AALBORG UNIVERSITY

REPREATED MEASUREMENTS

THE EFFECT OF IBUPROFEN ON WRIST FUNCTIONS AFTER COLLES FRACTURE

Author:
Lotte WOLSTED

Advisor:
Jakob G. RASMUSSEN

May 31st 2018

Front page is illustrated by Casper Vagnø Wolsted

School of Engineering and Science
Institut for Matematiske Fag
Skjernvej 4A
DK-9220 Aalborg Ø
http://www.math.aau.dk

AALBORG UNIVERSITY
STUDENT REPORT

Title: Repeated measurements
Subject: The effect of Ibuprofen on wrist functions after Colles fracture
Author: Lotte Wolsted
Advisor: Jakob G. Rasmussen
Copies: 2
Pages: 104
Project periode: Spring semester 2018
Handin date: May 31st 2018

Synopsis:

The purpose of this report is to determine whether Ibuprofen has any beneficial effect, wrt. recovery of wrist functions, on patients having been through surgery to correct a Colles fracture. The patients have been treated with either Ibuprofen or a placebo drug. At three timepoints after the surgery, measurements of how far the patients are able to bend and rotate their injured wrist have been recorded. If Ibuprofen does have a beneficial effect, we should be able to see, that the patients, who were given Ibuprofen, faster regain full mobility of the injured wrist than the patients, who were given a placebo drug.

To determine whether there is a difference between treatment groups, I will use three approaches. First, I will use ANOVA and multivariate ANOVA. This approach is a little too simple to say anything final about the relationship between the treatment groups, but it does give some idea of whether there might be a difference in the groups. Next, I will use linear mixed effects models (LMMs). The treatment may have an effect on the measurements, but how well the patients are able to bend and rotate their wrist may also be influenced by themselves, by some genetic effect. Without having to actually specify any genetic markers, the LMM takes these individual effects into account by allowing for random effects in the setup of the model. With the LMM, it is also possible to account for any kind of correlation between observations on the same subject. In order to use the LMM, a handful of assumptions about the model needs to be satisfied. Lastly, I will use generalized estimating equations models (GEE-models). This approach resembles the LMM and it also allows for specification of correlation between observations on the same subject. An advantage of the GEE-model is that there are no model assumptions to be met. A disadvantage is, that it is not possible to specify random effects.

The results from both ANOVA, multivariate ANOVA, the LMMs and the GEE-models are all unanimous; there is no difference in the treatment groups, i.e. Ibuprofen has no significant beneficial effect.

Referat

I denne rapport undersøges det om indtagelse af Ibuprofen forbedrer helningsprocessen for patienter opereret for Colles fraktur. Med helningsproces menes der, hvor hurtigt de genvinder mobilitet af den skadede hånd. Patienterne er inddelt i tre grupper ift. den medicin, de har taget i ugen efter operationen. På tre tidspunkter i løbet af det første år efter operationen, er der målt i grader, hvor meget patienterne kan bøje og rotere den skadede hånd. Samme målinger er lavet på den ikke-skadede hånd for at have et mål for, hvor meget hænderne normalt bør kunne bøjes og roteres.

For at undersøge om der er forskel i målingerne på de tre patientgrupper, udføres først både en ANOVA test og en multivariat ANOVA test. Disse test er blot med for at give en indikation af, om der er forskel på grupperne.

Dernæst beskrives teori om lineære mixed modeller. Det er muligt, at raten for patienternes helningsproces er påvirket af deres gener. Det anvendte datasæt indeholder ingen genetiske variable, så det er ikke muligt at opstille en almindelige lineær model, hvori der indgår variable for diverse gener for på den måde at tage højde for evt. genetisk indflydelse. Vha. den lineære mixed model kan vi dog tage højde for, at helningsprocessen kan være påvirket af ikke-målte faktorer via *random effects*. Vi er nødt til at tage højde for, at målingerne hen over tiden for hver patient kan være korrelerede på en eller anden måde. Dermed er det nødvendigt at anvende en model, der kan indkorporere diverse former for korrelation mellem målingerne. I den almindelige lineære model antages samtlige målinger at være uafhængige. Den lineære mixed model tillader en række af forskellige korrelationsstrukturer.

Slutteligt beskrives teori om generaliserede estimationsligninger (GEE-modeller). Denne tilgang minder lidt om lineære mixed modeller i den forstand, at de er anvendelige i tilfælde med korreleret GLM-agtigt data. Fordelen ved GEE-modeller er, at man ikke behøver gøre sig antagelser om fordelingen af hverken data eller fejlledene, hvilket er nødvendigt, hvis man anvender lineære mixed modeller. En ulempe er dog, at man ikke kan specificere random effects. Vha. random effects kan man komme helt ned og sige noget om de patient-specifikke effekts. I GEE-modeller kan vi "kun" undersøge den gennemsnitlige effekt i grupperne. Da vi ikke er interesserede i at følge hver enkelt patient for at finde den bedste behandling for ham/hende, men istedet gerne vil vide om den allerede anvendte behandling generelt har medført en forskel i helningsprocesserne blandt grupperne, så giver det ganske god mening at anvende en GEE-model.

Resultaterne fra både ANOVA, multivariat ANOVA, den lineære mixed model og GEE-modellen peger alle på, at der ingen forskel er på patientgrupperne. Dermed har indtagelse af Ibuprofen ingen signifikant indflydelse på helningsprocessen.

Contents

Preface

I have been given a data set with patients all of whom had suffered a Colles fracture. All received the same type of corrective surgery and, subsequently, were placed in one of three groups according to treatment. Some were given Ibuprofen for the pain while others were treated with a placebo drug. The objective is to find out whether there is a difference between the patients treated with Ibuprofen and the patients treated with a placebo drug wrt. how well and how fast they regain mobility of the injured hand in the course of a year after surgery. If the type of treatment plays no role in the patients' recovery, then the costs of treatment can be reduced as no pain medicin needs to be prescribed. It is suspected that Ibuprofen may cause bone deterioration, which is another reason to want to find out whether patients do just as well without the drug wrt. recovery of wrist functions. The hypotheses to be tested are

- H_0^{main} : *There is no difference wrt. recovery of wrist functions between the treatment groups.*
- H_A^{main} : *Patients treated with Ibuprofen have a better rate of recovery than the other patients.*

I will be working with a lot of hypotheses during the report. To distinguish these, I am adding appropriate superscripts. I should also note, that when I say a hypothesis is true, what I really mean is, that there is no evidence to reject it. Three different types of movement of the injured hand have been recorded. This means that each patients' response is multivariate. Furthermore, these recordings were made at three different timepoints after the surgery. This means that we have not just one response per patient per type of movement, but instead have a vector of responses for each patient, i.e. we have *repeated measurements*. While the responses between patients certainly are independent, the entries within each vector of responses may very well be correlated. An ordinary linear model or a generalized linear model cannot handle responses being correlated, and as such a different approach is needed. One approach is to use *analysis of variance* in which the means of each group of patients are compared. This approach, however, does not allow for comparison across timepoints. Another approach is the *mixed model*, which is essentially just a linear model with random effects added into the linear predictors. The idea is that the dependence between observations within and between groups are effected by some latent variables. The mixed model allows for a wide range of correlation patterns. A third approach is to use *quasi likelihood* and *generalized estimating equations*. When using quasi likelihood, no assumptions about the distribution of the observations are needed; specifying the mean and variance is all that is required. In order to estimate the unknown parameters, the generalized estimating equations are used, in which the correlation structure of the observations is specified. I will use all three approaches.

The theory behind each of the three approaches are described in seperate chapters. Each of these chapters are followed by a chapter in which the data is analysed in `R` using the theory just described. Each of these chapters end with a conclusion as to whether H_0^{main} is rejected or not based on the analysis.

Notation and other helpful information

All calculations and figures are made in the statistical software program `R`. References to `R`-code will be written in a certain font such as `geeglm()`.

To save myself a lot of repeated coding, I have made three functions that do just about everything I need in my analysis with the three approaches. These functions are called `superANOVA()`, `superLMM()` and `superGEE()`. Depending on which of the three approaches, I am working with, these functions do anything from testing model assumptions to plotting residuals and calculating standard errors. The in-

terested reader can find the source code for these functions in the chapters, where they are used.

References to various literature are noted in brackets [ ], e.g. [9].

For a smoother read, some of the theoretical calculations and results are omitted from the pages, where they are used. These calculations/results will be listed in Appendix A. All references to the appendix will be denoted by e.g. A.3; this is a reference to the 3rd result in Appendix A. Appendix B contains additional theory. The reader may use this to acquaint or re-acquaint themselves with some of the theory otherwise left out of the chapters. References to Appendix B is denoted by e.g. B.3.

All vectors or matrices are noted in bold letters. Vectors are always in small letters, and matrices are always in capital letters. Random variables are noted in non-bold capital letters. Random vectors are noted in bold capital letters, like matrices. It should be clear from context which is which. If nothing else is specified, all vectors will be column vectors, i.e.

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} x_1, \dots, x_n \end{bmatrix} \in \mathbb{R}^{n \times 1}.$$

A row vector will be denoted as a transposed column vector, e.g. $\mathbf{x}^T = \begin{bmatrix} x_1 & \dots & x_n \end{bmatrix} \in \mathbb{R}^{1 \times n}$.

Let $\mathbf{I}_{n \times n} = \mathbf{I} \in \mathbb{R}^{n \times n}$ denote the identity matrix, and let $\mathbf{1}_n = [1, \dots, 1] \in \mathbb{R}^{n \times 1}$ denote a vector of all 1s.

At times during calculations in the report, I will refer to other equations in the calculations themselves. These references will be noted above $=, \leq, \propto$ and the likes. For example

$$ab \overset{(4.1)}{=} c.$$

Here, I am using equation 1 from Chapter 4 in order to show than ab can be written as c.

A diagonal matrix is noted by

$$\mathbf{A} = \begin{bmatrix} a_1 & & \\ & \ddots & \\ & & a_n \end{bmatrix} = \text{diag}\{a_i\}_{i=1,\dots,n}.$$

An entry in a matrix may be denoted by $\mathbf{A}_{j,k}$. This is the element in the jth row and kth column in $\mathbf{A}$.

The determinant of a matrix is noted by $|\mathbf{A}|$.

All examples ends with a square, □.

At times, I will digress from whatever theory is being described to make a comment about my data. To avoid having to write "Wrt. my data..." too many times, I use a "**!**" to stress that I am making a comment about my data.

Acknowledgements

I would like to thank my advisor, Jakob G. Rasmussen, for constructive critisisme and ideas during the course of this project. I would also like to thank Marius Aliuskevicius, M. D. at Aalborg University Hospital, for lending me the data for analysis.

1 Description of data

The data consists of 83 patients, whom have all suffered a Colles fracture (broken bone in the wrist). All have received the same surgery from the same surgeon. They have all been treated by the same medical team, and all data on the patients have been recorded by the same people. In other words, all patients have received the same kind of treatment all through the experiment, apart from the type of drug they have been treated with. The patients were randomized into three groups and had to take medicine the first seven days after surgery. Group 1 took a placebo drug, group 2 took Ibuprofen for three days and a placebo drug for the next four days and group 3 took Ibuprofen. There are 28 patients in group 1 and group 2, and 27 patients in group 3. Three pairs of movement of the wrist were recorded on each patient at 6 weeks, 3 months and one year after the surgery. The movements are pronation and supination (rotating hand), extension and flexion (bending hand forward and backward) and ulnar and radial ("waving" hand from side to side). The same movements have been recorded on the uninjured hand, though only once.

