m$ and $q = 1/2$

<table>
<thead>
<tr>
<th>$m$</th>
<th>Threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.6990558</td>
</tr>
<tr>
<td>2</td>
<td>0.5427912</td>
</tr>
<tr>
<td>3</td>
<td>0.4436141</td>
</tr>
<tr>
<td>4</td>
<td>0.3750000</td>
</tr>
<tr>
<td>5</td>
<td>0.3247220</td>
</tr>
</tbody>
</table>

23
3.2.5 Autoregressive approach

The autoregressive approach regresses the outcomes on treatment indicator and previous measurement, and is thus in a sense similar to the ANCOVA. We assume

\[ Y_{i,k} = \mu + \beta \delta_i + \alpha Y_{i,k-1} + \epsilon_{i,k}, \]  

making the same (unrealistic) assumptions as we did before. This type of model was used in a comparative analysis in Twisk & de Vente (2008) [7]. Before that, it was described in more detail and illustrated with an example in Rosner, Muñoz, Tager, Speizer & Weiss (1985) [5]. It is not obvious that this approach would lead to an unbiased estimation of the treatment effect, since there is only an effect between the baseline and first follow-up measurement. The second follow-up measurement will always be around the level of the first follow-up, however the autoregressive approach as given does not accommodate this. Therefore we might expect the true treatment effect to be underestimated.

To show the bias, we start with

\[
X = \begin{pmatrix}
1 & 1 & Y_{1,0} \\
& \vdots & \vdots \\
1 & 1 & Y_{n_{\text{trt}},0} \\
& \vdots & \vdots \\
1 & 1 & Y_{n_{\text{trt}},m-1} \\
0 & \vdots & \vdots \\
1 & 0 & Y_{n_{\text{trt}},m+1,0} \\
& \vdots & \vdots \\
1 & 0 & Y_{n_{\text{trt}},m-1}
\end{pmatrix},
\]  

which leads to

\[
X^T X \approx nm \begin{pmatrix}
1 & q & qY_{0,\ldots,m-1}^{\text{trt}} + (1-q)Y_{0,\ldots,m-1}^{\text{trt}} \\
q & q & qY_{0,\ldots,m-1}^{\text{trt}} + (1-q)Y_{0,\ldots,m-1}^{\text{trt}} \\
qY_{0,\ldots,m-1}^{\text{trt}} + (1-q)Y_{0,\ldots,m-1}^{\text{trt}} & qY_{0,\ldots,m-1}^{\text{trt}} & qY_{0,\ldots,m-1}^{\text{trt}} + (1-q)Y_{0,\ldots,m-1}^{\text{trt}}
\end{pmatrix},
\]  

where the averages are taken over all measurements (including baseline) apart from the final follow-up measurement. For the autoregressive approach, we will first only consider the first order approximation of the estimated treatment effect. Replacing averages with expected values, we obtain

\[
X^T X \approx nm \begin{pmatrix}
1 & q & \frac{m-1}{m}q\beta + \mu_0 \\
q & q & \frac{m-1}{m}q\beta + q\mu_0 \\
\frac{m-1}{m}q\beta + \mu_0 & \frac{m-1}{m}q\beta + q\mu_0 & \frac{m-1}{m}q^2 + \sigma^2 + \tau^2 + \mu_0^2 + q\beta \frac{m-1}{m}[\beta + 2\mu_0]
\end{pmatrix},
\]  

which has inverse
\[
\begin{pmatrix}
\frac{1}{n} \int \frac{1}{\theta^q} + \frac{\mu_0^2}{\sigma^2 + \tau^2} q(m-1) \\
\frac{1}{n} \int \frac{1}{\theta^q} + \frac{(m-1)\beta\mu_0/m}{\sigma^2 + \tau^2} q(m-1) \\
\frac{1}{n} \int \frac{1}{\theta^q} + \frac{(m-1)\beta\mu_0/m}{\sigma^2 + \tau^2} q(m-1) \\
\frac{1}{n} \int \frac{1}{\theta^q} + \frac{(m-1)\beta\mu_0/m}{\sigma^2 + \tau^2} q(m-1)
\end{pmatrix}
\]

(3.76)

Similarly, we can compute

\[
X^TY \approx nm \begin{pmatrix}
\mu_0 + q\beta \\
q\mu_0 + q\beta \\
\mu_0^2 + \tau^2 + q(2 - \frac{1}{m})\beta\mu_0 + \frac{m-1}{m}q\beta^2
\end{pmatrix}
\]

(3.77)

from which we find

\[
\hat{\beta} = [(X^TX)^{-1}X^TY]_2 \approx \beta \left[ 1 - \rho \left( \frac{m}{m-1} + \frac{q\beta^2}{m(\sigma^2 + \tau^2)} \right)^{-1} \right].
\]

(3.78)

Note that this estimator is biased and it would thus be unwise to use the autoregressive approach to estimate the treatment effect. For this reason, we will not consider this estimator as suitable and choose to exclude it from further theoretical analyses.

### 3.2.6 Random intercept model

In this section we will finally do a full treatment of the random intercepts. We will look at the mixed model approach, using the maximum likelihood method to estimate the parameters. Note that this is the true model that we have assumed for this chapter. Thus we assume

\[
Y_{i,k} = \mu_i + \beta \delta_{i1_{k>0}} + \epsilon_{i,k}
\]

(3.79)

for \( k = 0, \ldots, m \) with independent \( \epsilon_{i,k} \sim N(0, \sigma^2) \) as well as \( \mu_i \sim N(\mu_0, \tau^2) \) independent from the \( \epsilon \). The likelihood for this model (being the one we assumed in (3.1)) can be written as

\[
L(\hat{\theta}) = \prod_{i=1}^{m} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi} \sigma} e^{-\frac{(y_{i,k} - \mu_i)^2}{2\sigma^2}} \prod_{k=0}^{m} \frac{1}{\sqrt{2\pi} \tau} e^{-\frac{(y_{i,k} - \mu_i - \beta \delta_{i1_{k>0}})^2}{2\tau^2}} d\mu_i
\]

\[
= (2\pi)^{-nm/2} \sigma^{-nm(1)} \tilde{\sigma}^{-n} \prod_{i=1}^{m} \int_{-\infty}^{\infty} e^{-\frac{1}{2} \sum_{k=0}^{m} (y_{i,k} - \mu_i - \beta \delta_{i1_{k>0}})^2 / \sigma^2 + (y_{i,k} - \mu_0)^2 / \tau^2} d\mu_i,
\]

(3.80)

where \( \tilde{\theta} \) denotes the vector \((\tilde{\mu}_0, \tilde{\beta}, \tilde{\sigma}, \tilde{\tau})\). Using the Gaussian integral, we can compute the identity

\[
\int_{-\infty}^{\infty} e^{-\frac{1}{2} \sum_{j=0}^{m} a_j (y_{i,j} - \beta \delta_{i1_{j>0}})^2} d\mu_j = \sqrt{\sum_{j=0}^{m} a_j} e^{\frac{1}{2} \sum_{j=0}^{m} a_j} e^{-\frac{1}{2} \sum_{j=0}^{m} a_j a_j},
\]

(3.81)

which we will use to compute the log-likelihood

\[
\ell(\hat{\theta}) = -\frac{n(m+1)}{2} \log(2\pi) - n(m+1) \log(\tilde{\sigma}) - n \log(\tilde{\tau}) - \frac{n}{2} \log \left( \frac{m+1}{\tilde{\sigma}^2} + \frac{1}{\tilde{\tau}^2} \right)
\]

\[
+ \frac{\tilde{\sigma}^2 + \tilde{\tau}^2}{2(m+1)\tilde{\tau}^2 + 2\tilde{\sigma}^2} \sum_{i=1}^{n} \left( \frac{1}{\tilde{\sigma}^2} \sum_{k=0}^{m} (y_{i,k} - \mu_0 - \beta \delta_{i1_{k>0}})^2 + \frac{\mu_0}{\tilde{\tau}^2} \right)^2
\]

\[
- \frac{1}{2\tilde{\sigma}^2} \sum_{i=1}^{n} \sum_{k=0}^{m} (y_{i,k} - \beta \delta_{i1_{k>0}})^2 - \frac{\mu_0^2}{2\tilde{\tau}^2}.
\]

(3.82)
Taking expectations yields

\[
E(\ell(\tilde{\theta})) = -\frac{n(m + 1)}{2} \log(2\pi) - n(m + 1) \log(\tilde{\sigma}) - n \log(\tilde{\tau}) - \frac{n}{2} \log \left(\frac{m + 1}{\tilde{\sigma}^2} + \frac{1}{\tilde{\tau}^2}\right) \\
+ \frac{\tilde{\sigma}^2 \tilde{\tau}^2}{2(m + 1)\tilde{\sigma}^2 + 2\tilde{\sigma}^2} \left[ \frac{n(m + 1)^2(\mu_0^2 + \tau^2)}{\tilde{\sigma}^4} + 2n_{\text{trt}}m(m + 1)\mu_0(\beta - \tilde{\beta}) \right] \\
+ \frac{2n(m + 1)\beta_0\mu_0}{\tilde{\sigma}^2 + \tilde{\tau}^2} + \frac{n(m + 1)\sigma^2}{\tilde{\sigma}^4} + \frac{2n_{\text{trt}}m\beta_0(\beta - \tilde{\beta})}{\tilde{\tau}^2} + \frac{n\mu_0^2}{\tilde{\tau}^4} \\
- \frac{1}{2\tilde{\tau}^2} \left( n(m + 1)(\mu_0^2 + \tau^2 + \sigma^2) + 2n_{\text{trt}}m\beta_0(\beta - \tilde{\beta}) + n_{\text{trt}}m(\beta - \tilde{\beta})^2 \right) - \frac{n\mu_0^2}{2\tilde{\tau}^2}. 
\]

(3.83)

To obtain the asymptotic variance of the estimator \(\hat{\beta}\), we compute the Fisher information matrix by computing the second derivative of (3.83) with respect to \(\tilde{\theta}\) in the true parameter \(\theta\) and multiplying by -1. Due to the nature of the expression, it does not seem feasible to do this analytically, since even the first derivatives with respect to \(\tilde{\sigma}\) and \(\tilde{\tau}\) become quite complex expressions. Instead, we will opt to use a numerical approach. Note that for a vector \(\delta a\) and small \(\delta\), we have (using that \(\ell\) is maximised in \(\theta\), and thus that the gradient is equal to the zero vector in \(\theta\))

\[
\ell(\theta + \delta a) \approx \ell(\theta) + \frac{\delta^2}{2} a^T D^2 \ell(\theta) a.
\]

The matrix multiplication on the r.h.s. is equal to

\[
\sum_{i,j} a_i D^2 \ell(\theta)_{i,j} a_j = \sum_i a_i^2 D^2 \ell(\theta)_{i,i} + 2 \sum_{i<j} a_i a_j D^2 = \ell(\theta)_{i,j} = \text{vec}(a)^T \text{vec}(D^2 \ell(\theta)),
\]

where (noting that \(a\) is a vector of length 4)

\[
\text{vec}(a) = (a_1^2, a_2^2, a_3^2, a_4^2, 2a_1 a_2, 2a_1 a_3, 2a_1 a_4, 2a_2 a_3, 2a_2 a_4, 2a_3 a_4)^T,
\]

\[
\text{vec}(D^2 \ell(\theta)) = (D^2 \ell(\theta)_{1,1}, D^2 \ell(\theta)_{2,2}, D^2 \ell(\theta)_{3,3}, D^2 \ell(\theta)_{4,4}, D^2 \ell(\theta)_{1,2}, D^2 \ell(\theta)_{1,3}, D^2 \ell(\theta)_{1,4}, D^2 \ell(\theta)_{2,3}, D^2 \ell(\theta)_{2,4}, D^2 \ell(\theta)_{3,4})^T.
\]

(3.84) (3.85)

Using multiple direction vectors \(a^{(1)}, ..., a^{(p)}\) we obtain a system of equations

\[
\begin{pmatrix}
\text{vec}(a^{(1)})^T \\
\text{vec}(a^{(2)})^T \\
\vdots \\
\text{vec}(a^{(p)})^T
\end{pmatrix}
\text{vec}(D^2 \ell(\theta)) = \frac{2}{\delta^2}
\begin{pmatrix}
\ell(\theta + \delta a^{(1)}) - \ell(\theta) \\
\ell(\theta + \delta a^{(2)}) - \ell(\theta) \\
\vdots \\
\ell(\theta + \delta a^{(p)}) - \ell(\theta)
\end{pmatrix}.
\]

(3.86)

From this we can obtain \(\text{vec}(D^2 \ell(\theta))\) and thus the approximated Fisher information matrix by a least squares procedure. More specifically, we use \(\delta = 10^{-4}\) and randomly generated direction vectors. First, we generate five hundred vectors of length four with every component distributed uniformly on \([-1, 1]\), independent from all other components. This set of direction vectors is combined with the set of vectors in the opposite directions, obtained by multiplying each vector by -1, to obtain a set of one thousand randomly generated direction vectors. Using this, we can compute \(\text{Var}(\hat{\beta})\) for any choice of \(\theta\).

In the case of the simple linear model we used before, we saw that for \(n_{\text{trt}} = n/2\), \(n\text{Var}(\hat{\beta})/(\sigma^2 + \tau^2)\) is only a function of \(m\) and the correlation \(\rho\). Therefore, plotting \(n\text{Var}(\hat{\beta})/(\sigma^2 + \tau^2)\) as a function of \(\rho\) for fixed \(m\) and different choices of \(\mu_0, \beta, \sigma^2 + \tau^2\) and \(n\), we hope to see the same\(^1\)

\(^1\)Note that we have exchanged the order of the second derivative and expectation operators.
curves every time. Although it is not a formal proof, we have observed this to be the case by making a few plots.

The downside of this numerical approach is that we do not obtain a closed analytic expression that we can compare directly with other such expressions we have already obtained. However, the same information can still be shown graphically by making the plots just described.

### 3.2.7 ANCOVA WITH RANDOM INTERCEPT

Instead of using all measurements (including the baseline) as outcomes, we may also use the baseline measurement for an individual as an input variable. We have done this before with the ANCOVA approach, but there we ignored the random intercept. Here, we will both use a random intercept as well as regress on the baseline measurements. Thus now we assume

\[ Y_{i,k} = \mu_i + \beta \delta_i + \alpha Y_{i,0} + \epsilon_{i,k} \]  

for \( k = 1, \ldots, m \) and independent \( \epsilon_{i,k} \sim N(0, \sigma^2) \), and finally \( \mu_i \sim N(\mu_0, \tau^2) \) independent of the \( \epsilon \). For the likelihood function we have

\[
L(\tilde{\theta}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi} \tau} \int_{-\infty}^{\infty} e^{-\frac{(y_{i,k} - \mu_i - \beta \delta_i - \alpha Y_{i,0})^2}{2\tau^2}} \prod_{k=1}^{m} \frac{1}{\sqrt{2\pi} \sigma} e^{-\frac{(y_{i,k} - \mu_i - \beta \delta_i - \alpha Y_{i,0})^2}{2\sigma^2}} \, d\mu_i
\]

\[ = (2\pi)^{-\frac{n(m+1)}{2}} \bar{\sigma}^{-nm} \bar{\tau}^{-n} \prod_{i=1}^{n} \int_{-\infty}^{\infty} e^{-\frac{n}{2} \left( \sum_{k=1}^{m} (y_{i,k} - \mu_i - \beta \delta_i - \alpha Y_{i,0})^2 / \sigma^2 + (\mu_i - \mu_0)^2 / \tau^2 \right)} \, d\mu_i,
\]

where \( \tilde{\theta} \) now denotes the vector \((\tilde{\mu}_0, \tilde{\beta}, \tilde{\alpha}, \tilde{\sigma}, \tilde{\tau})\). Using identity (3.81) from the previous section, we obtain the log-likelihood

\[
\ell(\tilde{\theta}) = -\frac{nm}{2} \log(2\pi) - nm \log(\bar{\sigma}) - n \log(\bar{\tau}) - \frac{n}{2} \log \left( \frac{m}{\bar{\sigma}^2} + \frac{1}{\bar{\tau}^2} \right)
+ \frac{\bar{\sigma}^2 \bar{\tau}^2}{2m\bar{\tau}^2 + 2\bar{\sigma}^2} \sum_{i=1}^{n} \left( \frac{1}{\bar{\sigma}^2} \left( \sum_{k=1}^{m} y_{i,k} - m \bar{\delta}_i \bar{\beta} - m \bar{\alpha} Y_{i,0} \right) + \frac{\mu_0}{\bar{\tau}} \right)^2
- \frac{1}{2\bar{\sigma}^2} \sum_{i=1}^{n} \sum_{k=1}^{m} (y_{i,k} - \bar{\beta} \bar{\delta}_i - \bar{\alpha} Y_{i,0})^2 - \frac{n\bar{\mu}_0^2}{2\bar{\tau}^2}.
\]

Taking expectations yields

\[
E(\ell(\tilde{\theta})) = -\frac{nm}{2} \log(2\pi) - nm \log(\bar{\sigma}) - n \log(\bar{\tau}) - \frac{n}{2} \log \left( \frac{m}{\bar{\sigma}^2} + \frac{1}{\bar{\tau}^2} \right)
+ \frac{\bar{\sigma}^2 \bar{\tau}^2}{2m\bar{\tau}^2 + 2\bar{\sigma}^2} \left[ \frac{nm^2 (1 - \bar{\alpha})^2 (\bar{\mu}_0^2 + \tau^2)}{\bar{\sigma}^4} + \frac{2n_{trt} m^2 (1 - \bar{\alpha}) \mu_0 (\beta - \bar{\beta})}{\bar{\sigma}^4} \right.
+ \frac{2nm(1 - \bar{\alpha}) \bar{\mu}_0}{\bar{\sigma}^2 \bar{\tau}^2} \left[ \frac{n_{trt} m^2 (1 + \bar{\alpha}^2)}{\bar{\sigma}^4} + \frac{n_{trt} m \bar{\mu}_0 (\beta - \bar{\beta})}{\bar{\sigma}^2 \bar{\tau}^2} + \frac{n\bar{\mu}_0^2}{\bar{\tau}^4} \right]
- \frac{1}{2\bar{\sigma}^2} \left[ \frac{nm(1 - \bar{\alpha})^2 (\bar{\mu}_0^2 + \tau^2)}{\bar{\sigma}^4} + \frac{2nm m \bar{\mu}_0 (1 - \bar{\alpha}) (\beta - \bar{\beta}) + 2n_{trt} m \bar{\mu}_0 (\beta - \bar{\beta})^2 + nm(1 + \bar{\alpha}^2) \sigma^2}{\bar{\sigma}^2 \bar{\tau}^2} \right]
- \frac{n\bar{\mu}_0^2}{2\bar{\tau}^2}.
\]

Under this misspecified model, it is not obvious what parameter the (quasi-)maximum likelihood estimator converges to. However, we claim the following:
Claim 3.1. The parameter \( \theta_* \) that maximises (3.90) is equal to

\[
\theta_* = (1 - \rho) \mu_0, \beta, \rho, \sigma, \sqrt{1 - \rho^2}
\]  

(3.91)

Proof. Through a somewhat long-winded calculation that we will not write out, it can be seen that the gradient of (3.90) is equal to 0 in \( \theta_* \). A useful strategy for the derivatives with respect to \( \sigma \) and \( \tau \) is to derive all terms separately and then add all coefficients for the additive terms containing \( \mu_0^2 \). One should see that the result is zero, which simplifies the addition of all the other terms by a lot. \( \Box \)

Using this knowledge, one way to compute the asymptotic variance for the maximum likelihood estimate of \( \beta \) is by using the methods described in Section 2.2. In particular, we would need the matrix \( C(\theta_*) \). Fortunately, we can compute \( A(\theta_*) \) by the numerical methods described in the previous section. However, we still need \( B(\theta_*) \). As before, we note that trying to compute the first derivatives is a tedious exercise (especially with respect to \( \tilde{\sigma} \) and \( \tilde{\tau} \)), and computing expectations of products of these functions is even worse. Therefore, it is easier to compute \( B_n(\theta_*) \) for a large \( n \) by simulating a data set of size \( n \) and computing products of first derivatives for each individual by a first order approximation. Then, this can be used to approximate \( B(\theta_*) \) and thus \( C(\theta_*) \).

The downside of this approach using simulation is that computing \( B_n(\theta_*) \) is no trivial task in terms of computation time. Where before we could easily and very quickly (in a matter of minutes if not less) compute the inverse of the Fisher information matrix of a thousand different correlations, simulating a data set of size \( n = 10^4 \) or larger for each of these correlations and computing derivatives takes significantly longer. Another downside is that since \( B_n(\theta_*) \) will depend on the generated data, and in particular will fluctuate somewhat between different simulations. Luckily, there is something that may help us after all.