I received three datasets; one with the measurements in degrees of the uninjured hand, one with the measurements in degrees of the injured hand, and one that describes in percentages how close the ranges of movement of the injured hand is to the ranges of the uninjured hand. By "range", I mean the degrees from, for instance, the hand bending as far forward as possible to the hand bending as far backward as possible. I will be working only with the dataset with measurements given in percentages. I have gathered all these measurements in a new dataset, I call `Wrist`. Below are the first nine rows of `Wrist`:

```
> head(Wrist,9)
  subject       ps       ef        ur time group
1    id01 72.22222 36.000000  40.00000    0     1
2    id01 86.11111 68.000000  80.00000    7     1
3    id01 94.44444 68.000000  50.00000   46     1
4    id02 91.17647 47.058824  37.50000    0     1
5    id02 91.17647 76.470588  62.50000    7     1
6    id02 97.05882 76.470588 100.00000   46     1
7    id03 68.57143  4.166667  22.22222    0     1
8    id03 94.28571 79.166667  77.77778    7     1
9    id03 97.14286 91.666667 100.00000   46     1
```

The first column, `subject`, contains an id identifying each patient. The second column, `ps`, are all the measurements for the ranges in percentages of pronation and supination, which I will henceforth just call the pro/sup-movement. There are three measurements of the pro/sup-movement for each patient; one taken at 6 weeks after surgery (1st timepoint), one taken 3 months after surgery (2nd timepoint) and one taken one year after surgery (3rd timepoint), and they are listed in that order. The same goes for the ex/flex- and uln/rad-movements (the columns `ef` and `ur`). The fifth column, `time`, indicate at what timepoint the measurement was taken. In order to indicate that the time between timepoints is not equidistant, I have decided not to code the timepoints as "1", "2", "3", which would have been an obvious way of coding the timepoints. Instead, I have coded the timepoints so they show how many (approximately) weeks have passed since the first timepoint. For the 2nd timepoint, approximately 7

weeks have passed. This is calculated by estimating how many weeks are in three months:

$$\frac{52 \text{ weeks in a year}}{12 \text{ months in a year}} \approx 4.333 \text{ weeks per month}$$
$$\Downarrow$$
$$4.333 \cdot 3 = 13 \text{ weeks in three months.}$$

Then the initial 6 weeks until the 1st timepoint must be subtracted, which gives 7 weeks between the 1st and 2nd timepoint. Between the 1st and 3rd timepoint, 46 weeks have passed, which is calculated simply by $52 - 6 = 46$ weeks. The last column, `group`, indicate which of the three treatment groups, the patient belongs to, either "1", "2" or "3".

Just to get an initial idea of whether there might be a difference in the responses from the treatment groups, I have made Figure 1. This figure shows for each type of movement, how well in percentages the injured hand on average per group performs in comparison to the uninjured hand on average per group. There are no obvious differences between the groups. Only in the first plot do group 3 seem to do a little better.

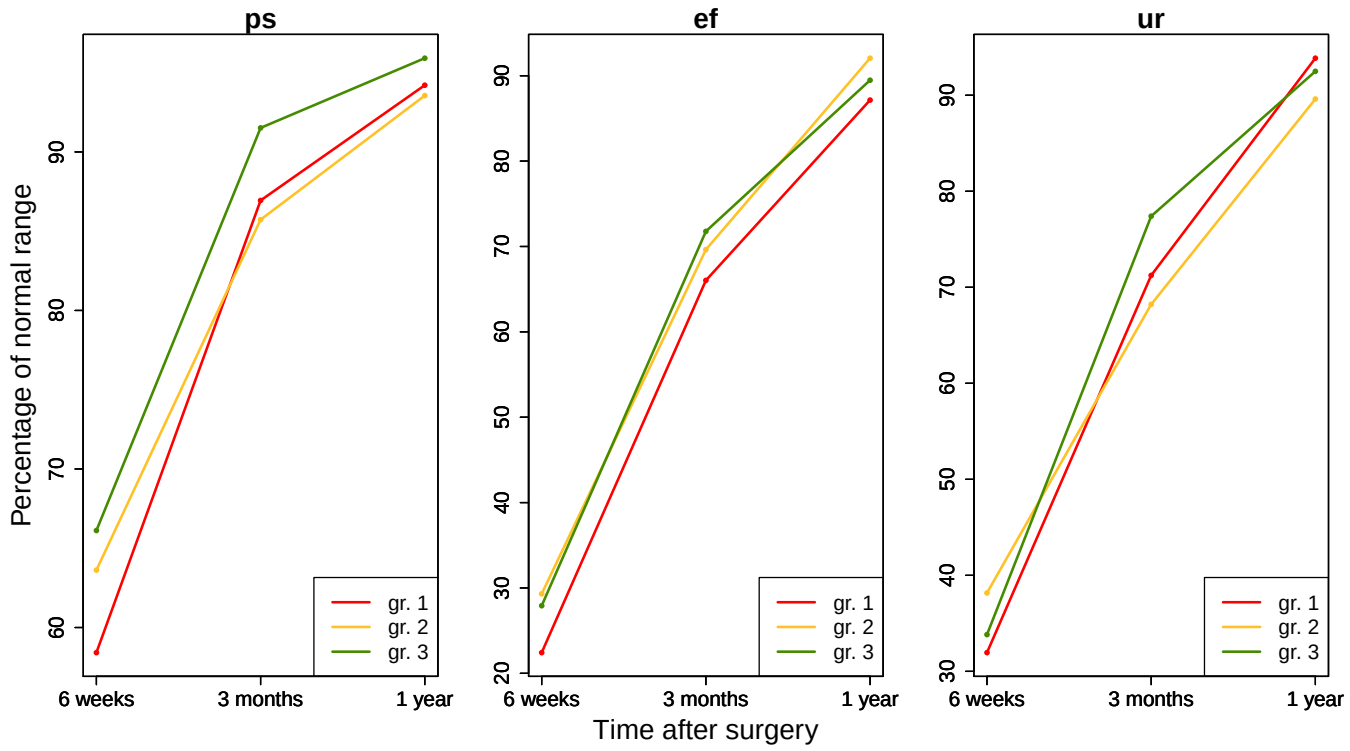

Figure 1: *Each line represent the average percentage of normal range of the pro/sup-, ex/flex- and uln/rad-movements for each group.*

Figure 2 show boxplots of the ranges of movement at the different timepoints. If H_0^{main} is true, we might expect to see, that the median for group 3 is larger than for the other two groups. This is actually the case in the majority of the boxplots, but the differences in the medians are only very small.

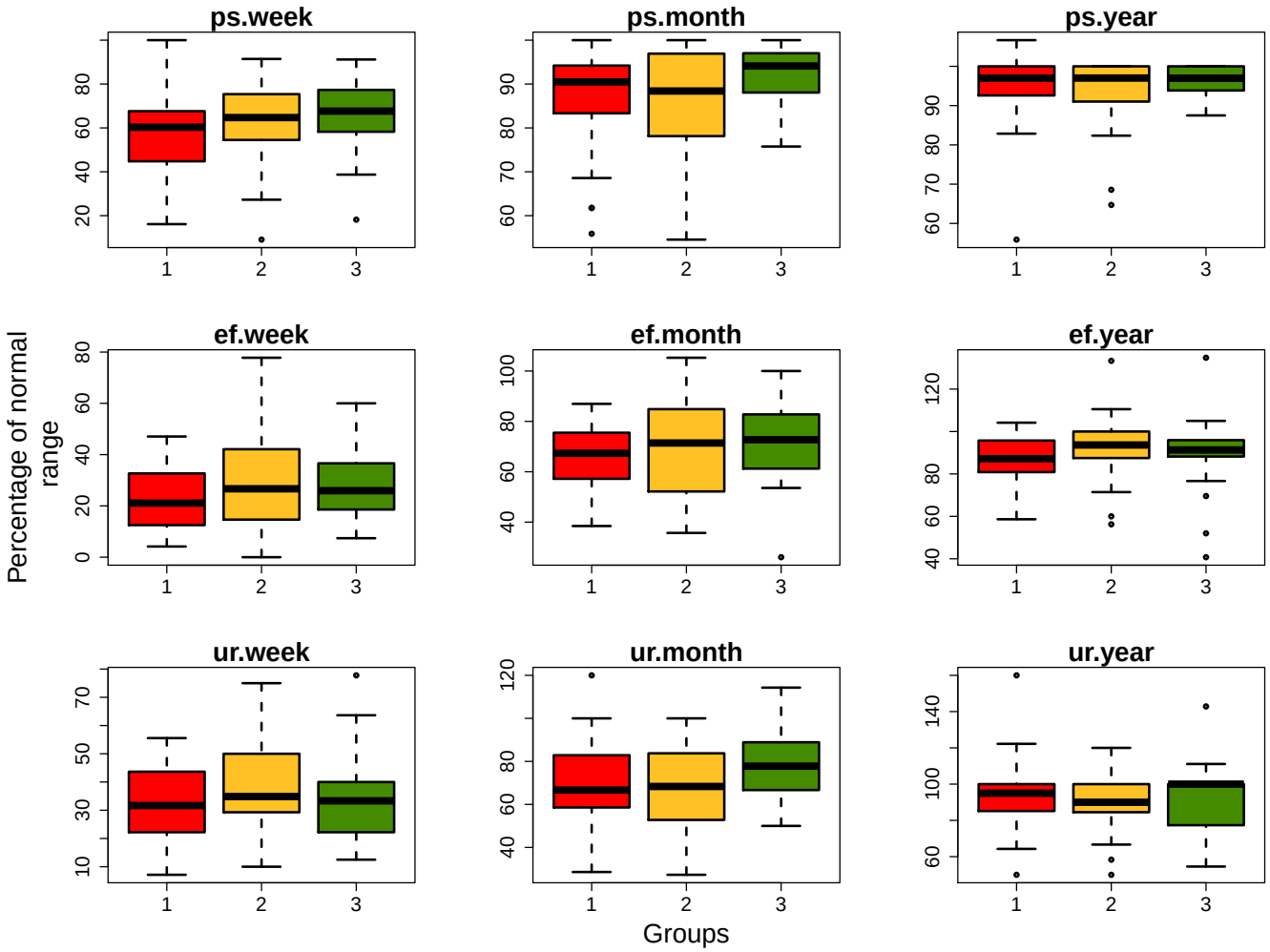

Figure 2: *Boxplots of the percentage of normal range of the movements at different timepoints.*

A quick and very simple way of getting an idea of the results, we may encounter later on, is to set up a linear model for each type of movement and compare it to an equivalent linear model without the `group`-term. For comparison, I use `anova()`[1], which tests the hypothesis[2]

$$H_0^{\text{anova}} : \textit{the models are not significantly different.}$$

The relationship between the types of movement and the covariates, `group` and `time`, is not linear, but, having investigated this further, linearity can in this case be achieved by adding a squared term of the timepoints to the models. Below are the results from comparing each of the linear models with their reduced counterparts:

```
> anova(lm(ps ~ time + I(time^2), data = Wrist),
+       lm(ps ~ group + time + I(time^2), data = Wrist))$"Pr(>F)"[2]
[1] 0.06431

> anova(lm(ur ~ time + I(time^2), data = Wrist),
+       lm(ur ~ group + time + I(time^2), data = Wrist))$"Pr(>F)"[2]
[1] 0.57625

```

[1] See B.1.