It turns out that the methods using matrices \( A \) and \( B \) are not necessary in this case. In particular, the estimates for \( \mu_0, \beta \) and \( \alpha \) can be seen to be exactly the same as when using the naive ANCOVA approach. We will now show the reason for this. It can be most easily seen using the multivariate normal notation for the likelihood. First, note that when using the naive linear model in the ANCOVA approach, one patient contributes a term

\[
- \frac{1}{2\sigma^2} \|Y_i - c_i(\mu_0, \beta, \alpha)\|^2 - \frac{m}{2} \log(2\pi) - \frac{1}{2} \log(m\sigma^2)
\]  

(3.92)

to the total log-likelihood, where \( Y_i \) is the vector of all follow-up measurements for the patient, and \( c_i(\mu_0, \beta, \alpha) = \mu_0 + \beta \delta_i + \alpha Y_{i,0} \) is the mean of all follow-ups for that patient. Thus, maximising the log-likelihood with respect to \( \mu_0, \beta \) or \( \alpha \) is equivalent to minimising the sum of all

\[
\|Y_i - c_i(\mu_0, \beta, \alpha)\|^2 = (Y_i - c_i(\mu_0, \beta, \alpha)\|^T (Y_i - c_i(\mu_0, \beta, \alpha)) = Y_i^T Y_i - c_i(\mu_0, \beta, \alpha)^T Y_i - c_i(\mu_0, \beta, \alpha)^T 1^T Y_i + m c_i(\mu_0, \beta, \alpha)^T (3.93)
\]

with respect to \( \mu_0, \beta \) or \( \alpha \), from which the first term can be ignored since it is independent of these parameters. On the other hand, for the random intercept ANCOVA a single patient contributes

\[
- \frac{1}{2\sigma^2} (Y_i - c_i(\mu_0, \beta, \alpha))\|^T \Sigma^{-1} (Y_i - c_i(\mu_0, \beta, \alpha)) - \frac{m}{2} \log(2\pi) - \frac{1}{2} \log(\|\Sigma\|)
\]  

(3.94)

to the total log-likelihood, where \( \Sigma = \tilde{\sigma}^2 I_m + \tilde{\tau}^2 11^T \) is the covariance matrix for a single patient, with terms \( \sigma^2 \) and \( \tau^2 \) on the diagonal and \( \tau^2 \) everywhere outside the diagonal. We used \( I_m \) for the identity matrix of size \( m \) to not confuse it with the Fisher information matrix. It can be shown that

\[
\Sigma^{-1} = \frac{1}{\tilde{\sigma}^2} \left(I_m - \frac{\tilde{\tau}^2}{\tilde{\sigma}^2 + m\tilde{\tau}^2} 11^T \right),
\]  

(3.95)
so that $\Sigma^{-1} \mathbf{1} = 1/(\tilde{\sigma}^2 + m\tilde{\tau}^2) := \lambda \mathbf{1}$. Note that $\Sigma^{-1}$ being symmetric implies that also $\Sigma^{-1} \mathbf{1} = \lambda \mathbf{1}^T$. Using this, we can see that maximising the log-likelihood with respect to $\tilde{\mu}_0, \tilde{\beta}$ or $\tilde{\alpha}$ is equivalent to minimising the sum of all

$$(Y_i - c_i(\tilde{\mu}_0, \tilde{\beta}, \tilde{\alpha})\mathbf{1})^T \Sigma^{-1}(Y_i - c_i(\tilde{\mu}_0, \tilde{\beta}, \tilde{\alpha})\mathbf{1})$$

$$= Y_i^T \Sigma^{-1} Y_i - c_i(\tilde{\mu}_0, \tilde{\beta}, \tilde{\alpha})Y_i^T \Sigma^{-1} \mathbf{1} - c_i(\tilde{\mu}_0, \tilde{\beta}, \tilde{\alpha})\mathbf{1}^T \Sigma^{-1} Y_i + c_i(\tilde{\mu}_0, \tilde{\beta}, \tilde{\alpha})\mathbf{1}^T \Sigma^{-1} \mathbf{1} + m c_i(\tilde{\mu}_0, \tilde{\beta}, \tilde{\alpha})^2$$

from which we can again ignore the first term. Since $\lambda$ is independent of $\tilde{\mu}_0, \tilde{\beta}$ and $\tilde{\alpha}$, we see that maximising this likelihood is equivalent to maximising the naive likelihood. Thus we indeed see that both approaches will yield the same estimator and will in particular have the same variance properties.

The estimated variance, however, will still be different. In the random intercept case the variance of $\beta$ will be estimated through the inverse Fisher information matrix. In general, under misspecification it will be biased. However, it could be the case that the corresponding diagonal elements of $C$ and $\Gamma^{-1}$ (see Section 2.2) are equal or very close, in which case the misspecification does not matter when estimating this parameter. If this is the case, then the estimated variance of $\beta$ is correct.

### 3.3 Discussion

The results of the calculations in the previous sections are summarised in Table 3.2 (as far as closed expressions were obtained) and visualised for a few values of $m$, for $q = 1/2$ in Figure 3.2. The curves for the random intercept model were obtained using points $(\rho, n \text{Var}(\beta)/(\sigma^2 + \tau^2))$ for different $\rho$ in $[0, 1]$ and fitting a polynomial through the points. The resulting polynomials (that fit the points very well, with residuals in the order of $10^{-4}$) are plotted in Figure 3.2. Since both ANCOVA methods are equivalent, only one curve is shown that represents both approaches.

This curve is also exactly the same as the Fisher information curve we would obtain from ignoring misspecification. This means that in the case of ANCOVA, using random effects does not affect the estimate of the treatment effect, but it does affect the estimated variance of the treatment effect. The random intercept ANCOVA will give an estimated variance through the (naive) Fisher information matrix, which turns out to be correct despite the misspecification. On the other hand, we have seen that the ANCOVA without random intercept will provide a biased estimate of this variance.

We see that, as expected by the Cramér-Rao lower bound, the random intercept model always provides the most stable estimate of the treatment effect, at least for large $n$. Between the other three estimators, the interval $[0, 1]$ can be divided in two or three sub-intervals where a different estimator performs the best. For large correlations, the naive “outcome” estimator performs badly compared to the other estimators, which are very close in performance. Similarly, for small correlations the change score estimator will perform badly, while the other estimators all have a similar performance. Finally, we see that the ANCOVA (both the naive and the random intercept version) never really performs badly (for $m > 1$), and thus there is no real harm done when choosing this approach if there is more than one follow-up measurement.

<table>
<thead>
<tr>
<th>Method</th>
<th>Variance</th>
<th>Estimated variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Changes</td>
<td>$(1-\rho)(m+1)(\sigma^2 + \tau^2)$</td>
<td>$(1-\rho)(m+1)(\sigma^2 + \tau^2)$</td>
</tr>
<tr>
<td></td>
<td>$\frac{qnm}{qnm(1-q)m+1} \left[ 1 + \frac{(1-q)m^2-1}{(1-q)m+1} \right]$</td>
<td>$\frac{qnm}{qnm(1-q)m+1} \left[ 1 + \frac{(1-q)m^2-1}{(1-q)m+1} \right]$</td>
</tr>
<tr>
<td>Outcomes</td>
<td>$\frac{\sigma^2 + \tau^2}{nmq(1-q)} \left[ 1 + (m-1)\rho - m\rho^2 \right]$</td>
<td>$\frac{\sigma^2 + \tau^2}{nmq(1-q)} \left[ 1 + (m-1)\rho - m\rho^2 \right]$</td>
</tr>
<tr>
<td>ANCOVA</td>
<td></td>
<td>$\frac{(m+1)(\sigma^2 + \tau^2)}{nmq(1-q)m+1}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\frac{(1-\rho^2)(\sigma^2 + \tau^2)}{nmq(1-q)}$</td>
</tr>
</tbody>
</table>
Figure 3.2: Variance of the estimated treatment effect as a function of $\rho$ for $q = 1/2$
Chapter 4

Model 2: random intercepts and treatment effect as slope

In this chapter we will describe and analyse a model that contains the previous model as a special case. We will start by describing the model, both conceptually and by giving the precise expressions of the distributions. Then, we will compute the maximum likelihood estimator and its variance for three different approaches using linear models. Afterwards, we will use a numerical method to compute the asymptotic variance of the maximum likelihood estimator for two random intercept approaches.

4.1 The model

This model is very similar to the first, except that we will assume the treatment to have a lasting effect on the behaviour of the outcome variable of interest throughout time. More specifically, untreated subjects will see on average no development through time, while treated subjects will show a constant increase or decrease in their outcome variable throughout time. The model will be as follows:

\[
\begin{align*}
\mu_i &\sim N(\mu_0, \tau^2) \\
Y_{i,0} | \mu_i &\sim N(\mu_i, \sigma^2) \\
Y_{i,k} | \mu_i &\sim N(\mu_i + \beta \delta_i t_{i,k}, \sigma^2) \quad \text{for } k = 1, ..., m.
\end{align*}
\]

Here we have again assumed a fixed number of follow-up measurement \(m\) per individual, and we will further assume that the study is designed such that every individual is measured at the same time points, so that \(t_{1,k} = t_{2,k} = \ldots = t_{n,k} := t_k\). Note that the model described in Chapter 3 is equivalent to this model if we take all follow-up times equal to 1. Thus, this chapter might be seen as a generalisation of that chapter. The introduction of the linear effect through time somewhat complicates the computations compared to the previous model. The model is illustrated in Figure 4.1. The following properties still hold for this model:

\[
\begin{align*}
\text{Var}(Y_{i,k}) &= \sigma^2 + \tau^2 \quad \text{for } k = 0, ..., m, \quad (4.2) \\
\text{Cov}(Y_{i,k}, Y_{i,l}) &= \tau^2 \quad \text{for } k, l = 0, ..., m \text{ and } k \neq l, \quad (4.3) \\
\text{Cov}(Y_{i,k}, Y_{i,l}) &= \frac{\tau^2}{\sigma^2 + \tau^2} = \rho \quad \text{for } k, l = 0, ..., m \text{ and } k \neq l. \quad (4.4)
\end{align*}
\]
4.2 Estimating the treatment effect

Here, we start by giving some useful expressions for averages of outcomes, which we will then use to compute the (variance of) three different estimators of the treatment effect. After obtaining these analytic expressions, we will use a numerical approach to compute the variance of the maximum likelihood estimator when using the random intercept model.

4.2.1 Another note on averages

As in Chapter 3, it will sometimes be convenient to express some averages of random variables as a sum of their mean and a standard normal random variable multiplied by the standard deviation. Since the baseline outcomes follow the same model, their expressions will be the same as in that chapter. For completeness, we will list them again here. We will not write out full computations, since they are very similar to the ones in Chapter 3. We have the following identities:

\[ Y_{\text{trt}} = \mu_0 + \sqrt{\tau^2 + \frac{\sigma^2}{m}} Z_{\text{trt}}, \]  

\[ Y_{-\text{trt}} = \mu_0 + \sqrt{\tau^2 + \frac{\sigma^2}{m - n_{\text{trt}}}} Z_{-\text{trt}}, \]  

\[ Y_{1,\ldots,m} = \mu_0 + \beta t + \sqrt{\frac{\sigma^2}{m} + \tau^2} Z_{1,\ldots,m}, \]  

\[ Y_{-1,\ldots,m} = \mu_0 + \sqrt{\frac{\sigma^2}{m} + \tau^2} \sqrt{n_{\text{trt}}} Z_{-1,\ldots,m}. \]  

where \( t := (t_1 + \ldots + t_m)/m \) is the average of the follow-up times. As before, the \( Z \)-variables have mean zero and variance one. Further, the \( Z \)-variables for the treatment group are independent from those for the control group, while those for the same group are not:

\[ \text{Cov}(Z_{0,\ldots,m}^{\text{trt}}, Z_{1,\ldots,m}^{\text{trt}}) = \text{Cov}(Z_{0,\ldots,m}^{\text{trt}}, Z_{1,\ldots,m}^{\text{trt}}) = \frac{\tau^2}{\sqrt{\frac{\sigma^2}{m} + \tau^2} \sqrt{\sigma^2 + \tau^2}} \]  

A new average we will come across that we have not seen before is the average of outcomes multiplied by time. We have

\[ tY_{1,\ldots,m}^{\text{trt}} = \mu_0 t + \beta t^2 + \sqrt{\frac{\sigma^2 t^2}{m} + \tau^2 t^2} Z_{1,\ldots,m}^{\text{trt}}. \]
Again, the variable $Z_{trt}$ is independent from the other $Z$-variables dealing with the control group, but we can compute

$$\text{Cov}(Z_{0 \text{ trt}}, Z_{t \text{ trt}}) = \frac{\tau^2}{\sqrt{\sigma^2 + \tau^2 \sqrt{(t^2/f)^2 \sigma^2/m + \tau^2}}} := \gamma_1,$$  

(4.11)

$$\text{Cov}(Z_{1, \ldots, m \text{ trt}}, Z_{t \text{ trt}}) = \frac{\sqrt{\sigma^2/m + \tau^2}}{\sqrt{(t^2/f)^2 \sigma^2/m + \tau^2}} := \gamma_2.$$  

(4.12)

### 4.2.2 Change scores

As before, we can define the change scores $\Delta_{i,k} := Y_{i,k} - Y_{i,0} = t_k \beta \delta_i + \epsilon_{i,k} - \epsilon_{i,0}$, which are normally distributed with mean $t_k \beta \delta_i$ and

$$\text{Var}(\Delta_{i,k}) = 2\sigma^2$$  

for $k = 1, \ldots, m$,  

(4.13)

$$\text{Cov}(\Delta_{i,k}, \Delta_{i,l}) = \sigma^2$$  

for $k, l = 1, \ldots, m$ and $k \neq l$.  

(4.14)

We can again choose whether or not we include an intercept. First, we will look into the more natural case of not including one.

#### Without intercept

In this case we are interested in the coefficient for the interaction of treatment and time, so that we obtain $X = (t_1, \ldots, t_m, t_1, \ldots, t_m, \ldots, 0, \ldots, 0)^T$, where $(t_1, \ldots, t_m)$ is repeated $n_{\text{trt}}$ times and followed by $(n - n_{\text{trt}})m$ zeroes.

Before computing an estimator, however, we would again like to transform the data again so that the transformed data are uncorrelated. The same method as in Chapter 3 can be used, since the correlation structure of the errors is the same. Recall that we defined

$$T_0 = I + \frac{1}{m} \left( \frac{1}{\sqrt{m+1}} - 1 \right) 11^T,$$  

(4.15)

and $T$ as the block-diagonal matrix with $n \times m$-sized blocks $T_0$. Noting that we want to model $E(\Delta_{i,k}) = \beta \delta_i t_k$, we can obtain the linear model

$$T \Delta = TX \beta + \nu,$$  

(4.16)

with $\nu \sim N(0, \sigma^2 I)$. We have

$$T_0 X_i = X_i + \delta_i T \left( \frac{1}{\sqrt{m+1}} - 1 \right) 1,$$  

(4.17)

from which we can compute

$$(TX)^T TX = n_{\text{trt}} m \left( t^2 - \frac{m}{m+1} \right).$$  

(4.18)

The estimator for the treatment effect will then be normally distributed with

$$E(\hat{\beta}) = E(\{(TX)^T TX\}^{-1}(TX)^T \Delta)$$

$$= \{(TX)^T TX\}^{-1}(TX)^T E(\Delta)$$

$$= \frac{1}{n_{\text{trt}} m (t^2 - m t^2 / (m + 1))} \sum_{i=1}^{n_{\text{trt}}} \sum_{k=1}^{m} \left( t_k - \frac{m}{m+1} \right) E(\Delta_{i,k})$$

$$= \frac{1}{n_{\text{trt}} m (t^2 - m t^2 / (m + 1))} \sum_{i=1}^{n_{\text{trt}}} \sum_{k=1}^{m} \left( t_k - \frac{m}{m+1} \right) \beta_t k$$

$$= \beta$$  

(4.19)
and

\[
\text{Var}(\hat{\beta}) = \sigma^2((TX)^T TX)^{-1} = \frac{\sigma^2}{n_{\text{trt}} m (m + 1)\bar{t}^2 - m t^2} = \frac{(1 - \rho)(m + 1)(\sigma^2 + \tau^2)}{n_{\text{trt}} (m + 1)\bar{t}^2 - m t^2}.
\]

(4.20)

We write \(\bar{t}^2\) for the average of the squared follow-up times. Note that if \(t_m \gg t_m - 1\), then the last fraction in the above expression is approximately equal to \(1/t_m^2\), meaning that it is favourable to have the final follow-up measurement at a relatively large time point. This is also intuitively clear from the fact that we are estimating an intercept that is assumed to stay constant throughout time. With the variance not increasing over time, measuring a slope over a longer time period is more accurate than measuring it in a small window. Since the system satisfies all assumptions of the linear model, the estimated variance that is given by statistical software will be a consistent estimate of the true variance.

Now we will look into the consequences of including an intercept in the model. Recall that in Chapter 3, we saw that (for \(n_{\text{trt}} = n/2\)), including an intercept doubled the variance of the estimator for the treatment effect. In general, the variance was increased by a factor \(1/(1 - q)\) where \(q = n_{\text{trt}}/n\). We might expect a similar outcome here.

**WITH INTERCEPT**

Adding an intercept, we obtain the linear model

\[
T \Delta = TX \theta + \eta,
\]

(4.21)

for parameter vector \(\theta = (1, \beta)^T\), residuals \(\eta\) that satisfy the assumptions of the linear model and

\[
X = \begin{pmatrix}
1 & t_1 \\
\vdots & \vdots \\
1 & t_m \\
\vdots & \vdots \\
1 & 0 \\
1 & 0 \\
\vdots & \vdots \\
1 & 0
\end{pmatrix},
\]

(4.22)

Using previous observations we can compute

\[
(TX)^T TX = \frac{qnm}{m + 1} \begin{pmatrix}
1/q & 1 \\
1 & (m + 1)\bar{t}^2 - m t^2
\end{pmatrix}.
\]

(4.23)

and thus

\[
((TX)^T TX)^{-1} = \frac{m + 1}{nm} \frac{1}{(m + 1)\bar{t}^2 - (m + q) t^2} \begin{pmatrix}
(m + 1)\bar{t}^2 - m t^2 & -\bar{t} \\
-\bar{t} & 1/q
\end{pmatrix}.
\]

(4.24)

Finally, we can find

\[
(TX)^T T = \frac{1}{m + 1} \begin{pmatrix}
1^T \\
(1 + 1)\bar{t}^T - m t^T & \cdots & \cdots & 1^T \\
(1 + 1)\bar{t}^T - m t^T & \cdots & 1^T & 1^T \cdots 1^T
\end{pmatrix}.
\]

(4.25)
where we used vectors $\mathbf{t} = (t_1, ..., t_m)^T$ and $\mathbf{0} = (0, ..., 0)^T$ of length $m$ for convenience. Using what we have learned here we can find that $\hat{\beta} = [(TX)^T TX]^{-1}(TX)^T \Delta]_2$ is normally distributed with

$$E(\hat{\beta}) = \beta$$  \hspace{1cm} (4.26)

and

$$\text{Var}(\hat{\beta}) = \frac{\sigma^2(m + 1)}{qnm} \frac{1}{(m + 1)t^2 - (m + q)t^2}$$

$$= \frac{(1 - \rho)(m + 1)(\sigma^2 + \tau^2)}{qnm} \frac{1}{(m + 1)t^2 - (m + q)t^2}. \hspace{1cm} (4.27)$$

The difference between this and (4.20) is that the term $mt^2$ changed into $(m + q)t^2$, increasing the variance of the estimator by a factor

$$\frac{(m + 1)t^2 - mt^2}{(m + 1)t^2 - (m + q)t^2}. \hspace{1cm} (4.28)$$

Plugging in $t_1 = ... = t_m = 1$ we obtain the same factor $1/(1 - q)$ we saw in Chapter 3.

Since the variance is correctly specified, the estimator $\hat{\sigma}$ for $\sigma$ used by statistical software will be consistent. This means that the estimated variance of $\hat{\beta}$ will also be consistent for the estimator of the treatment effect.

### 4.2.3 Using outcomes and ignoring random intercept

Just like when we used this approach in Chapter 3, we can be naive and ignore the presence of a random intercept and correlated outcomes. Thus we assume

$$Y_{i,k} = \mu + \beta t_k \delta_i 1_{(k>0)} + \epsilon_{i,k} \hspace{1cm} (4.29)$$

for $k = 0, ..., m$ and uncorrelated errors $\epsilon_{i,k} \sim N(0, \sigma^2)$. Then we have the linear model

$$Y = X(1, \beta)^T + \epsilon \hspace{1cm} (4.30)$$

with

$$X = \begin{bmatrix}
1 & 0 \\
1 & t_1 \\
1 & \vdots \\
1 & t_m \\
1 & 0 \\
1 & 0 \\
1 & \vdots \\
1 & 0
\end{bmatrix} \begin{bmatrix}
\vdots
\vdots
\vdots
\vdots
\vdots
\vdots
\vdots
\vdots
\end{bmatrix}$$

$n_{trt}(m + 1)$ rows. \hspace{1cm} (4.31)
which leads to

\[ X^T X = \begin{pmatrix} n(m+1) & n_{trt} \sum_{k=1}^m t_k \\ n_{trt} \sum_{k=1}^m t_k & n_{trt} \sum_{k=1}^m t_k^2 \end{pmatrix}, \tag{4.32} \]

\[ (X^T X)^{-1} = \frac{1}{mn_{trt}(m+1)\sum_{k=1}^m t_k^2 - n_{trt}^2(\sum_{k=1}^m t_k)^2} \begin{pmatrix} n_{trt} \sum_{k=1}^m t_k^2 & -n_{trt} \sum_{k=1}^m t_k \\ -n_{trt} \sum_{k=1}^m t_k & n(m+1) \end{pmatrix}, \tag{4.33} \]

\[ X^T Y = \begin{pmatrix} \sum_{i=1}^n \sum_{k=0}^m Y_{i,k} \\ \sum_{i=1}^{n_{trt}} \sum_{k=1}^m t_k Y_{i,k} \end{pmatrix}. \tag{4.34} \]

A straightforward calculation shows that

\[ \hat{\beta} = [(X^T X)^{-1} X^T Y]_2 \sim N \left( \beta, \sigma^2 + \tau^2 \frac{n_{trt}}{nmU} \left[ 1 + \frac{1 - \frac{qm}{mn}}{m^2U} \right] \right), \tag{4.35} \]

where

\[ U = \frac{\tau^2}{m+1} - \frac{qm}{m+1} \tau^2. \tag{4.36} \]

We will not compare this expression to the previously derived variance of the change score estimator, due to the relatively complex dependence on the choice of follow-up time points. Instead, we will refer to the discussion at the end of this chapter for a visual overview.

It can also be shown that the estimator for \( \mu \) has expected value \( \mu_0 \). This means that the residuals are approximately equal to \( \mu_i - \mu_0 + \epsilon_{i,k} \), which in turn implies that \( \hat{\sigma}^2 \approx \sigma^2 + \tau^2 \).

The estimated variance \( \hat{\text{Var}}(\hat{\beta}) \) is then approximately equal to

\[ \hat{\sigma}^2 [(X^T X)^{-1}]_{2,2} \approx \sigma^2 + \tau^2 \frac{n_{trt}}{mnU}, \tag{4.37} \]

which means the variance given by statistical software when using this method will be biased.