[2] H_0^{anova} is equavalent to $H_0^{\text{anova}_1}$ in B.1.

```
> anova(lm(ef ~ time + I(time^2), data = Wrist),
+       lm(ef ~ group + time + I(time^2), data = Wrist))$"Pr(>F)"[2]
[1] 0.05631
```

With a significance level of 0.05, each of these p-values indicate that the `group`-term is insignificant to the models, although for the pro/sup- and ex/flex-movement, we are not far from rejecting H_0^{anova}. The linear model is a little too simple to model the types of movement properly, but it does give some idea of the relevance of the treatment groups.

To get a clearer indication of whether there might be a difference between the groups, I will now move on to my first idea of testing H_0^{main}. In Chapter 2, the theory behind *analysis of variance* is explained and in Chapter 3, I test to see if H_0^{main} should be rejected when using analysis of variance.

2 | Analysis of variance

In order to determine whether there is a difference in the responses from a study where subjects are seperated into groups, one can perform a test to compare the group means. *Analysis of variance* (ANOVA for short) is a method for multiple comparisons, meaning two or more groups can be compared. In ANOVA, the variation within the samples and the variation between the samples are used to detect differences in the means. In a one-way ANOVA, the objective is to compare means of two or more samples, where groups are formed according to levels in one factor (hence *one*-way ANOVA).

The ANOVA is also known as the *univariate* ANOVA, as it assumes a univariate response. An extension of the univariate ANOVA is the *multivariate* ANOVA (MANOVA), which assumes a multivariate response. Both ANOVA and MANOVA are methods for comparing group means, however, neither ANOVA nor MANOVA allows us to compare group means across timepoints.

> **!** It may seem, as I have a multivariate response for each patient in my data, that only MANOVA is of interest, but I shall use both ANOVA and MANOVA to cement my conclusion about whether treatment plays a significant role in the patients' recovery. With ANOVA, I will just have to test each type of movement seperately.

A group can be several things. If a study consisted of recording some patients' blood pressure, say, once a month for a year, then each patient can be thought of as forming a group (or sample). In this case, the response for each patient is a vector. A study could also consist of the scores from some test distributed to, say, 7th graders. In this case, we only have one response per subject in the study. A possible choice for groups could be boys and girls. An observation, y_{ij}, can then either be an observation on the ith subject at the jth timepoint, or an observation on the jth subject in the ith group. In the following, y_{ij} is thought of as the observation on subject j in the ith group, but it need not be.

2.1 ANOVA

This section is based on [1] and [2].

In the univariate case, we have g independent samples (or groups)

$$\begin{aligned}\mathbf{y}_1 &= \left[y_{11}, y_{12}, \ldots, y_{1n_1}\right]\\ \mathbf{y}_2 &= \left[y_{21}, y_{22}, \ldots, y_{2n_2}\right]\\ &\vdots\\ \mathbf{y}_g &= \left[y_{g1}, y_{g2}, \ldots, y_{gn_g}\right].\end{aligned}$$

Here, n_i is the number of subjects in group i. The assumptions of the ANOVA are that $Y_{ij} \overset{\text{iid.}}{\sim} N\left(\mu_i, \sigma^2\right)$. That is, each group has an individual mean, μ_i, all groups have a common variance, σ^2, the data is Gaussian, and all subjects are independent. If $n_1 = n_2 = \ldots = n_g$, we call it a *balanced* design, and unbalanced otherwise. An advantage of ANOVA over MANOVA, is that ANOVA can handle unbalanced data. In the following, a balanced design, i.e. $n_i = n$ for $i = 1, \ldots, g$, is assumed.

> **!** With three treatment groups, I have $\mathbf{y}_1$, $\mathbf{y}_2$ and $\mathbf{y}_3$ with $n_1 = n_2 = 28$ and $n_3 = 27$ for each type of movement. Hence, I have an unbalanced design. For the sake of simplicity, I focus on a balanced design when describing the theory of ANOVA. It is also worth noting, that when performing ANOVA in `R`, the function `aov()`, that I will be using, can actually handle an unbalanced design.

The hypotheses to be tested are

$$H_0^{\text{ANOVA}} : \mu_1 = \mu_2 = \ldots = \mu_g$$
$$H_A^{\text{ANOVA}} : \mu_i \neq \mu_j \textit{ for at least one } i \neq j.$$

That is, we want to test whether all groups have the same mean. We do this by investigating the variances. The independent estimators of the variances for each sample are

$$s_i^2 = \frac{1}{n-1} \sum_{j=1}^{n} \left(y_{ij} - \bar{y}_{i\cdot} \right)^2, \quad i = 1, \ldots, g \tag{2.1}$$

with $\bar{y}_{i\cdot} = \frac{1}{n} \sum_{j=1}^{n} y_{ij}$ being the sample mean for group i. Having a balanced design means we can define an average variance for each group, i.e. the *within-group variance*:

$$s_w^2 = \frac{1}{g} \sum_{i=1}^{g} s_i^2 \;=\; \frac{1}{g(n-1)} \sum_{i=1}^{g} \sum_{j=1}^{n} \left(y_{ij} - \bar{y}_{i\cdot} \right)^2.$$

If H_0^{ANOVA} is true, we can regard the sample means as being $\bar{Y}_{i\cdot} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$. This leads to an overall average of the variance across groups, i.e. the *between-group variance*:

$$s_b^2 = \frac{n}{g-1} \sum_{i=1}^{g} \left(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot} \right)^2 \;=\; \frac{1}{g-1} \sum_{i=1}^{g} \sum_{j=1}^{n} \left(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot} \right)^2,$$

with $\bar{y}_{\cdot\cdot} = \frac{1}{g} \sum_{i=1}^{g} \bar{y}_{i\cdot} = \frac{1}{gn} \sum_{i=1}^{g} \sum_{j=1}^{n} y_{ij}$ being the total sample mean. We have that s_w^2 always is an unbiased estimator of σ^2, but s_b^2 is only an unbiased estimator if H_0^{ANOVA} is true (see A.1). If H_A^{ANOVA} is true, then $\mathrm{E}\left[s_b^2\right] > \sigma^2$. We only have that $\mathrm{E}\left[s_b^2\right] = \mathrm{E}\left[s_w^2\right]$ when H_0^{ANOVA} is true, and so H_0^{ANOVA} should be rejected if s_b^2 is significantly larger than s_w^2, i.e. if the ratio s_b^2 / s_w^2 is larger than 1 as it would indicate a significant difference between the groups.

A disadvantage of ANOVA is that it assumes that the correlation is the same for any pair of observations within a group. The MANOVA makes no assumptions about the correlation between observations.

2.1.1 Sphericity

This subsection is based on [2] and [3].

In this subsection, data is assumed to be *longitudinal*, that is, y_{ij} is now the observation for subject i at the jth timepoint. When conducting ANOVA for repeated measurements, *sphericity* is assumed. Sphericity refers to the variances of all pairwise differences between variables being equal. That is,

$$\mathrm{Var}\left[Y_{ij} - Y_{ik}\right] = \mathrm{Var}\left[Y_{ij}\right] + \mathrm{Var}\left[Y_{ik}\right] - 2\mathrm{Cov}\left[Y_{ij}, Y_{ik}\right] = c, \quad \forall j, k$$

where c is a constant. In other words, sphericity assumes that

$$\mathrm{Var}\left[Y_{ij}\right] = \mathrm{Var}\left[Y_{ik}\right], \quad \forall k$$
$$\mathrm{Cov}\left[Y_{ij}, Y_{ik}\right] = a, \quad \forall k \neq j$$

where a is a constant. This structure is know as *compound symmetry* or *exchangeability* (more on that in Section 4.3). For longitudinal data, sphericity is often unrealistic, as variances tend to increase with time, i.e. $\mathrm{Var}\left[Y_{ij}\right] < \mathrm{Var}\left[Y_{ik}\right]$ when $j < k$.

! As `Wrist` consists of several measurements on patients recorded over time, it is likely, that sphericity does not hold for my data.

The assumption of sphericity must be met to avoid an increase in Type I Errors[3]. Say we had a study in which we observed the weight (kg) of some patients at three seperate timepoints. We could have the following data set, where e.g. T1 stands for the 1st timepoint:

Patient	T1	T2	T3	T1 − T2	T1 − T3	T2 − T3
1	64	65	69	-1	-5	-4
2	71	74	74	-3	-3	0
3	72	72	73	-10	-11	1
4	56	55	57	1	-1	-2
5	62	66	65	-4	-3	1
			Variance:	17.3	14.8	4.7

Let

$$\sigma^2_{T1-T2} = 17.3, \quad \sigma^2_{T1-T3} = 14.8 \quad \text{and} \quad \sigma^2_{T2-T3} = 4.7$$

be the variances for the three differences. The null-hypothesis of sphericity in a case with just three groups (A, B and C) is

$$H_0^{\text{sphericity}} : \sigma^2_{A-B} = \sigma^2_{A-C} = \sigma^2_{B-C},$$

the validity of which is tested via an F-test[4]. In the example with the weight of the patients, we would expect not have sphericity as $\sigma^2_{T1-T2} \neq \sigma^2_{T1-T3} \neq \sigma^2_{T2-T3}$. As mentioned, violations of sphericity lead to an increase in Type I Errors. When testing the model assumptions, e.g. the normality assumption, if any of these assumptions are not met, it could be a result of the variances of the differences between groups not being (sufficiently) equal. Hence, it is important to test for sphericity.

2.2 MANOVA

This section is based on [4] and [5].

In the multivariate case, we still have g independent samples, but now each observation has a 3rd subscript representing one of the p variables. That is, y_{ijk} is the observation on the kth variable from the jth subject in group i. In the univariate case, each observation on a subject were one-dimensional because there were just that one variable. Now each observation on a subject is p-dimensional; one observation per variable, i.e. the observation on subject j in group i is

$$\mathbf{y}_{ij} = \begin{bmatrix} y_{ij1} \\ y_{ij2} \\ \vdots \\ y_{ijp} \end{bmatrix}.$$

To have a balanced design in the multivariate case, we no longer need the group sizes to be equal. Instead, we need all the $\mathbf{y}_{ij}$'s to be p-dimensional.

[3] A Type I Error is when you reject a true hypothesis (false positive). We have that $P(\text{making an error}) = \alpha$, where α is the significance level, often 0.05.

[4] See B.2

! In the MANOVA setup, I have a balanced design, as I have three variables per patient and no missing values.

The assumptions of the MANOVA are that all data in group i have a common mean $\boldsymbol{\mu}_i = [\mu_{i1}, \mu_{i2}, \ldots, \mu_{ip}]$, and a common covariance matrix, $\boldsymbol{\Sigma}$. It is also assumed, that $\mathbf{y}_{ij}$ is independent of $\mathbf{y}_{ik}$ whenever $k \neq j$, and that the data is multivariate normal. The hypotheses to be tested are

$$H_0^{\text{MANOVA}} : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \ldots = \boldsymbol{\mu}_g$$
$$H_A^{\text{MANOVA}} : \mu_{ik} \neq \mu_{jk} \textit{ for at least one } i \neq j \textit{ and at least one variable } k.$$

We reject H_0^{MANOVA} if even just one pair of group means differ on just one variable. Unlike ANOVA, in MANOVA the covariance structure is not restricted to variances being equal and covariances being constant. Hence, MANOVA may actually be used in a repeated measurements setup, where the ANOVA fails due to lack of sphericity.

The sample mean for group i is

$$\bar{\mathbf{y}}_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i}\mathbf{y}_{ij} = \begin{bmatrix} \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij1} \\ \vdots \\ \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ijp} \end{bmatrix} = \begin{bmatrix} \bar{y}_{i\cdot 1} \\ \vdots \\ \bar{y}_{i\cdot p} \end{bmatrix},$$

where $\bar{y}_{i\cdot k}$ is the sample mean for the kth variable in group i. The total sample mean is

$$\bar{\mathbf{y}}_{\cdot\cdot} = \frac{1}{N}\sum_{i=1}^{g}\sum_{j=1}^{n_i}\mathbf{y}_{ij} = \begin{bmatrix} \frac{1}{N}\sum_{i=1}^{g}\sum_{j=1}^{n_i} y_{ij1} \\ \vdots \\ \frac{1}{N}\sum_{i=1}^{g}\sum_{j=1}^{n_i} y_{ijp} \end{bmatrix} = \begin{bmatrix} \bar{y}_{\cdot\cdot 1} \\ \vdots \\ \bar{y}_{\cdot\cdot p} \end{bmatrix},$$

where $\bar{y}_{\cdot\cdot k}$ is the total mean for variable k.

In the multivariate case, we have something called the *total sum of squares and cross product* matrix, $\mathbf{T}$. The total sum of squares is a cross product matrix:

$$\mathbf{T} = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(\mathbf{y}_{ij} - \bar{\mathbf{y}}_{\cdot\cdot})(\mathbf{y}_{ij} - \bar{\mathbf{y}}_{\cdot\cdot})^T$$

This is a $p \times p$-matrix. In $\mathbf{T}$, we are looking at differences in the observations $\mathbf{y}_{ij}$ and the total sample mean. We may split this matrix into a sum of matrices:

$$\begin{aligned}\mathbf{T} &= \sum_{i=1}^{g}\sum_{j=1}^{n_i}(\mathbf{y}_{ij} - \bar{\mathbf{y}}_{\cdot\cdot})(\mathbf{y}_{ij} - \bar{\mathbf{y}}_{\cdot\cdot})^T \\ &\overset{\text{A.2}}{=} \sum_{i=1}^{g}\sum_{j=1}^{n_i}(\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i\cdot})(\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i\cdot})^T + \sum_{i=1}^{g} n_i(\bar{\mathbf{y}}_{i\cdot} - \bar{\mathbf{y}}_{\cdot\cdot})(\bar{\mathbf{y}}_{i\cdot} - \bar{\mathbf{y}}_{\cdot\cdot})^T = \mathbf{E} + \mathbf{H},\end{aligned}$$

where $\mathbf{E}$ is called the *error sum of squares and cross product* and $\mathbf{H}$ is called the *hypothesis sum of squares and cross product*. The matrix $\mathbf{T}$ forms a covariance matrix for total variability; $\mathbf{E}$ is the covariance for the errors (or residuals) and $\mathbf{H}$ is the covariance for the hypothesis. For $k = l$,

$$\mathbf{E}_{k,l} = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(y_{ijk} - \bar{y}_{i\cdot k})(y_{ijk} - \bar{y}_{i\cdot k})$$

and

$$\mathbf{H}_{k,l} = \sum_{i=1}^{g} (\bar{y}_{i\cdot k} - \bar{y}_{\cdot\cdot k})(\bar{y}_{i\cdot k} - \bar{y}_{\cdot\cdot k})$$

measures the within- and between-group variation, respectively, for the kth variable. For $k \neq l$, $\mathbf{E}_{k,l}$ and $\mathbf{H}_{k,l}$ both measures the dependence between variables k and l, but $\mathbf{E}_{k,l}$ does it after taking into account the groups and $\mathbf{H}_{k,l}$ does it across groups. These matrices are of particular interest. In MANOVA, we are essentially testing the hypothesis, that $\mathbf{H} = \mathbf{E}$, which means we would want $\mathbf{H}\mathbf{E}^{-1} \approx \mathbf{I}_{p\times p}$. Notice the similarities between $\mathbf{E}$ and s_w^2 from the univariate case, and between $\mathbf{H}$ and s_b^2. In the univariate case, when $\mathrm{E}\left[s_w^2\right] = \mathrm{E}\left[s_b^2\right]$, we would accept H_0^{ANOVA}. It then makes sense, that in the multivariate case, we want $\mathbf{H}\mathbf{E}^{-1} \approx \mathbf{I}_{p\times p}$.

When a grouping factor has more than two levels, a single test statistic cannot detect all types of departures from H_0^{MANOVA}. Hence, several different test statistics are used. Let λ_i denote the ith eigenvalue[5] of $\mathbf{H}\mathbf{E}^{-1}$. The most popular test statistics for MANOVA are

- *Hotelling-Lawley trace* given by
$$\Lambda_{HL} = \mathrm{tr}\left\{\mathbf{H}\mathbf{E}^{-1}\right\} = \sum_{i=1}^{p} \lambda_i.$$
If $\mathbf{H}$ is large relative to $\mathbf{E}$, then Λ_{HL} will be a large value. If the sum of the eigenvalues is large, then we won't have $\mathbf{H}\mathbf{E}^{-1} \approx \mathbf{I}_{p\times p}$. Thus, H_0^{MANOVA} is rejected when Λ_{HL} is large.

- *Pillai trace* given by
$$\Lambda_P = \mathrm{tr}\left\{\mathbf{H}(\mathbf{H}+\mathbf{E})^{-1}\right\} \overset{\text{A.3}}{=} \sum_{i=1}^{p} \frac{\lambda_i}{1+\lambda_i}.$$
If $\mathbf{H}$ is large relative to $\mathbf{E}$, then Λ_P will be a large value. Thus, H_0^{MANOVA} is rejected when Λ_P is large.

- *Wilk's lambda* given by
$$\Lambda_W = \frac{|\mathbf{E}|}{|\mathbf{E}+\mathbf{H}|} \overset{\text{A.3}}{=} \prod_{i=1}^{p} \frac{1}{1+\lambda_i}.$$
If $\mathbf{H}$ is large relative to $\mathbf{E}$, then the denominator will be large relative to the numerator. Hence, H_0^{MANOVA} is rejected if $\Lambda_W \approx 0$.

- *Roy's greatest root* given by
$$\max\{\lambda_i\}.$$
If $\mathbf{H}$ is large relative to $\mathbf{E}$, then λ_i will be a large value. Thus, H_0^{MANOVA} is rejected when λ_i is large.

We do not want $\mathbf{H}$ to be large relative to $\mathbf{E}$. Once again we can draw parallels to the univariate case. If $\mathrm{E}\left[s_b^2\right] > \mathrm{E}\left[s_w^2\right]$, then we would reject H_0^{ANOVA}. In the same way, if $\mathbf{H} > \mathbf{E}$, we reject H_0^{MANOVA}.

2.3 The ANOVA model

This section is based on [6].

I will end this chapter by explaining how ANOVA is connected to mixed models. In the simple one-way ANOVA, we may express observations as

$$y_{ij} = \beta_i + \epsilon_{ij}, \quad i = 1,\ldots,m, j = 1,\ldots,n_i \tag{2.2}$$

[5] See B.3

with m being the number of subjects, n_i being the number of observations for the ith subject, and the errors ϵ_{ij} being iid. with zero mean and constant variance. The parameters $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_m]$ are fixed, which means that observations for the ith subject are all basically the same (only the small error term, ϵ_{ij}, seperates them), i.e. $\mathrm{E}\left[Y_{ij}\right] = \beta_i, \forall j$. It makes sense then, that the hypothesis to be tested in ANOVA is

$$H_0^{\text{ANOVA}} : \beta_1 = \ldots = \beta_m,$$

It is expressed slightly differently than before, but we are still testing that the means of all subjects/-groups are the same.

Equation (2.2) is called the *ANOVA model*. Gathering all observations in an $N \times 1$ vector, with $N = \sum_{i=1}^{m} n_i$, we can see that the ANOVA model is a special case of a linear regression

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where

$$\mathbf{y} = \begin{bmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ \vdots \\ y_{2n_2} \\ \vdots \\ \vdots \\ y_{m1} \\ \vdots \\ y_{mn_m} \end{bmatrix} \in \mathbb{R}^{N\times 1}, \quad \mathbf{X} = \begin{bmatrix} 1 & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \cdots & \cdots & \vdots \\ 1 & 0 & \cdots & \cdots & 0 \\ 0 & 1 & \cdots & \cdots & 0 \\ \vdots & \vdots & \cdots & \cdots & \vdots \\ 0 & 1 & \cdots & \cdots & 0 \\ \vdots & \vdots & \cdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \cdots & \vdots \\ 0 & 0 & \cdots & \cdots & 1 \\ \vdots & \vdots & \cdots & \cdots & \vdots \\ 0 & 0 & \cdots & \cdots & 1 \end{bmatrix} \in \mathbb{R}^{N\times m}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{bmatrix} \in \mathbb{R}^{m\times 1}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{1n_1} \\ \epsilon_{21} \\ \vdots \\ \epsilon_{2n_2} \\ \vdots \\ \vdots \\ \epsilon_{m1} \\ \vdots \\ \epsilon_{mn_m} \end{bmatrix} \in \mathbb{R}^{N\times 1}.$$

The ANOVA model is in fact a *fixed effect model*. It is called *fixed* because the parameters are fixed. If the parameters, β_i, instead were assumed to be random and iid. with a common variance, then we would have the *variance component model* (VARCOMP model for short):

$$y_{ij} = \beta + u_i + \epsilon_{ij},$$

where $\beta_i = \beta + u_i$, and u_i is called a random effect. The VARCOMP model is a *random effects model*. Combining the ANOVA model and the VARCOMP model results in a *mixed model*, which will be discussed in Chapter 4.

3 Results of using ANOVA and MANOVA on the data

Performing the ANOVA and MANOVA tests in `R` is fairly simple. But before we can get into it, we must make sure, that the assumptions about Gaussian and homoscedastic data holds.

3.1 Testing assumptions

The assumptions will be tested seperately for each timepoint and for each type of movement. That is a total of 27 samples when the data is seperated further into groups. To test the assumptions, I use `superANOVA(test = "assumption")`.

Normality assumption

The normality assumption is assessed both visually and via a test (both implemented in `superANOVA()`). I will start with the visual assessment. Figure 3 shows three of the 27 Q-Q plots and histograms used for detecting normality. The points needs to follow the straight red line and most of the points must lie withing the 95%-confidence interval. The majority of the Q-Q plots resembles the leftmost, which means that the majority of the samples may be considered Gaussian judging from the Q-Q plots. The rightmost Q-Q plot in Figure 3 is the only one of the 27, that looks as if the sample is not Gaussian.