### 4.2.4 ANCOVA

In Chapter \[3\] we used an intercept, treatment indicator and baseline measurement to predict the outcomes. Here, instead of a treatment indicator, we use a treatment-time interaction. Thus we will assume

\[ Y_{i,k} = \mu + \beta t_k \delta_i + \alpha Y_{i,0} + \epsilon_{i,k} \tag{4.38} \]

for \( k = 1, \ldots, m \) and uncorrelated errors \( \epsilon_{i,k} \sim N(0, \sigma^2) \). Written shortly, this is the linear model

\[ Y = X(1, \beta, \alpha)^T + \epsilon, \tag{4.39} \]
where $Y$ now is the vector of all follow-up measurements (instead of all measurements), and

$$
X = \begin{pmatrix}
1 & t_1 & Y_{1,0} \\
\vdots & \vdots & \vdots \\
1 & t_m & Y_{1,0} \\
1 & t_{n_{trt},0} & \vdots \\
1 & t_m & Y_{n_{trt},0} \\
1 & 0 & Y_{n_{trt}+1,0} \\
\vdots & \vdots & \vdots \\
1 & 0 & Y_{n,0}
\end{pmatrix},
$$

(4.40)

This gives us

$$
X^T X = nm \begin{pmatrix}
1 & q\overline{t} & \overline{Y}_0 \\
q\overline{t} & q\overline{t}^2 & q\mu_0 \\
\mu_0 & q\mu_0 & \sigma^2 + \tau^2 + \mu_0^2 \\
\overline{B}_0 & \overline{D}_0
\end{pmatrix} + \epsilon \begin{pmatrix}
0 & 0 & \sqrt{qZ_{0_{trt}}^{trt} + \sqrt{1-\bar{q}Z_{0_{trt}}^{-trt}}} \\
0 & 0 & \sqrt{qZ_{0_{trt}}^{trt}} \sqrt{qZ_{0_{trt}}^{trt}} \\
\sqrt{1-\bar{q}Z_{0_{trt}}^{-trt}} & \sqrt{qZ_{0_{trt}}^{trt}} & 2\sigma^2 + 2\mu_0^2 \\
\epsilon & \epsilon & 0
\end{pmatrix} \begin{pmatrix}
qY_{1_{trt}}^{trt} + (1-q)\overline{Y}_{1_{trt}}^{-trt} \\
qY_{1_{trt}}^{trt} \\
\overline{Y}_0 Y_{1_{trt}}
\end{pmatrix},
$$

(4.41)

where $q = n_{trt}/n$ and $\overline{t}$ and $\overline{t}^2$ are again the average of the regular and squared follow-up times, respectively. By the results of Section 3.2.1, we can write the matrix in the equation above as

$$
\begin{pmatrix}
1 & q\overline{t} & \mu_0 \\
q\overline{t} & q\overline{t}^2 & q\mu_0 \\
\mu_0 & q\mu_0 & \sigma^2 + \tau^2 + \mu_0^2 \\
\overline{B}_0 & \overline{D}_0
\end{pmatrix} + \epsilon \begin{pmatrix}
0 & 0 & \sqrt{qZ_{0_{trt}}^{trt} + \sqrt{1-\bar{q}Z_{0_{trt}}^{-trt}}} \\
0 & 0 & \sqrt{qZ_{0_{trt}}^{trt}} \sqrt{qZ_{0_{trt}}^{trt}} \\
\sqrt{1-\bar{q}Z_{0_{trt}}^{-trt}} & \sqrt{qZ_{0_{trt}}^{trt}} & 2\sigma^2 + 2\mu_0^2 \\
\epsilon & \epsilon & 0
\end{pmatrix} \begin{pmatrix}
qY_{1_{trt}}^{trt} + (1-q)\overline{Y}_{1_{trt}}^{-trt} \\
qY_{1_{trt}}^{trt} \\
\overline{Y}_0 Y_{1_{trt}}
\end{pmatrix},
$$

(4.42)

We can write this as

$$
X^T Y = nm \begin{pmatrix}
\mu_0 + q\overline{t}\beta \\
q\mu_0 + q\overline{t}\beta \\
\mu_0^2 + \tau^2 + q\overline{t}\mu_0 \\
D_0
\end{pmatrix} + \frac{1}{\sqrt{n}} \begin{pmatrix}
\sqrt{\sigma^2/m + \tau^2} \left(\sqrt{qZ_{1_{trt}}^{trt} + \sqrt{1-\bar{q}Z_{1_{trt}}^{-trt}}} \right) \\
\sqrt{qZ_{0_{trt}}^{trt}} \sqrt{qZ_{0_{trt}}^{trt}} \\
\sqrt{\sigma^2/m + \tau^2} Z_{0_{trt}}^{trt} \\
CZ_x
\end{pmatrix} \begin{pmatrix}
qY_{1_{trt}}^{trt} + (1-q)\overline{Y}_{1_{trt}}^{-trt} \\
qY_{1_{trt}}^{trt} \\
\overline{Y}_0 Y_{1_{trt}}
\end{pmatrix}
$$

(4.43)

As we have done multiple times already, we can compute the maximum likelihood estimator by

$$
\hat{\beta} = [(X^T X)^{-1} X^T Y]_2 = [(B_0 + \epsilon A)^{-1}(D_0 + D_1/\sqrt{n})]_2.
$$

(4.44)
For small $\epsilon$, we have \((B_0 + \epsilon A)^{-1} \approx B_0^{-1} - \epsilon B_0^{-1} A B_0^{-1}\). If we ignore terms of order smaller than \(n^{-1/2}\), we obtain

\[
\hat{\beta} = [B_0^{-1} D_0]_2 = \epsilon[B_0^{-1} A B_0^{-1} D_0]_2 + [B_0^{-1} D_1]_2 / \sqrt{n}
\]

as \(n \to \infty\). Note that as before, the vector \(B_0^{-1} D_0\) is the first order term of the estimator of all three parameters. We have

\[
B_0^{-1} = \begin{pmatrix}
\frac{\mu_0^2}{\sigma^2 + \tau^2} + \frac{\tau}{\sigma^2 - qt^2} - \frac{\mu_0}{\sigma^2 + \tau^2} & -\frac{\tau}{\sigma^2 - qt^2} \\
-\frac{\tau}{\sigma^2 - qt^2} & \frac{1}{\sigma^2 - qt^2} \\
-\frac{\mu_0}{\sigma^2 + \tau^2} & 0 & \frac{1}{\sigma^2 + \tau^2}
\end{pmatrix}
\]

which leads to

\[
B_0^{-1} D_0 = \begin{pmatrix}
\mu_0 (1 - \rho) \\
\beta \\
\rho
\end{pmatrix}
\]

which is (maybe not so) miraculously exactly the same as found in Chapter 3. Through a straightforward calculation we find

\[
[B_0^{-1} D_1]_2 = -\frac{\sqrt{\sigma^2/m + \tau^2}}{t^2 - qt^2}(\sqrt{q} Z_{1,\ldots,m}^{\text{trt}} + \sqrt{1 - q} Z_{1,\ldots,m}^{-\text{trt}}) + \frac{1}{\sqrt{q}} \frac{\sqrt{\sigma^2/m + \tau^2}}{t^2 - qt^2} Z_t^{\text{trt}}.
\]

Employing the same strategy that we have already used in Chapter 3, we further find

\[
[B_0^{-1} A B_0^{-1} D_0]_2 = \rho \frac{\sqrt{1 - q}}{t^2 - qt^2} \left[ \sqrt{\frac{1 - q}{q}} Z_0^{\text{trt}} - Z_0^{-\text{trt}} \right].
\]

Combining these results with the distribution of the vector of \(Z\)-variables,

\[
\begin{pmatrix}
Z_{1,\ldots,m}^{\text{trt}} \\
Z_{1,\ldots,m}^{-\text{trt}} \\
Z_0^{\text{trt}} \\
Z_t^{\text{trt}} \\
Z_0^{-\text{trt}}
\end{pmatrix} \sim \mathcal{N}(0, \begin{pmatrix}
1 & 0 & \gamma & \gamma_2 \\
0 & 1 & \gamma & 0 \\
\gamma & 0 & 1 & \gamma_1 \\
0 & \gamma & 0 & 1 \\
\gamma_2 & 0 & \gamma_1 & 0
\end{pmatrix}),
\]

we find

\[
\hat{\beta} \sim \mathcal{N} \left( \beta, \frac{\tau^2(\sigma^2 + \tau^2)}{amq(t^2 - qt^2)^2} \left[ -m\rho^2(1 - q) + \rho \left( m - \frac{\tau^2}{t^2} - q(m - 1) \right) + \frac{\tau^2}{t^2} - q \right] \right).
\]

Using the first order approximations of the parameters, we see that the residuals are approximately equal to \((1 - \rho)(\mu_t - \mu_0) + \epsilon_{i,k} - \rho \epsilon_{i,k}\), so that \(\hat{\sigma}^2 \approx (1 - \rho)^2 \tau^2 + (1 + \rho^2)\sigma^2 = \sigma^2 + (1 - \rho)\tau^2\). This gives us

\[
\text{Var}(\hat{\beta}) \approx \frac{\hat{\sigma}^2 [B_0^{-1}]_2^2}{nm} \approx \frac{(1 - \rho^2)(\sigma^2 + \tau^2)}{qnm} \frac{1}{t^2 - qt^2},
\]

which is a biased estimate of the true variance.
4.2.5 Autoregressive approach

Recall that the autoregressive approach takes the form

\[ Y_{i,k} = \mu + \beta t_k \delta_i + \alpha Y_{i,k-1} + \epsilon_{i,k}, \quad (4.53) \]

for \( k = 1, \ldots, m \). As input matrix we have

\[
X = \begin{pmatrix}
1 & t_1 & Y_{1,0} \\
\vdots & \vdots & \vdots \\
1 & t_m & Y_{1,m-1} \\
\vdots & \vdots & \vdots \\
1 & t_1 & Y_{ntrt,0} \\
\vdots & \vdots & \vdots \\
1 & t_m & Y_{ntrt,m-1} \\
1 & 0 & Y_{ntrt+1,0} \\
\vdots & \vdots & \vdots \\
1 & 0 & Y_{n,m-1}
\end{pmatrix}
\quad (4.54)
\]

which leads to

\[
X^T X \approx nm \begin{pmatrix}
1 & q\overline{t} & q\overline{Y_{0,\ldots,m-1}} (1 - q) \overline{Y_{0,\ldots,m-1}} \\
q\overline{t} & q\overline{t^2} & q\overline{t^2} Y_{k-1} \\
q\overline{Y_{0,\ldots,m-1}} (1 - q) \overline{Y_{0,\ldots,m-1}} & q\overline{t^2} Y_{k-1} & Y_{0,\ldots,m-1}
\end{pmatrix}
\quad (4.55)
\]

where the averages are taken over all measurements (including baseline) apart from the final follow-up measurement. Again we will first only consider the first order approximation of the estimated treatment effect in order to expose the biased nature. Replacing averages with expected values, we obtain

\[
X^T X \approx nm \begin{pmatrix}
1 & q\overline{t} & \frac{m-1}{m} q\overline{t^m} \beta + \mu_0 \\
q\overline{t} & q\overline{t^2} & \frac{m-1}{m} q\overline{t^m} \beta + \mu_0 \\
\frac{m-1}{m} q\overline{t^m} \beta + \mu_0 & \frac{m-1}{m} q\overline{t^m} \beta + \mu_0 & \frac{m-1}{m} q\overline{t^m} \beta + \mu_0 \\
\frac{m-1}{m} q\overline{t^m} \beta + \mu_0 & \frac{m-1}{m} q\overline{t^m} \beta + \mu_0 & \frac{m-1}{m} q\overline{t^m} \beta + \mu_0 \\
\frac{m-1}{m} q\overline{t^m} \beta + \mu_0 & \frac{m-1}{m} q\overline{t^m} \beta + \mu_0 & \frac{m-1}{m} q\overline{t^m} \beta + \mu_0 \\
\frac{m-1}{m} q\overline{t^m} \beta + \mu_0 & \frac{m-1}{m} q\overline{t^m} \beta + \mu_0 & \frac{m-1}{m} q\overline{t^m} \beta + \mu_0 \\
\end{pmatrix}
\quad (4.56)
\]

where \( \overline{t^m} = (t_1 + \ldots + t_{m-1})/(m-1) \), \( \overline{t^2} = (t_1^2 + \ldots + t_{m-1}^2)/(m-1) \) and \( \overline{t_k t_{k-1}} = (t_k t_{k-1})/(m-1) \). Similarly, we can compute

\[
X^T Y \approx nm \begin{pmatrix}
\mu_0 + q\overline{t} \beta \\
q\overline{t^2} \beta + \mu_0 \\
\mu_0 + q\overline{t} \beta \\
\mu_0 + q\overline{t^2} \beta \\
\mu_0 + q\overline{t} \beta \\
\mu_0 + q\overline{t^2} \beta
\end{pmatrix}
\quad (4.57)
\]