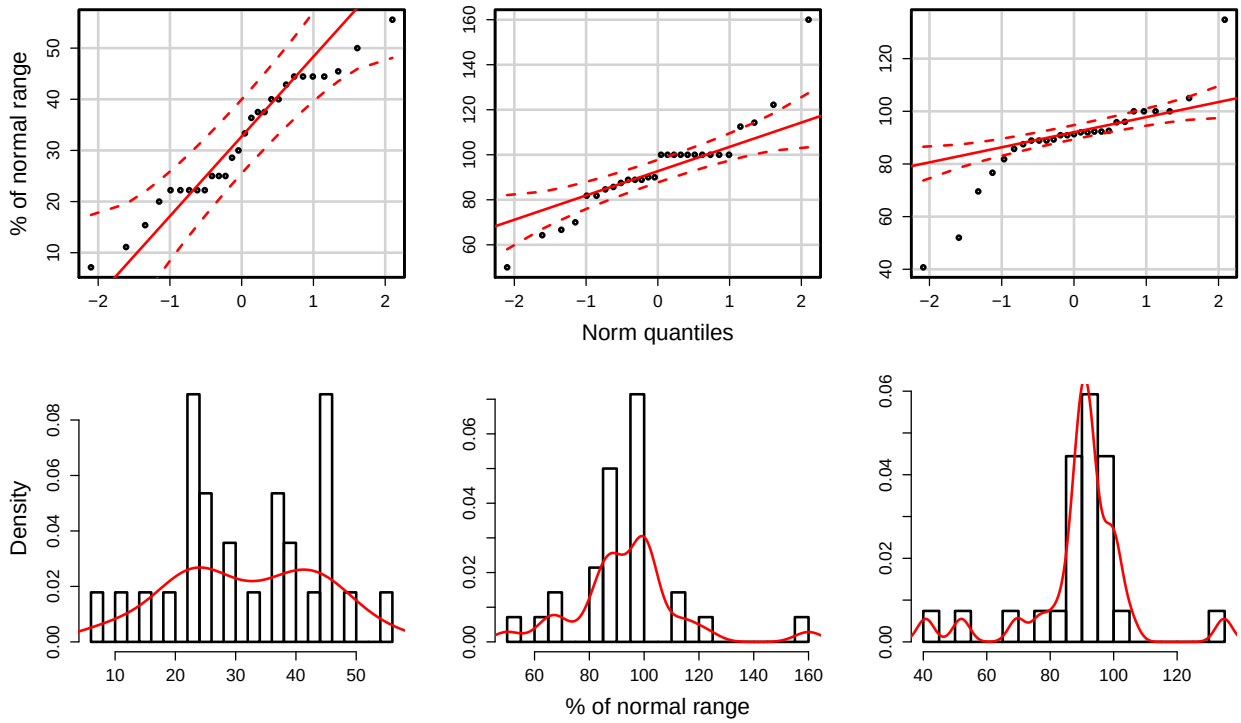

Figure 3: *1st row: Q-Q plots with 95%-confidence intervals (red punctured lines). The first two plots are of the uln/rad movement for group 1 at the 1st and 3rd timepoints, respectively. Last plot is of the ex/flex-movement for group 3 at the 3rd timepoint. 2nd row: Histograms with density curves of the same.*

Wrt. the histograms, the majority resembles either the 2nd or the 3rd. Generally, I find there is no clear connection between the Q-Q plots and the histograms; if the Q-Q plot looks good, the histogram does not, and vice versa.

Next, I will perform a Shapiro-Wilk test[6] with `shapiro.test()`. The null-hypothesis in this test is

$$H_0^{\text{shapiro}} : \textit{sample is from a normally distributed population,}$$

i.e. the sample given as input is Gaussian. All the p-values from performing the Shapiro-Wilk test on all the 27 sample are gathered in the following output:

```
> superANOVA(test = "assumption")$shapiro
      ps.week ps.month     ps.year ur.week ur.month ur.year ef.week ef.month ef.year
gr.1 0.83614   0.00081 3.7422e-06 0.38844   0.79319 0.00981 0.25980   0.54905 0.23157
gr.2 0.15199   0.02526 6.0753e-06 0.67586   0.50036 0.00521 0.18377   0.15872 0.10294
gr.3 0.23995   0.00283 8.1111e-04 0.01153   0.41630 0.00328 0.43686   0.32926 0.00056
```

The Shapiro-Wilk test rejects H_0^{shapiro} for just less than half the samples. Some of these too low p-values could be a result of a Type I Error. The probability of making a Type I Error increases as the number of tests increases:

$$P(\text{making at least 1 error in } m \text{ test}s) = 1-(1-\alpha)^m = 1-(1-0.05)^{27} \approx 0.75.$$

With a probability of 75%, certainly some of the too low p-values could be a result of a Type I Error, but it could also just be that not all the samples are Gaussian. According to several sources, the ANOVA test is not very sensitive to moderate deviations from normality. Hence, we may continue with the ANOVA test. But, just to be on the safe side, I will also perform a Kruskal-Wallis test[7], which is an alternative to the ANOVA test for the situation with non-normal data.

The breaches of normality could be due to breaches of sphericity. I will now test if this could be true. First, I calculated the differences between observations at the three timepoints. Then I calculated the variances of these differences. I do this for `ps`, `ur` and `ef`. The results are shown below:

```
> superANOVA(test = "assumption")$var.diff
      t1-t2  t1-t3  t2-t3
ps   219.26 345.53  71.65
ur   328.36 412.91 316.96
ef   208.57 296.88 240.76
```

The rows in the output are the variances of `ps`, `ur` and `ef` for the differences in observations between timepoints 1 and 2, 1 and 3, and 2 and 3. We see that there is a difference in the variances (more so for `ps`), but are these differences big enough, that we cannot say they are approximately equal and thus satisfy sphericity? To answer this, we use Mauchy's test for sphericity. Here, one can either use `mauchly.test()` or `Anova()`. The set up in `mauchly.test()` is quite complicated and requires both a transformation matrix and a projection matrix. The easier choice is to extract the result of the Mauchly's test in the summary of `Anova()`. In doing so, I get the following p-values:

```
> superANOVA(test = "assumption")$sphericity
             ps      ur      ef
[1,] 4.907e-15 0.29816 0.16826
```

The p-values for `ur` and `ef` are above 0.05, meaning sphericity is satisfied for these two types of movement. For `ps`, the p-value is well below 0.05, meaning sphericity is not satisfied. These results seem

[6] See B.4.
[7] See B.5.

reasonable as there is a greater difference in the variances for ps, according to `superANOVA(test = "assumption")$var.diff`, than there is for the other two types of movement. When sphericity fails, there are ways to make adjustments in order to achieve sphericity. I could try to make adjustments, but to me it seems reasonable that we do not have complete sphericity, as I believe the covariance structure used in ANOVA may not be the most fitting for my data, as it is longitudinal.

Homoscedasticity assumption

The homoscedasticity assumption is first assessed by calculating the variances for each sample:

```
> superANOVA(test = "assumption")$varSample
      ps.week ps.month ps.year ur.week ur.month ur.year ef.week ef.month ef.year
gr.1  365.89   144.25   84.36  152.65   362.75  407.92  127.50   132.36  102.44
gr.2  390.05   154.90   84.44  253.14   406.50  212.50  319.69   443.08  238.74
gr.3  279.78    57.68   18.32  202.52   258.38  327.23  170.19   245.07  280.13
```

To achieve homoscedasticity, the ranges in each column cannot be too large. In the column for `ef.month`, we find the largest range, but is it too big? To answer this, I will perform two tests; a Bartlett test[8], `bartlett.test()`, and a Levene's test[9], `leveneTest()`, which tests the null-hypothesis that

$$H_0^{\text{bartlett}} = H_0^{\text{levene}} : \textit{variances across samples are equal},$$

i.e. $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma^2$ for groups 1, 2, and 3. The following output shows the p-values from these tests:

```
> superANOVA(test = "assumption")$BartLeve
      ps.week ps.month ps.year ur.week ur.month ur.year ef.week ef.month ef.year
Bart  0.67737  0.03212 0.00034 0.43156  0.50058 0.24661 0.04870  0.00922 0.03258
Leve  0.74429  0.15922 0.35857 0.63868  0.64556 0.62689 0.15330  0.00119 0.58765
```

The Levene's test accepts H_0^{levene}, except for `ef.month`. The Bartlett test rejects H_0^{bartlett} in about half the tests. The Bartlett test is sensitive to violations of normality, which may explain those low p-values.

Because H_0^{levene} was rejected for `ef.month`, I will expect ANOVA to also reject H_0^{ANOVA} for this sample. Wrt. the normality assumption, I find the results inconclusive, and will therefore rely on both the ANOVA test and the Kruskal-Wallis test in the following section.

3.2 Analysis using `aov()`

Whether data is balanced or unbalanced plays no role in the setup in R when using the `aov`-function to perform the ANOVA test. The function tests H_0^{ANOVA}, i.e. $\mu_1 = \mu_2 = \mu_3$ for groups 1, 2 and 3. The input in `aov()` is a linear model, for instance

```
aov(ps[which(time == 0)] ~ group[which(time == 0)], data = Wrist).
```

Afterwards, the `summary()` is used to extract the p-value telling us whether or not H_0^{ANOVA} has been rejected. The Kruskal-Wallis test is performed with `kruskal.test()`, which is set up the same way as `aov()`, but we may extract the p-value directly from `kruskal.test()`. In Kruskal-Wallis, we investigate the median of the groups instead of the means, and the null-hypothesis is

$$H_0^{\text{kruskal}} : \textit{the medians of all the groups are equal.}$$

[8] See B.6.
[9] See B.7.

The p-values from `summary(aov())` and `kruskal.test()` are shown in the following output:

```
> superANOVA(test = "anova")
        ps.week ps.month ps.year ef.week ef.month ef.year ur.week ur.month ur.year
aov     0.29704  0.12560 0.52692 0.17276  0.43184 0.44481 0.25237  0.18101 0.65978
kruskal 0.14800  0.21032 0.96337 0.22180  0.32195 0.22175 0.34471  0.24098 0.67808
```

Neither the ANOVA test nor the Kruskal-Wallis test rejects their respective null-hypotheses for any of the types of movement at the three timepoints. But with breaches of the normality assumption (and thus breaches with sphericity), a better test might be the MANOVA test, which I will now move on to.

3.3 Analysis using `manova()`

The function `manova()` is used for performing MANOVA in `R`. With `manova()`, we get *one* p-value per timepoint indicating whether H_0^{MANOVA} should be rejected. This means that we no longer have to test the three types of movemet individually, like in `aov()`; the `manova`-function takes into account that the responses are multivariate. The setup is almost the same as in `aov()`, only now all three types of movement are tested together, for instance

```
manova(cbind(ps[which(time == 0)],
             ef[which(time == 0)],
             ur[which(time == 0)]) ~ group[which(time == 0)], data = Wrist).
```

Through `summary(manova(), test = "Pillai")`, we get the p-value from having used the test statistic of the Pillai trace. We can change `test` to whichever test from Section 2.2, we want. This is all implemented in `superANOVA()`, and the results are shown below:

```
> superANOVA(test = "manova")
                    week   month    year
Hotelling-Lawley 0.42032 0.17653 0.53641
Pillai           0.40507 0.17406 0.52835
Wilks            0.41264 0.17522 0.53232
Roy              0.27634 0.06509 0.24271
```

With no p-values below 0.05, we have no reason to reject H_0^{MANOVA} at any of the timepoints.

3.4 Conclusion

From these initial tests on the data, I conclude that there is no evidence to reject H_0^{main}. These tests are very simple and do not take into account measurement on the same patient possibly being correlated in some way. In the next chapter, Chapter 4, I will present a different way of modelleing the data, that allows for different specfications of the relationship between observations on the same subject.

3.5 Source code: `superANOVA()`

```
superANOVA = function(test, plot = NULL){
  cols = c("ps.week", "ps.month", "ps.year", "ur.week", "ur.month", "ur.year", "ef.week", "ef.
    month", "ef.year")
  TOM = list(Wrist$ps, Wrist$ur, Wrist$ef)

```

```
5   if(test == "assumption"){
6     ## Homoscedastic
7     varSample = matrix(0, ncol = 3, nrow = 9)
8     rownames(varSample) = cols; colnames(varSample) = c("gr.1", "gr.2", "gr.3")
9     n = 1
10    for(i in 1:3){
11      for(j in c(0,7,46)){
12        varSample[n,] = tapply(TOM[[i]][which(Wrist$time == j)], Wrist$group[which(Wrist$time ==
       j)], var)
13        n = n+1
14      }
15    }
16    varSample = t(varSample)
17
18    BartLeve = matrix(0, ncol = 9, nrow = 2); colnames(BartLeve) = colnames(varSample)
19    rownames(BartLeve) = c("Bart", "Leve")
20    nn = 1
21    for(i in 1:3){
22      for(j in c(0,7,46)){
23        BartLeve[1,nn] = bartlett.test(TOM[[i]][which(Wrist$time == j)], Wrist$group[which(Wrist
      $time == j)])$p.value
24        BartLeve[2,nn] = car::leveneTest(TOM[[i]][which(Wrist$time == j)], Wrist$group[which(
      Wrist$time == j)])$"Pr(>F)"[1]
25        nn = nn + 1
26      }
27    }
28
29    ## Normality
30    if(!is.null(plot)){
31      par(mar = c(4.5,4.5,2,1), mfrow = c(3,9))
32      for(i in 1:3){
33        for(j in c(0,7,46)){
34          for(k in 1:3){
35            if(plot == "qq"){
36              car::qqp(TOM[[i]][which(Wrist$time == j & Wrist$group == k)], "norm", main = paste
      ("M",i,"T",j,"G",k), ylab = "%␣of␣normal␣range")
37            }
38            else if(plot == "hist"){
39              hist(TOM[[i]][which(Wrist$time == j & Wrist$group == k)], 20, probability = T,
      main = paste("M",i,"T",j,"G",k), xlab = "%␣of␣normal␣range")
40              lines(density(TOM[[i]][which(Wrist$time == j & Wrist$group == k)]), col = 2)
41            }
42          }
43        }
44      }
45    }
46    shapiro = matrix(0, nrow = 3, ncol = 9); colnames(shapiro) = cols; rownames(shapiro) = c("gr
      .1", "gr.2", "gr.3")
47    s = 1
48    for(i in 1:3){
49      for(j in c(0,7,46)){
50        for(k in 1:3){
51          shapiro[k,s] = shapiro.test(TOM[[i]][which(Wrist$time==j & Wrist$group==k)])$p.value
52        }
53        s = s + 1
54      }
55    }
```

```
    ## Sphericity
    X1 = t(matrix(Wrist$ps, ncol = 83)); X2 = t(matrix(Wrist$ur, ncol = 83)); X3 = t(matrix(
    Wrist$ef, ncol = 83))
    Xps = cbind(X1, X1[,1] - X1[,2], X1[,1] - X1[,3], X1[,2] - X1[,3])
    Xur = cbind(X2, X2[,1] - X2[,2], X2[,1] - X2[,3], X2[,2] - X2[,3])
    Xef = cbind(X3, X3[,1] - X3[,2], X3[,1] - X3[,3], X3[,2] - X3[,3])
    var.diff = rbind(c(var(Xps[,4]), var(Xps[,5]), var(Xps[,6])),
                     c(var(Xur[,4]), var(Xur[,5]), var(Xur[,6])),
                     c(var(Xef[,4]), var(Xef[,5]), var(Xef[,6])))
    colnames(var.diff) = c("t1-t2", "t1-t3", "t2-t3"); rownames(var.diff) = c("ps", "ur", "ef")

    Mps = lm(Xps[,1:3] ~ 1); Mur = lm(Xur[,1:3] ~ 1); Mef = lm(Xef[,1:3] ~ 1)
    design = factor(c("t1", "t2", "t3"))
    options(contrasts=c("contr.sum", "contr.poly"))
    Rps = car::Anova(Mps, idata = data.frame(design), idesign = ~design, type = "III")
    Rur = car::Anova(Mur, idata = data.frame(design), idesign = ~design, type = "III")
    Ref = car::Anova(Mef, idata = data.frame(design), idesign = ~design, type = "III")
    sphericity = matrix(c(summary(Rps, multivariate = F)$sphericity.tests[2],
                          summary(Rur, multivariate = F)$sphericity.tests[2],
                          summary(Ref, multivariate = F)$sphericity.tests[2]),
                        ncol = 3, nrow = 1); colnames(sphericity) = c("ps", "ur", "ef")

    out = list(varSample = varSample, BartLeve = BartLeve, shapiro = shapiro, var.diff = var.
    diff, sphericity = sphericity); out
  }
  else if(test == "anova"){
    p = c(); kw = c()
    m = matrix(0, ncol = 9, nrow = 2); colnames(m) = cols; rownames(m) = c("aov", "kruskal")
    for(i in 2:4){
      for(t in c(0,7,46)){
        mod = aov(Wrist[,i][which(time == t)] ~ group[which(time == t)], data = Wrist)
        p = c(p, summary(mod)[[1]][["Pr(>F)"]][[1]])
        kw = c(kw, kruskal.test(Wrist[,i][which(time == t)] ~ group[which(time == t)], data =
    Wrist)$p.value)
      }
    }
    m[1,] = p; m[2,] = kw; m
  }
  else if(test == "manova"){
    m = matrix(0, ncol = 3, nrow = 4)
    n = 1
    for(t in c(0,7,46)){
      p = c()
      mod = manova(cbind(ps[which(time == t)], ef[which(time == t)], ur[which(time == t)]) ~
    group[which(time == t)], data = Wrist)
      p = c(p, summary(mod, test = "Hotelling-Lawley")$stats[1,6])
      p = c(p, summary(mod, test = "Pillai")$stats[1,6])
      p = c(p, summary(mod, test = "Wilks")$stats[1,6])
      p = c(p, summary(mod, test = "Roy")$stats[1,6])
      m[,n] = p
      n = n + 1
    }
    rownames(m) = c("Hotelling-Lawley", "Pillai", "Wilks", "Roy")
    colnames(m) = c("week", "month", "year")
    print(m)
  }
}
```

4 | Mixed models

Mixed effects models, or just mixed models, are a class of models used for analyzing data with repeated measurements. I will distinguish between *clustered* data and *longitudinal* data. In clustered data, the repeated measurements accur from having several measurements within each cluster. Each cluster can be thought of as forming a group. In longitudinal data, the repeated measurements accur from having several measurements on each subject. Each subject can be thought of as forming a group. This means that, like in Chapter 2, the observation y_{ij} is either the jth measurement in group i (clustered data), or the measurement of subject i at the jth timepoint (longitudinal data). In the mixed model, it is assumed that the conditions within each group are the same, but may vary between groups.

Clustered data can be exemplified as data from a medical experiment, where patients are grouped according to the type of medicine, they are treated with. Each patient has a response to the medicine (e.g. cured/not cured). The data then consists of these responses. The goal may be to determine whether there is a significant difference in the responses from the groups.

> **!** In this sense, one could view my data as clustered data, seeing as patients have been randomized into groups according to the three types of treatment.

Longitudinal data can be exemplified as data from a medical experiment, where a response from each patient in the experiment is recorded, say, once a month. Each patient thus have a vector of responses creating a time series for each patient. The goal could be to determine whether or not there is a difference between the men and the women in the experiment.

> **!** My data can be viewed as longitudinal and will be modelled as such. Each patient have three 3-dimensional vectors of responses for each type of movement. Each patient then have three time series. Figure 4 shows the three time series for a random patient in `Wrist`.

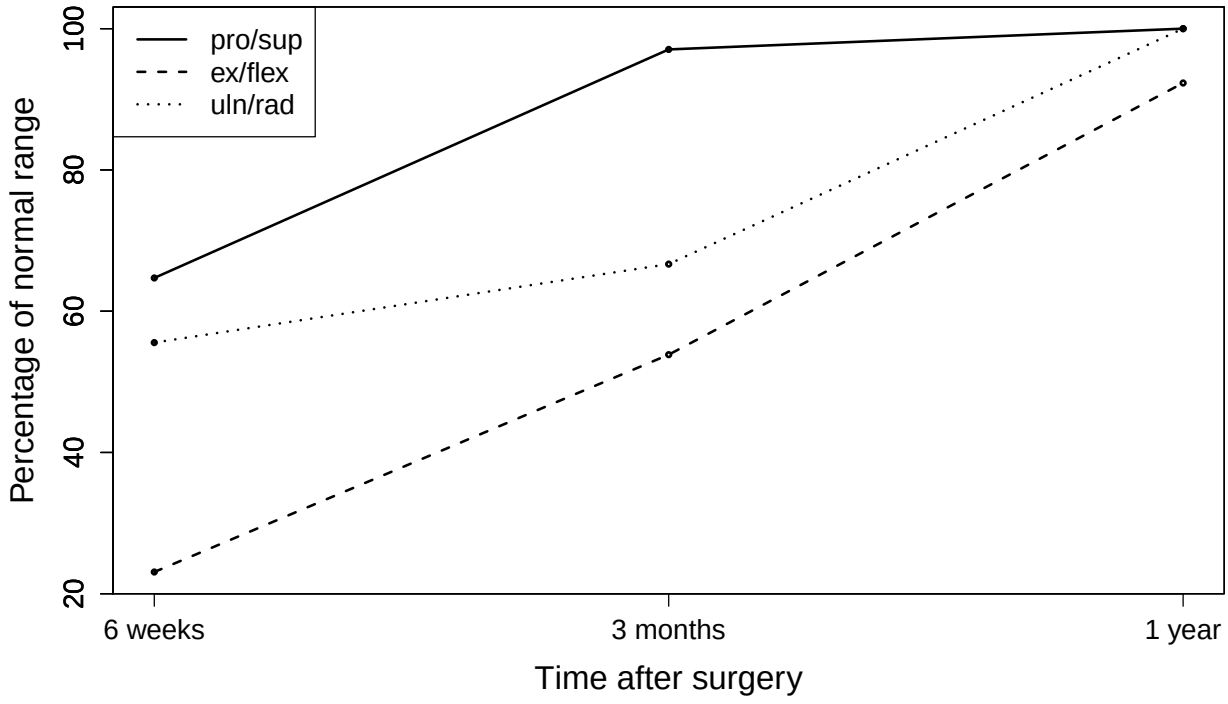

Figure 4: *Time series of the progression of wrist function (measured in degrees) for a random patient in* `Wrist`.

Whether or not the design is balanced depends on the type of data. For clustered data, n_i is the number of observations in the ith group. So if $n_i = n$, $\forall i$, i.e. if all groups are of equal size, the design is balanced. For longitudinal data, n_i is the number of observations for the ith subject. So if $n_i = n$, $\forall i$, i.e. if we have equal number of observations per subject, the design is balanced. It is of interest to have a balanced design, as it simplifies parameter estimation for the model.

> **!** The treatment groups in my data are not of equal size, but since the data is longitudinal data, and there is 3 measurements per patient per type of movement, the design is balanced.

I will now move on to explaining the structure of the mixed model.

4.1 Setting up the mixed model

This section is based on [6].