Combining these expressions to obtain the approximate value of \( \hat{\beta} = [(X^T X)^{-1} X^T Y]_2 \) is rather time-consuming to do analytically, so we will opt to demonstrate the bias of the estimator by some example. We use \( \sigma = 3, \tau = 6, \mu_0 = 130 \) and \( \beta = -1 \). The results can be found in Table 4.1.

The estimator is clearly biased, and we will again not put more effort into computing the variance properties. Instead, we will now look at some random intercept approaches.
4.2.6 Random intercept model

In this section, we will look at the mixed model approach, using the maximum likelihood method to estimate the parameters using all outcomes. Note that as in the previous chapter, this model is the true assumed model. Repeating here in a slightly different form, the model is

\[ Y_{i,k} = \mu_i + \beta t_k \delta_1 \{ k > 0 \} + \epsilon_{i,k}, \quad (4.58) \]

for \( k = 0, \ldots, m \), where \( \mu_i \sim N(\mu_0, \tau^2) \) is independent of the errors \( \epsilon_{i,k} \sim N(0, \sigma^2) \), which are also assumed to be independent from each other. The likelihood for this model can be written as

\[
L(\hat{\theta}) = \prod_{i=1}^{n} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi} \hat{\sigma}} e^{-\frac{(y_{i,k} - \hat{\mu}_i - \hat{\beta} t_k \delta_1 \{ k > 0 \})^2}{2 \hat{\sigma}^2}} \left( \frac{1}{\sqrt{2\pi} \hat{\tau}} e^{-\frac{(\sum_{k=0}^{m} (y_{i,k} - \mu_0 - \hat{\beta} t_k \delta_1 \{ k > 0 \})^2 + (\mu_i - \hat{\mu}_0)^2 + \tau^2)}{2 \tau^2}} \right) d\mu_i,
\]

where \( \hat{\theta} \) denotes the vector \((\hat{\mu}_0, \hat{\beta}, \hat{\sigma}, \hat{\tau})\). Recall the identity given in Chapter 3:

\[
\int_{-\infty}^{\infty} e^{-\frac{1}{2} \sum_j a_j (\mu - b_j)^2} d\mu = \sqrt{\frac{2\pi}{\sum_j a_j}} e^{-\frac{1}{2} \sum_j a_j b_j^2}. \quad (4.60)
\]

We can use this to compute the log-likelihood

\[
\ell(\hat{\theta}) = -\frac{n(m+1)}{2} \log(2\pi) - n(m+1) \log(\hat{\sigma}) - n \log(\hat{\tau}) - \frac{n}{2} \log \left( \frac{m+1}{\hat{\beta}^2} + \frac{1}{\hat{\tau}^2} \right) + \frac{\hat{\sigma}^2 \hat{\tau}^2}{2(m+1)^2 \hat{\sigma}^2 + 2\hat{\tau}^2} \sum_{k=0}^{m} \left( \frac{1}{\hat{\sigma}^2} \left( \sum_{i=1}^{n} (y_{i,k} - \mu_0 - \hat{\beta} t_k \delta_1 \{ k > 0 \})^2 + \frac{\hat{\mu}_0}{\hat{\tau}^2} \right) \right)^2 - \frac{1}{2\hat{\sigma}^2} \sum_{i=1}^{n} \sum_{k=0}^{m} (y_{i,k} - \hat{\beta} t_k \delta_1 \{ k > 0 \})^2 - \frac{n\hat{\mu}_0^2}{2\hat{\tau}^2}. \quad (4.61)
\]

Taking expectations yields

\[
E(\ell(\hat{\theta})) = -\frac{n(m+1)}{2} \log(2\pi) - n(m+1) \log(\hat{\sigma}) - n \log(\hat{\tau}) - \frac{n}{2} \log \left( \frac{m+1}{\hat{\beta}^2} + \frac{1}{\hat{\tau}^2} \right) + \frac{\hat{\sigma}^2 \hat{\tau}^2}{2(m+1)^2 \hat{\sigma}^2 + 2\hat{\tau}^2} \left[ \frac{n(m+1)^2 (\hat{\mu}_0^2 + \tau^2)}{\hat{\sigma}^4} + \frac{2n_{trt} m (m+1) \hat{\mu}_0 (\beta - \hat{\beta})}{\hat{\sigma}^4} + \frac{2n_{trt} \hat{\mu}_0 \hat{\beta}^2 (\beta - \hat{\beta})}{\hat{\sigma}^4} + \frac{n(m+1) \sigma^2}{\hat{\sigma}^4} + \frac{2n_{trt} \hat{\mu}_0 \hat{\beta} (\beta - \hat{\beta})}{\hat{\sigma}^4} + \frac{n\hat{\mu}_0^2}{2\hat{\tau}^2} \right] - \frac{1}{2\hat{\sigma}^2} \left[ n(m+1)(\hat{\mu}_0^2 + \tau^2 + \sigma^2) + 2n_{trt} \hat{\mu}_0 \hat{\beta} (\beta - \hat{\beta}) + n_{trt} \hat{\beta}^2 (\beta - \hat{\beta})^2 - \frac{n\hat{\mu}_0^2}{2\hat{\tau}^2} \right]. \quad (4.62)
\]
Now we can use the numerical scheme described in Section 3.2.6 to compute the second derivative and thus the asymptotic variance of $\hat{\beta}$. Again, for fixed time vector $(t_1, \ldots, t_m)$, the curve of the variance of $n \text{Var}(\beta) / (\sigma^2 + \tau^2)$ as a function of $\rho$ does not seem to depend on the choice of $\sigma^2 + \tau^2$, $n$, $\beta$ or $\mu_0$.

4.2.7 ANCOVA with random intercept

For the ANCOVA with random intercept we model

$$Y_{i,k} = \mu_i + \beta_0 t_k + \alpha Y_{i,0} + \epsilon_{i,k}$$

(4.63)

for $k = 1, \ldots, m$ and independent $\epsilon_{i,k} \sim N(0, \sigma^2)$, and finally $\mu_i \sim N(\mu_0, \tau^2)$ independent of the $\epsilon$. For the likelihood function we have

$$L(\tilde{\theta}) = \prod_{i=1}^{n} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{(y_{i,k} - \mu_{i,k})^2}{2\sigma^2}} \prod_{k=1}^{m} \frac{1}{\sqrt{2\pi}} e^{-\frac{(y_{i,k} - \mu_i - \beta_0 t_k - \alpha y_{i,0})^2}{2\tau^2}} \, d\mu_i$$

$$= (2\pi)^{-nm(m+1)/2} \sigma^{-nm} \tau^{-m} \prod_{i=1}^{n} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(\sum_{k=1}^{m} (y_{i,k} - \mu_i - \beta_0 t_k - \alpha y_{i,0})^2 + (\mu_i - \mu_0)^2 / \tau^2)} \, d\mu_i,$n

(4.64)

where $\tilde{\theta}$ now denotes the vector $(\mu_0, \beta_0, \alpha, \widetilde{\sigma}, \widetilde{\tau})$. Again using identity (3.81), we obtain the log-likelihood

$$\ell(\tilde{\theta}) = -\frac{nm}{2} \log(2\pi) - nm \log(\sigma) - n \log(\tau) - \frac{n}{2} \log \left( \frac{m}{\sigma^2} + \frac{1}{\tau^2} \right)$$

$$+ \frac{2m \tau^2 \sigma^2}{2m \tau^2 + 2\sigma^2} \sum_{i=1}^{n} \left( \frac{1}{\sigma^2} \left( \sum_{k=1}^{m} (y_{i,k} - \mu_i - \beta_0 t_k - \alpha y_{i,0}) \right) + \frac{\mu_i}{\tau^2} \right)^2$$

(4.65)

$$- \frac{1}{2\sigma^2} \sum_{i=1}^{n} \sum_{k=1}^{m} (y_{i,k} - \beta_0 t_k - \alpha y_{i,0})^2 - \frac{n\mu_0^2}{2\tau^2}.$$n

Taking expectations yields

$$E(\ell(\tilde{\theta})) = -\frac{nm}{2} \log(2\pi) - nm \log(\sigma) - n \log(\tau) - \frac{n}{2} \log \left( \frac{m}{\sigma^2} + \frac{1}{\tau^2} \right)$$

$$+ \frac{2m \tau^2 \sigma^2}{2m \tau^2 + 2\sigma^2} \left[ \frac{nm^2 (1 - \alpha^2) (\mu_0^2 + \tau^2)}{\sigma^4} + \frac{2nm \tau^2 (1 - \alpha) \mu_0 (\beta - \beta_0)}{\sigma^4} \right.$$n

$$+ \frac{2nm (1 - \alpha) \mu_0 \mu_0}{\sigma^4} + \frac{nm \tau^2 (1 - \alpha) \mu_0 (\beta - \beta_0)}{\sigma^4} + \frac{nm \sigma^2 (1 + m\alpha^2)}{\sigma^4} + \frac{2nm \tau^2 \mu_0 \mu_0 (\beta - \beta_0)}{\sigma^2 \tau^2} \right.$$n

$$+ \frac{nm (1 - \alpha) (\mu_0^2 + \tau^2) + 2nm \tau^2 \mu_0 (1 - \alpha) (\beta - \beta_0) + nm \tau^2 (\beta - \beta_0)^2 + nm (1 + \alpha^2) \sigma^2}{\sigma^2 \tau^2} + \frac{\eta \mu_0^2}{\tau^4} \right] - \frac{1}{2\sigma^2} \left[ nm (1 - \alpha) (\mu_0^2 + \tau^2) + 2nm \tau^2 \mu_0 (1 - \alpha) (\beta - \beta_0) + nm \tau^2 (\beta - \beta_0)^2 \right.$$

(4.66)

$$+ \frac{\eta \mu_0^2}{\tau^4} \left] - \frac{\eta \mu_0^2}{\tau^4} \right].$$

We claim that the maximum of this expression is achieved in the same parameter vector as we saw in Chapter 3 for the ANCOVA with random intercept.