Why can we not just use a simple classical model

$$y_i = \mu + x_i\beta + \epsilon_i, \quad i = 1, \ldots, N \tag{4.1}$$

with N being the number of subjects, and where the error terms, ϵ_i, are independent and identically distributed with zero mean and constant variance σ^2? Suppose one wanted to test the vocabulary of all students of ages 13 to 16 in a specific school. A test is destributed to these students, and their scores are recorded. Let x_i be the age of student i, and let y_i be the ith students' score on the test. Then (x_i, y_i) is a sample of observations collected on ages and scores, and μ is the overall average of scores. We can then use OLS to estimate μ and β:

$$(\hat{\mu}, \hat{\beta}) = \underset{\mu,\beta\in\mathbb{R}}{\arg\min} \sum_{i=1}^{N} \left(y_i - \mu - x_i\beta\right)^2. \tag{4.2}$$

The model (4.1) assumes that the variation in scores is the same no matter the students' ages. Thus minimizing Equation (4.2) does not take into account the possible within-age correlation. It is possible that the older the student, the higher the score, i.e. there may be a correlation between the scores of students of the same age (see A.4). Although the OLS-estimates are unbiased, accounting for the within-age correlation may give us more efficient estimates of μ and β. Thus it would be more appropriate to assume that each age group has its own age-specific scores:

$$y_{ij} = \mu_i + \mathbf{x}_i^T\boldsymbol{\beta} + \epsilon_{ij}, \quad i = 1,2,3,4, j = 1, \ldots, n_i. \tag{4.3}$$

The notation has changed a bit now. The y_{ij} is the score of the jth student in group i, with a total of n_i students in group i, where each group represents one of the four age groups. The $\mathbf{x}_i$ is a vector that is zero in all entries except for the ith entry indicating that this subject belongs to group i. The $\boldsymbol{\beta}$ is then a vector with entries giving the weight of belonging to each group. The $\boldsymbol{\beta}$ is called a fixed effect as it is the same for all subjects. The μ_i is an age-specific average of scores. The assumption is still that the ϵ_{ij}'s are identically distributed with zero mean and constant variance σ^2. The assumption of the mixed model is that the intercepts, μ_i, are random and can be expressed as

$$\mu_i = \mu + u_i, \tag{4.4}$$

where μ is the overall average of scores (just like in Equation (4.1)) and u_i is a random effect (the deviation of the ith groups' average score from the overall average score). With "μ" we are assuming that the students are representative of 13-16 year olds, but we are still allowing for age-specific variation with

"u_i". Equation (4.3) is actually a special mixed model known as a *random intercept model* (more on that type of model in Section 4.6). Combining Equation (4.3) and (4.4) leads to the linear mixed model

$$y_{ij} = \mu + \mathbf{x}_i^T \boldsymbol{\beta} + u_i + \epsilon_{ij}, \tag{4.5}$$

where u_i and ϵ_{ij} are independent. Let $\text{Var}[U_i] = \sigma_u^2$, $\forall i$, and $\text{Var}\left[\epsilon_{ij}\right] = \sigma^2$. In the mixed model, we then have two sources of variation; σ_u^2, the variation *between* age-groups, and σ^2, the variation *within* age-groups. The parameter $\boldsymbol{\beta}$ is a fixed effect, which means it is constant for all ages. The parameter u_i is an age-specific effect, which means it is the same within the groups, but varies between the groups.

Using Equation (4.5), the model for all scores can be written as

$$\begin{bmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ \vdots \\ y_{2n_2} \\ y_{31} \\ \vdots \\ y_{3n_3} \\ y_{41} \\ \vdots \\ y_{4n_4} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{bmatrix} + \begin{bmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{1n_1} \\ \epsilon_{21} \\ \vdots \\ \epsilon_{2n_2} \\ \epsilon_{31} \\ \vdots \\ \epsilon_{3n_3} \\ \epsilon_{41} \\ \vdots \\ \epsilon_{4n_4} \end{bmatrix}$$

$$\Downarrow$$

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \mathbf{y}_3 \\ \mathbf{y}_4 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \mathbf{X}_3 \\ \mathbf{X}_4 \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \mathbf{1}_{n_1} & & & \\ & \mathbf{1}_{n_2} & & \\ & & \mathbf{1}_{n_3} & \\ & & & \mathbf{1}_{n_4} \end{bmatrix} \mathbf{u} + \begin{bmatrix} \boldsymbol{\epsilon}_1 \\ \boldsymbol{\epsilon}_2 \\ \boldsymbol{\epsilon}_3 \\ \boldsymbol{\epsilon}_4 \end{bmatrix}$$

$$\Downarrow$$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}.$$

Let $\mathbf{V}_i$ denote the covariance matrix of $\mathbf{y}_i$. The more efficient estimates of μ and the β_i's, now gathered in $\boldsymbol{\beta}$, are found as

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^{5\times 1}}{\arg\min} \sum_{i=1}^{4} (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta})^T \mathbf{V}_i^{-1} (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta}),$$

which gives a generalized least squares estimate. These estimates now account for the variation both within and between age-groups.

In the example, I mentioned the words "fixed" and "random" effects. It is important to understand the difference between the two. A fixed effect is a variable whose levels are of particular interest. In other words, we are interested in what effect the levels in the factor have on some outcome. It is explained very well in [7] how to understand the difference between random and fixed effects: Say you want to toast some bread. You want to test the quality of the toasted bread from baking the bread at three different temperatures. Temperature is then of particular interest, and is then a fixed effect. To test these temperatures, we select four slices of bread from each of six batches of bread. The slices of bread represent

a sample from a larger population and thus form a random variable. Hence, the factor describing what batch a slice comes from will be a random factor.

! We are interested in what effect the treatment has on the patients. The groups are then of particular interest, hence, `group` is a fixed effect. The time between observations is not equidistant (the distance is the same for all patients, though). It is possible that these distances between observations affects how observations are correlated for each patient. Hence, also `time` is a fixed effect. It is likely that every patient has some kind of individual influence on the model, something which cannot be controlled (unlike type of medicine, which can be controlled). To include this as a random effect in the model, I must specify for each measurement which patient this measurement comes from. For this, I use `subject` as the random effect.

4.1.1 The linear mixed effects model

This subsection is based on [8].

In the following, I will assume a balanced design, i.e. $n_i = n, \forall i$. Let m be the total number of groups and let $N = \sum_{i=1}^{m} n = mn$. The linear mixed model is defined as follows:

Definition 4.1 (Linear mixed model) The linear mixed model is given as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}, \tag{4.6}$$

where $\mathbf{X}$ and $\mathbf{Z}$ are known matrices, $\boldsymbol{\epsilon} \sim N_N(\mathbf{0}, \mathbf{R})$ and $\mathbf{U} \sim N_{qm}(\mathbf{0}, \mathbf{G})$ are independent, and $\mathbf{G}$ and $\mathbf{R}$ may depend on some unknown variance parameters $\boldsymbol{\varphi}$. The parameters $\boldsymbol{\beta}$ and $\mathbf{u}$ are called the fixed and random effects, respectively.

In Equation (4.6), we have that

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_m \end{bmatrix} \in \mathbb{R}^{N\times 1}, \text{ with } \mathbf{y}_i \in \mathbb{R}^{n\times 1}, \forall i, \qquad \mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_m \end{bmatrix} \in \mathbb{R}^{N\times p}, \text{ with } \mathbf{X}_i \in \mathbb{R}^{n\times p}, \forall i, \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \in \mathbb{R}^{p\times 1},$$

$$\mathbf{Z} = \operatorname{diag}\{\mathbf{Z}_i\}_{i=1,\dots,m} \in \mathbb{R}^{N\times qm}, \text{ with } \mathbf{Z}_i \in \mathbb{R}^{n\times q}, \forall i, \qquad \mathbf{u} = \begin{bmatrix} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_m \end{bmatrix} \in \mathbb{R}^{qm\times 1}, \text{ with } \mathbf{u}_i \in \mathbb{R}^{q\times 1}$$

$$\boldsymbol{\epsilon} = \begin{bmatrix} \boldsymbol{\epsilon}_1 \\ \vdots \\ \boldsymbol{\epsilon}_m \end{bmatrix} \in \mathbb{R}^{N\times 1}, \text{ with } \boldsymbol{\epsilon}_i \in \mathbb{R}^{n\times 1}, \forall i,$$

where q is the number of random effects, and

$$\begin{aligned} &\operatorname{Cov}[\mathbf{U}] = \mathbf{G} = \operatorname{diag}\{\mathbf{G}_0\} \in \mathbb{R}^{qm\times qm}, \text{ with } \mathbf{G}_0 = \operatorname{Cov}[\mathbf{U}_i], \forall i, \\ &\operatorname{Cov}[\boldsymbol{\epsilon}] = \mathbf{R} = \operatorname{diag}\{\mathbf{R}_i\}_{i=1,\dots,m} \in \mathbb{R}^{N\times N}. \end{aligned}$$

The model (4.6) has marginal distribution

$$\mathbf{Y} \overset{\text{A.8}}{\sim} N_N\left(\mathbf{X}\boldsymbol{\beta}, \mathbf{Z}\mathbf{G}\mathbf{Z}^T + \mathbf{R}\right) = N_N\left(\mathbf{X}\boldsymbol{\beta}, \mathbf{V}\right),$$

where $\mathbf{V} = \mathbf{ZGZ}^T + \mathbf{R}$, and where we may write $\mathbf{V} = \mathbf{V}(\boldsymbol{\varphi})$ when the covariance depends on $\boldsymbol{\varphi}$. Notice, in the mixed model, how the fixed effects are used to model the mean of $\mathbf{Y}$, while the random effects are used to shape to covariance of $\mathbf{Y}$. The model can be written as a so-called *two level heirarchical model*

$$\mathbf{Y} \mid \mathbf{U} = \mathbf{u} \sim N_N\left(\mathbf{X}\boldsymbol{\beta} + \mathbf{Zu}, \mathbf{R}\right) \tag{4.7}$$

$$\mathbf{U} \sim N_{qm}\left(\mathbf{0}, \mathbf{G}\right). \tag{4.8}$$

Letting $\boldsymbol{\mu}(\boldsymbol{\beta}, \mathbf{u}) = \mathrm{E}\left[\mathbf{Y} \mid \mathbf{U} = \mathbf{u}\right] = \mathbf{X}\boldsymbol{\beta} + \mathbf{Zu}$, we may term $\boldsymbol{\mu}(\boldsymbol{\beta}, \mathbf{u})$ the *mean function* of $\mathbf{y}$ conditioned on the outcome of the random effect, and we may write the model (4.6) as $\mathbf{y} = \boldsymbol{\mu}(\boldsymbol{\beta}, \mathbf{u}) + \boldsymbol{\epsilon}$. The model (4.6) is called the *linear* mixed effects model (LMM for short) as $\boldsymbol{\mu}(\boldsymbol{\beta}, \mathbf{u})$ is linear in $\boldsymbol{\beta}$. It is also possible to have a non-linear mixed model

$$g\left(\boldsymbol{\mu}(\boldsymbol{\beta}, \mathbf{u})\right) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Zu},$$

where the *link function*, $g(\cdot)$, is a function such that $g\left(\boldsymbol{\mu}(\boldsymbol{\beta}, \mathbf{u})\right)$ is linear in $\boldsymbol{\beta}$. The link function specifies how the mean depends on the covariates. Every type of distribution from the exponential family has a certain link function. For Gaussian data, $g(\cdot)$ is just the identity. I shall work only with the linear model.

4.2 Estimation and prediction of effects

This section and subsections are based on [8] and [9].

Unknown parameters such as $\boldsymbol{\beta}$ must be estimated. Seeing as $\mathbf{u}$ is random, and we predict rather than estimate random variables and vectors, $\mathbf{u}$ must be predicted, thus explaining the name of this section.