Claim 4.1. The parameter $\theta_*$ that maximises (4.66) is equal to

$$\theta_* = ((1 - \rho)\mu_0, \beta, \rho, \sigma, \sqrt{1 - \rho \tau})$$

(4.67)

41
Proof. This is proved in the same way as Claim 3.1. The derivatives with respect to \( \tilde{\mu}_0 \), \( \tilde{\beta} \) and \( \tilde{\alpha} \) are straightforward to compute, and for the derivatives with respect to \( \tilde{\sigma} \) and \( \tilde{\tau} \) it can be noted that Claim 3.1 implies they are equal to zero in \( \theta^* \). This is the case since the only terms where (3.90) differs from (4.66) also involve a factor \( \beta - \tilde{\beta} \), which means these terms disappear when taking the derivative with respect to \( \tilde{\sigma} \) or \( \tilde{\tau} \) and substituting \( \tilde{\theta} = \theta^* \).

Using this, we can use the numerical methods that we have already described in the random intercept ANCOVA section in Chapter 3. In this case it is not true that both ANCOVA approaches yield the same estimator, this was only true in Chapter 3 because the expected follow-up for a patient was the same at each follow-up time. Since we are dealing with a slope here, that does not hold true anymore. The results will be shown and discussed in the next section.

4.3 Discussion

In this chapter, we obtained more complicated and thus less readable results for the three estimators for which we could obtain analytic expressions. These expressions are summarised in Table 4.2 although admittedly they are not easily interpretable through the dependence on the follow-up times. Due to the length of the expressions and because they are not used in our discussion here, we have omitted the estimated variances. Recall that the change score approach will estimate the variance consistently while the other two provide biased estimators. To make the results more interpretable, we have plotted \( n \text{Var}(\tilde{\beta})/(\sigma^2 + \tau^2) \) as a function of \( \rho \)

<table>
<thead>
<tr>
<th>Method</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Changes</td>
<td>( \frac{(1-\rho)(m+1)(\sigma^2 + \tau^2)}{\sigma^2 + \tau^2} \frac{1}{m} ) ( \frac{n_{\text{trt}} m}{(m+1) \tilde{t}^2 - m \tilde{t}^2} )</td>
</tr>
<tr>
<td>Outcomes</td>
<td>( \frac{\sigma^2 + \tau^2}{n_{\text{trt}} m \bar{U}} \left[ 1 + \left( 1 - \frac{\bar{U}}{m} \right) \frac{\tilde{m}^2 - \tilde{t}^2}{m} \frac{\rho}{\tilde{t}^2} \right] )</td>
</tr>
<tr>
<td>ANCOVA</td>
<td>( \frac{\tilde{t}^2 (\sigma^2 + \tau^2)}{nmq(\tilde{t}^2 - q^2)^2} \left[ -m \rho^2 (1 - q) + \rho \left( m - \frac{\tilde{t}^2}{\tilde{t}^2 - q^2} - q (m - 1) \right) + \frac{\tilde{t}^2}{\tilde{t}^2 - q} \right] )</td>
</tr>
</tbody>
</table>

Table 4.2: True and estimated variance for the three estimators, where \( U = \tilde{t}^2 - \frac{sm^2}{m+1} \)

for all of the estimators, including the ones where numerical methods were needed. As was the case in Chapter 3, these curves do not depend on the choice of \( \sigma^2 + \tau^2 \), \( n, \mu_0 \) or \( \beta \). Some plots for \( q = 1/2 \) and a number of follow-up time vectors can be found in Figure 4.2.

Unlike in Chapter 3, we obtained different curves for the (naive) ANCOVA and for the random intercept ANCOVA. The numerically computed curve for the random intercept ANCOVA is again the same curve as one would obtain when ignoring the misspecification. This means that again, the misspecification is not relevant for the estimation of \( \beta \), and the variance of this parameter will be estimated consistently. However, we do see that including a random intercept in the model improves the estimate for the treatment effect.

As was to be expected from the Cramér-Rao lower bound, the maximum likelihood estimator using the true model cannot be beaten. For large correlations, all estimators apart from the naive "outcomes" estimator provide very similar performance. For small correlations, all estimators apart from the change score estimator provide similar performance.

If the correlation is unknown, neither of the random intercept models are a bad choice. Interestingly, for some regions of correlations, the naive estimators (which would not be used in practice), provide a very similar performance to the random intercept approaches. In particular, the naive ANCOVA is close to the random intercept ANCOVA. However, the naive approaches give biased estimates of the variance, leading to incorrect \( p \)-values when testing the significance of the treatment effect. Finally, we conclude that for multiple follow-ups, the
Figure 4.2: Variance of the estimated treatment effect as a function of $\rho$ for $q = 1/2$

random intercept ANCOVA cannot be said to be a bad choice under the assumed model. It does and cannot perform as well as the maximum likelihood estimator using the true model, but it is very close.
Chapter 5

Using an inclusion criterion based on baseline measurements

In this chapter, we will look at a different distribution for the random intercept. In Chapters 3 and 4 we have used a normal distribution for the random intercept variable. Here, we will look into a different distribution that might occur in practice when applying a selection criterion before admitting patients to a trial. This criterion, based on the baseline measurement, changes the distribution of both the baseline measurement as well as the random intercept. The distribution for the follow-up measurements is also changed, but only through the changed random intercept distribution. We will first give the details for these distributions, and then look at the consequences for the results already derived in previous chapters.

5.1 The effect on the distributions

We will again assume a normal distribution for the variable \( \mu_i \), namely

\[
\mu_i \sim N(\mu_0, \tau^2).
\]

However, now assume the trial is designed in such a way that only patients with a baseline measurement above a certain threshold can be included in the study. More specifically, patients are selected such that for their baseline measurement,

\[
Y_{i,0} = \mu_i + \epsilon_{i,0} \geq K
\]

for some threshold \( K \). We will assume the baseline measurement is done for a decently sized population, from which patients to be included in the study are randomly selected from those satisfying the criterion. For example, only patients with a baseline blood pressure above 140 could be included in the study. This changes the distribution for \( \mu_i \) as well as \( Y_{i,0} \). The distribution of the follow-up measurements, defined as \( \mu_i + \epsilon_{i,k} \), also changes, but only because the changed distribution of \( \mu_i \). Using

\[
\Pr(\mu_i \leq x | \mu_i + \epsilon_{i,k} \geq K) = \frac{\Pr(K - \epsilon_{i,k} \leq \mu_i \leq x)}{\Pr(\mu_i + \epsilon_{i,k} \geq K)}
\]

and conditioning on \( \epsilon_{i,k} \) for the numerator, we can find the probability density function

\[
\frac{1}{\tau} \phi \left( \frac{x - \mu_0}{\tau} \right) \left( 1 - \Phi \left( \frac{K - x}{\sigma} \right) \right) \left( 1 - \Phi \left( \frac{K - \mu_0}{\sqrt{\sigma^2 + \tau^2}} \right) \right),
\]

(5.3)
Figure 5.1: Densities of $\mu_i$ and $Y_{i,0}$ given that $Y_{i,0} \geq K$ for $\mu_0 = 130$, $\sigma = 3$, $\tau = 6$ and $K = 140$.

where $\phi$ and $\Phi$ are the standard normal probability density and distribution function respectively. For the baseline measurement, we find distribution function

$$Pr(Y_{i,0} \leq x|Y_{i,0} \geq K) = \frac{Pr(K \leq Y_{i,0} \leq x)}{Pr(Y_{i,0} \geq K)} = \frac{\Phi \left( \frac{x - \mu_0}{\sqrt{\sigma^2 + \tau^2}} \right) - \Phi \left( \frac{K - \mu_0}{\sqrt{\sigma^2 + \tau^2}} \right)}{1 - \Phi \left( \frac{K - \mu_0}{\sqrt{\sigma^2 + \tau^2}} \right)}$$

(5.4)

and density

$$1 \sqrt{\sigma^2 + \tau^2} \phi \left( \frac{x - \mu_0}{\sqrt{\sigma^2 + \tau^2}} \right),$$

(5.5)

for $x \geq K$. For $x < K$, both expressions are zero. The two densities are depicted in Figure 5.1 for a certain choice of parameters.

The full model is now given by

$$\mu_i \sim m(x) = \frac{1}{\tau} \phi \left( \frac{x - \mu_0}{\tau} \right) \left( 1 - \Phi \left( \frac{K - \mu_0}{\sqrt{\sigma^2 + \tau^2}} \right) \right),$$

$$Y_{i,0} \sim b(x) = \frac{1}{\sqrt{\sigma^2 + \tau^2}} \phi \left( \frac{x - \mu_0}{\sqrt{\sigma^2 + \tau^2}} \right),$$

$$Y_{i,k} = \mu_i + \beta \delta_i + \epsilon_{i,k} \text{ for } k = 1,...,m,$$

(5.6)

where the $\mu_i$ and $\epsilon_{i,k} \sim N(0,\sigma^2)$ are all independent variables. The given functions are the density functions for the respective random variables.

It is now useful to wonder about what will happen when we use this selection criterion. Of
course, since the distributions change, the correlations between measurements can be expected to change as well. Also, if $K - \mu_0$ is large compared to $\sigma^2 + \tau^2$, the distribution for $Y_{i0}$ will be sharply peaked right at $K$. This is because the probability of $Y_{i0}$ being larger than $K$ gets smaller with decreasing variance. Thus, the baselines of the included patients will all be very close to $K$, simply because it is very rare to exceed $K$ by a large amount. This means that theoretically, choosing such a strict criterion will cause the baseline variable to be almost useless in estimating the treatment effect. This would however not be encountered in practice, since it would be very hard to reach an acceptable number of patients included in the study when using too strict criteria.

Further, we note that the baseline measurements of every patient are taken to be at least $K$, while the characteristic blood pressure ($\mu_i$, the random intercept belonging to that patient) can also be smaller than $K$. This phenomenon can also be seen in Figure 5.1. However, the follow-up measurements of patients in the control group will be somewhere around $\mu_i$. This means it may happen that for some patients we see a downward change in blood pressure, without applying any treatment. Similarly, this effect may happen in the treatment group and add to the true treatment effect. This may cause estimates for the true treatment effect to be biased when using certain methods, and is known as "regression to the mean". In this chapter, we will look at what happens when using different methods, including the methods seen in previous chapters. We will do this using simulations. First, we will go into the simulation methods and approaches to estimate the treatment effect.

5.2 Simulation strategy

We start from the same two models as seen in Chapters 3 (the treatment causing a single sustained effect) and 4 (the treatment setting the outcome variable on a slope), but instead of simulating a single data set in every iteration, we simulate a population large enough so that there are enough patients satisfying $Y_{i0} \geq K$. Then we apply the estimating methods to a subset of the patients with baseline at least $K$ of the desired size. The R function lm is used to estimate random effects.

We will fix $n, q, \sigma^2 + \tau^2, \beta, \mu_0$ and $K$ and take fifty correlations ranging from 0.01 to 0.99. Note that we will use the pre-selection correlations $\tau^2/(\sigma^2 + \tau^2)$. As stated, the post-selection correlations will very likely be different. However, it may be more intuitive to use the correlations that might be expected to hold in general in a population, which is what the pre-selection correlation entails. For each correlation, we will repeatedly generate a population and data set as described, and apply all of the desired approaches to obtain estimators. After a number of iterations, we save the mean and variance of each estimator, and move on to the next correlation.

This is all done for four different scenarios (number of follow-ups or follow-up time points), for both models, for a total of eight scenarios. In particular, we choose $n = 1000, q = 1/2, \sigma^2 + \tau^2 = 40, \mu_0 = 130, K = 140, \beta = -5$ for the first model and $\beta = -1.5$ for the second (slope) model. Finally, we use a population size of a hundred thousand and one thousand iterations. The chosen $n$ is such that the whole process can be finished in reasonable times (a few hours per scenario), while still being large enough to be satisfied about approaching asymptotic results.

For each scenario, the mean squared error of the estimators will be computed for each correlation through the bias and variance found. From this, we can make the plots $n\text{MSE}/(\sigma^2 + \tau^2)$, choosing the scaling to be able to compare the plots to those found in previous chapters. Now we will describe some new approaches that will be used to estimate the treatment effect.
5.3 Estimating the treatment effect

In this simulation approach, we will use the same methods as seen in previous chapters along with two others. The familiar approaches are the (naive) ANCOVA, autoregressive and outcomes approaches, along with the random intercept, ANCOVA with random intercept and change score estimators. These all follow the same formulation as before. Apart from these approaches, we will also use two more that will hopefully be able to handle the regression to the mean.

Random intercept without baseline measurements

The first new approach is the random intercept approach where we ignore the baseline measurements. That is, the baseline measurements will only be used for the selection criterion, and not to estimate the treatment effect. This means there is no effect from regression to the mean between the baseline measurement and the first follow-up, which might help in avoiding the bias that can come from this phenomenon. This means for the first model we take

\[ Y_{i,k} = \mu_{i} + \beta\delta_{i} + \epsilon_{i,k} \]  

for \( k = 1, \ldots, m \). Here we have, as before, \( \mu_{i} \sim N(\mu_{0}, \tau^{2}) \) independent from the i.i.d. \( \epsilon_{i,k} \sim N(0, \sigma^{2}) \). Similarly, for the slope model,

\[ Y_{i,k} = \mu_{i} + \beta t_{k}\delta_{i} + \epsilon_{i,k} \]

for \( k = 1, \ldots, m \).

Autoregressive approach with random intercept

This approach is the same as the autoregressive approach, except that we accommodate for the random intercepts. Thus for model 1, the approach can be written as

\[ Y_{i,k} = \mu_{i} + \beta\delta_{i} + \alpha Y_{i,k-1} + \epsilon_{i,k}, \]  

and similarly for model 2

\[ Y_{i,k} = \mu_{i} + \beta t_{k}\delta_{i} + \alpha Y_{i,k-1} + \epsilon_{i,k}. \]

Here \( k = 1, \ldots, m \) and \( \mu_{i} \) and \( \epsilon_{i,k} \) are as we have already seen many times.

5.4 Results

The results for a particular choice of correlation, \( \rho = 0.71 \), can be found in Table 5.1 (model 1) and Table 5.2 (model 2). It is clear that some methods that were previously unbiased are now biased. In both tables, we see that including an intercept with the change scores ensures an unbiased estimate of the treatment effect. This means that indeed, the added intercept explains the regression to the mean, leaving the true treatment effect to be given by \( \beta \). Similarly, leaving out the baseline measurement altogether when using the random intercept approach yields an unbiased estimate. Both ANCOVA approaches seem to be unbiased, although using a random intercept increases the stability of the estimate for model 2 (for model 1, we have already seen that both the naive and random intercept ANCOVA yield the exact same estimator and thus the same variance). Finally, the autoregressive approach with random intercept seems usable for model 1 (albeit slightly biased for \( m = 2 \)), but for model 2 it performs significantly worse.

In Figures 5.2 and 5.3 the mean squared error of the competitive estimators are given. The reason we have used the mean squared error instead of the variance is to accommodate possible
bias. Only the (almost) unbiased estimators were seen as competitive, due to the bias severely increasing the mean squared error for the others. In Figure 5.2, only one of the ANCOVA methods is graphed since both approaches are equivalent in terms of estimating the treatment effect. For the second model this is not the case, and so both approaches are included in Figure 5.3.

<table>
<thead>
<tr>
<th>Method</th>
<th>1</th>
<th>2</th>
<th>5</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Change scores (without intercept)</td>
<td>-8.69(0.21)</td>
<td>-8.69(0.18)</td>
<td>-8.68(0.15)</td>
<td>-8.69(0.14)</td>
</tr>
<tr>
<td>Change scores (with intercept)</td>
<td>-5.01(0.30)</td>
<td>-4.99(0.25)</td>
<td>-5.00(0.21)</td>
<td>-5.00(0.19)</td>
</tr>
<tr>
<td>Outcomes</td>
<td>-7.47(0.22)</td>
<td>-6.84(0.21)</td>
<td>-6.05(0.20)</td>
<td>-5.62(0.20)</td>
</tr>
<tr>
<td>ANCOVA</td>
<td>-5.01(0.29)</td>
<td>-4.99(0.25)</td>
<td>-5.00(0.21)</td>
<td>-5.01(0.19)</td>
</tr>
<tr>
<td>ANCOVA with random intercept</td>
<td>-5.01(0.29)</td>
<td>-4.99(0.25)</td>
<td>-5.00(0.21)</td>
<td>-5.01(0.19)</td>
</tr>
<tr>
<td>Autoregressive approach</td>
<td>-5.01(0.29)</td>
<td>-4.24(0.23)</td>
<td>-3.53(0.18)</td>
<td>-3.13(0.15)</td>
</tr>
<tr>
<td>Autoregressive approach with r.i.</td>
<td>-5.01(0.29)</td>
<td>-4.82(0.27)</td>
<td>-4.95(0.24)</td>
<td>-4.99(0.22)</td>
</tr>
<tr>
<td>Random intercept (with baselines)</td>
<td>-7.72(0.23)</td>
<td>-7.62(0.19)</td>
<td>-7.56(0.16)</td>
<td>-7.55(0.15)</td>
</tr>
<tr>
<td>Random intercept (without baselines)</td>
<td>-5.00(0.31)</td>
<td>-5.00(0.26)</td>
<td>-5.00(0.23)</td>
<td>-5.00(0.22)</td>
</tr>
</tbody>
</table>

Table 5.1: Simulation results for the first model with different values for \( m \) and \( \rho = 0.71 \), given as mean (standard error) over 1000 iterations.

In the case of model 1, the random intercept autoregressive approach is not depicted for \( m = 2 \) for the reason that for correlations around 0.95, the MSE increases greatly. This is not a fault of the randomness in the simulation, since it could be replicated by trying again. In the other cases, it behaved normally. Further, as before, the results might be independent of \( \sigma^2 + \tau^2 \) and \( n \). In this case it is harder to verify through the computation time needed to run a simulation, but we did see very similar plots and the same conclusion from the plots when trying different values. The plots also seem robust under changing \( \beta \), although we need to be cautious with the autoregressive approach with random intercept being biased. Finally, we have not touched \( \mu_0 \) and \( K \), since these parameters together will most likely change the results through changing the distributions (most importantly, the variance of the baseline measurement).

In Chapter 5 we saw that some estimators perform better for small correlations while others do well for larger correlations. We again see similar results here, namely that for small correlations, the change score estimator (with intercept) performs worse than the other depicted methods, which all perform similarly. For large correlations, the random intercept (ignoring baselines) approach performs worse than the change score and ANCOVA estimators. The same holds true for the autoregressive approach with random intercept, except for \( m = 1 \) where it is (by definition) the same as the ANCOVA approach. For medium \( \rho \) (somewhere around \( \rho = 0.5 \)), the ANCOVA seems to consistently be a good choice, with the other estimators performing slightly worse. For other correlations, the ANCOVA also provides good performance, although then some others are comparable depending on \( \rho \).

For the second model, not as much can be said about the plots. Here we have also tried another
Figure 5.2: Mean squared errors for the relevant estimators, with \( \sigma^2 + \tau^2 = 40, \mu_0 = 130, K = 140, \beta = -5 \) and \( q = 1/2 \).

set of parameters for one of the scenarios, and the plot was very similar. The random intercept approach (ignoring baselines) is the clear winner here, at least for the sets of follow-up times we have chosen. The random intercept ANCOVA is quite close, especially for the last three scenarios given in Figure 5.3. The naive ANCOVA is the worst approach for relatively large correlations, and the change score estimator (again including an intercept) performs badly for small \( \rho \). This examples serves as a nice illustration that there can be no one-sided conclusion, since different approaches perform differently for different functional forms of the outcome variable through time.
Follow-up times (1, 2)

Follow-up times (1, 2, 3)

Follow-up times (1, 2, 3, 4)

Follow-up times (1, 3, 6, 12)

Figure 5.3: Mean squared errors for the relevant estimators, with $\sigma^2 + \tau^2 = 40, \mu_0 = 130, K = 140, \beta = -1.5$ and $q = 1/2$. 
Chapter 6

Conclusion and discussion

To form a conclusion seems somewhat daunting at first, since each chapter had many results on its own. The best approach to use heavily depends on the underlying data generating mechanism, the believed form of the behaviour through time of the outcome variable, the correlation between the measurements and the number of follow-ups or the times at which follow-up measurements are done. The work we have done only compares a few approaches for only a few such data generating mechanisms. In short, there is no one method that always performs best.

While this may look like a grim statement, there is also an upside, namely the widely used ANCOVA method (if necessary with random intercepts). While it cannot always be said to be the best choice, it turns out not to be a bad one either. In Chapters 3 and 4 it was observed that while it cannot be as good as using the true likelihood, it is still relatively close for every \( \rho \). For the simulated example shown in Chapter 5 this was also the case. It is able to correct for regression to the mean as seen in Chapter 5 and also does not lose much power if such a correction is not necessary. Overall, we can conclude that the ANCOVA is relatively robust under the specifics of the context that may be unknown. There is no evidence that the method cannot be used in the cases we have explored.

Many of the other methods were seen to have their own merits in specific situations. However, without background knowledge, it is hard to judge how the method will perform. For example, for large correlations between measurements, using the change score approach may be fruitful. On the other hand, the decision on whether or not to include an intercept may increase the variance or introduce bias when one is not careful. Thus, one should always be careful when choosing an estimation method. With good grounds, these methods can still be used and provide good results. Unfortunately, there simply is no holy grail method.

There is a lot of room for further research in this topic. We have only worked on continuous data where normal distributions are applicable, and on two specific examples of behaviour throughout time. Both choices can be expanded to other types of data (for example count data, skewed distribution) and more functional forms of the outcome variable through time. Secondly, the methodology can be expanded. As noted, even within the realm of maximum likelihood estimation, there are more methods that can be used. Besides maximum likelihood, there are many more estimation methods, such as restricted maximum likelihood and generalised estimating equations, that can be looked into. As the results differ case by case, it is hard to take a more generalised approach towards this problem.
BIBLIOGRAPHY