From the marginal distribution, we know that $\mathbf{Y} \sim N_N\left(\mathbf{X}\boldsymbol{\beta}, \mathbf{V}(\boldsymbol{\varphi})\right)$. The likelihood and log-likelihood functions are then

$$L(\boldsymbol{\beta}, \boldsymbol{\varphi}; \mathbf{y}) = \frac{1}{\sqrt{2\pi}^N} |\mathbf{V}(\boldsymbol{\varphi})|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{V}(\boldsymbol{\varphi})^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right) \tag{4.9}$$

$$\ell(\boldsymbol{\beta}, \boldsymbol{\varphi}; \mathbf{y}) \equiv -\frac{1}{2}\log(|\mathbf{V}(\boldsymbol{\varphi})|) - \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{V}(\boldsymbol{\varphi})^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \tag{4.10}$$

Maximizing $\ell(\boldsymbol{\beta}, \boldsymbol{\varphi}; \mathbf{y})$ wrt. the parameters is dependent on whether $\mathbf{V}$ is known. We then have two procedures for estimating $\boldsymbol{\beta}$ and predicting $\mathbf{u}$; a rather simple one for when $\mathbf{V}$ is known (Subsection 4.2.1), and a slightly more comprehensive one for when $\mathbf{V} = \mathbf{V}(\boldsymbol{\varphi})$ (Subsection 4.2.2). In Section 4.3, I will show a little more precisely how these expressions come to look once a specific structure for $\mathbf{V}$ is given.

4.2.1 Known covariance

Expression for $\hat{\boldsymbol{\beta}}$

Assuming $\mathbf{V}$ is known, $\boldsymbol{\beta}$ is found by solving $\frac{\partial}{\partial \boldsymbol{\beta}}\ell(\boldsymbol{\beta}; \mathbf{y}) = \mathbf{0}$, which gives the MLE

$$\hat{\boldsymbol{\beta}} \overset{\text{A.5}}{=} \left(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}. \tag{4.11}$$

Expression for $\hat{\mathbf{u}}$

Suppose we want to predict a random variable X as much as possible by a constant c. Then what we want to do is to minimize $\mathrm{E}\left[(X - c)^2\right]$:

$$\mathrm{E}\left[(X - c)^2\right] \overset{\text{A.6}}{=} \mathrm{Var}\left[X\right] + (\mathrm{E}\left[X\right] - c)^2.$$

This expression is clearly minimized when $c = \mathrm{E}[X]$, and thus $\mathrm{E}[X]$ is the best predictor for X. If we observe a random variable Z, then we can improve our prediction of X by conditioning on Z. That is, in the same way that $\mathrm{E}[X]$ was the best predictor for X when we did not know Z, $\mathrm{E}[X \mid Z = z]$ is now the best predictor of the unknown X, when Z is observed. And so, having observed $\mathbf{y}$, $\mathbf{u}$ has a conditional multivariate distribution[10] and is predicted by

$$\hat{\mathbf{u}} = \mathrm{E}\left[\mathbf{U} \mid \mathbf{Y} = \mathbf{y}\right] \;=\; \mathbf{0} + \mathbf{G}\mathbf{Z}^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \;=\; \mathbf{G}\mathbf{Z}^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \tag{4.12}$$

as

$$\begin{pmatrix}\mathbf{Y}\\ \mathbf{U}\end{pmatrix} \sim N_{N+qm}\left(\begin{bmatrix}\mathbf{X}\boldsymbol{\beta}\\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}\mathbf{V} & \mathbf{Z}\mathbf{G}\\ \mathbf{G}\mathbf{Z}^T & \mathbf{G}\end{bmatrix}\right).$$

Equation (4.12) is called the *best linear unbiased predictor* (BLUP for short) of $\mathbf{u}$, where $\boldsymbol{\beta}$ is replaced by $\hat{\boldsymbol{\beta}}$. It is called *linear* as it is a linear function of $\mathbf{y}$. It is unbiased by the Law of Total Expectation:

$$\mathrm{E}\left[\hat{\mathbf{U}}\right] = \mathrm{E}\left[\mathrm{E}\left[\mathbf{U} \mid \mathbf{Y} = \mathbf{y}\right]\right] \;=\; \mathrm{E}[\mathbf{U}].$$

The BLUP replaces the random effects, $\mathbf{u}$, by their conditional means, $\hat{\mathbf{u}}$, given the data. We can then make predictions on $\mathbf{y}$ by

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\mathbf{u}}.$$

4.2.2 Unknown covariance

Expression for $\hat{\boldsymbol{\beta}}$

Assume $\mathbf{R}$ and $\mathbf{G}$ are known up to the unknown $\boldsymbol{\varphi}$ in the marginal model, where $\mathbf{V}(\boldsymbol{\varphi}) = \mathbf{Z}\mathbf{G}(\boldsymbol{\varphi})\mathbf{Z}^T + \mathbf{R}(\boldsymbol{\varphi})$. We must now estimate $\boldsymbol{\beta}$ and $\boldsymbol{\varphi}$.

Maximizing $\ell(\boldsymbol{\beta}, \boldsymbol{\varphi}; \mathbf{y})$ (Equation (4.10)) wrt. $\boldsymbol{\beta}$ for fixed $\boldsymbol{\varphi}$ gives

$$\hat{\boldsymbol{\beta}}(\boldsymbol{\varphi}) = \left(\mathbf{X}^T\mathbf{V}(\boldsymbol{\varphi})^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{V}(\boldsymbol{\varphi})^{-1}\mathbf{y}. \tag{4.13}$$

The estimate is dependent on $\boldsymbol{\varphi}$, which must be found. This is done by profile likelihood. The profile log-likelihood for $\boldsymbol{\varphi}$ is

$$\ell_{\mathrm{p}}(\boldsymbol{\varphi}) = \ell\left(\hat{\boldsymbol{\beta}}(\boldsymbol{\varphi}), \boldsymbol{\varphi}; \mathbf{y}\right) \;\equiv\; -\frac{1}{2}\log\left(|\mathbf{V}(\boldsymbol{\varphi})|\right) - \frac{1}{2}\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}(\boldsymbol{\varphi})\right)^T\mathbf{V}(\boldsymbol{\varphi})^{-1}\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}(\boldsymbol{\varphi})\right). \tag{4.14}$$

Solving $\frac{\partial}{\partial\boldsymbol{\varphi}}\ell_{\mathrm{p}}(\boldsymbol{\varphi}) = \mathbf{0}$ gives the maximum likelihood estimate $\hat{\boldsymbol{\varphi}}_{\mathrm{ML}}$, which is a biased estimate. To get an unbiased estimate, we can instead use the *restricted maximum likelihood* method (REML-method for short). Here, we will need the marginal log-likelihood to estimate $\boldsymbol{\varphi}$:

$$\ell_{\mathrm{R}}(\boldsymbol{\varphi}) = \log\left(\int L(\boldsymbol{\beta}, \boldsymbol{\varphi}; \mathbf{y})\, d\boldsymbol{\beta}\right) \overset{\text{A.9}}{=} \ell_{\mathrm{p}}(\boldsymbol{\varphi}) - \frac{1}{2}\log\left(|\mathbf{X}^T\mathbf{V}(\boldsymbol{\varphi})^{-1}\mathbf{X}|\right) \tag{4.15}$$

Maximizing $\ell_{\mathrm{R}}(\boldsymbol{\varphi})$ wrt. $\boldsymbol{\varphi}$ gives the *restricted maximum likelihood estimate*, $\hat{\boldsymbol{\varphi}}_{\mathrm{R}}$.

Expression for $\hat{\mathbf{u}}$

As in the case where $\mathbf{V}$ is known, we can predict $\mathbf{u}$ by Equation (4.12). Only now we have to insert $\mathbf{G}(\hat{\boldsymbol{\varphi}}_{\mathrm{R}})$, $\mathbf{V}(\hat{\boldsymbol{\varphi}}_{\mathrm{R}})$ and $\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\varphi}}_{\mathrm{R}})$:

$$\hat{\mathbf{u}}(\hat{\boldsymbol{\varphi}}_{\mathrm{R}}) = \mathbf{G}(\hat{\boldsymbol{\varphi}}_{\mathrm{R}})\mathbf{Z}^T\mathbf{V}(\hat{\boldsymbol{\varphi}}_{\mathrm{R}})^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\varphi}}_{\mathrm{R}})).$$

[10] See B.8.

Simultaneous estimation

Notice that $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{u}}$ are dependent on $\hat{\boldsymbol{\varphi}}_{\mathrm{R}}$, which is dependent on $\hat{\boldsymbol{\beta}}$. This complicates parameter estimation. The solution is to do simultaneous estimation. The joint density for $(\mathbf{Y}, \mathbf{U})$ is a so-called *hierarchical likelihood* comprised of the pdf's of the two level heirarchical model, Equations (4.7) and (4.8):

$$\begin{aligned} f(\mathbf{y}, \mathbf{u}; \boldsymbol{\beta}, \boldsymbol{\varphi}) &= f_{\mathbf{y}|\mathbf{u}}(\mathbf{y}; \boldsymbol{\beta}, \boldsymbol{\varphi})\, f_{\mathbf{u}}(\mathbf{u}; \boldsymbol{\varphi}) \\ &= \left(\frac{1}{\sqrt{2\pi}^N} |\mathbf{R}(\boldsymbol{\varphi})|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})^T \mathbf{R}(\boldsymbol{\varphi})^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}) \right) \right) \\ &\quad \left(\frac{1}{\sqrt{2\pi}^{qm}} |\mathbf{G}(\boldsymbol{\varphi})|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} \mathbf{u}^T \mathbf{G}(\boldsymbol{\varphi})^{-1} \mathbf{u} \right) \right) \end{aligned}$$

The joint log-likelihood is thus

$$\ell(\boldsymbol{\beta}, \boldsymbol{\varphi}; \mathbf{u}, \mathbf{y}) \equiv -\frac{1}{2} \Big(\log(|\mathbf{R}(\boldsymbol{\varphi})|) - (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})^T \mathbf{R}(\boldsymbol{\varphi})^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}) \log(|\mathbf{G}(\boldsymbol{\varphi})|) - \mathbf{u}^T \mathbf{G}(\boldsymbol{\varphi})^{-1} \mathbf{u} \Big) \tag{4.16}$$

and the score functions wrt. $\boldsymbol{\beta}$ and $\mathbf{u}$ are

$$S_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\varphi}; \mathbf{u}, \mathbf{y}) = \frac{\partial}{\partial \boldsymbol{\beta}} \ell(\boldsymbol{\beta}, \boldsymbol{\varphi}; \mathbf{u}, \mathbf{y}) \;=\; \mathbf{X}^T \mathbf{R}(\boldsymbol{\varphi})^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}) \tag{4.17}$$

$$S_{\mathbf{u}}(\boldsymbol{\beta}, \boldsymbol{\varphi}; \mathbf{u}, \mathbf{y}) = \frac{\partial}{\partial \mathbf{u}} \ell(\boldsymbol{\beta}, \boldsymbol{\varphi}; \mathbf{u}, \mathbf{y}) \overset{\text{A.7}}{=} \mathbf{Z}^T \mathbf{R}(\boldsymbol{\varphi})^{-1} \big(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u} \big) - \mathbf{G}(\boldsymbol{\varphi})^{-1} \mathbf{u}. \tag{4.18}$$

Putting both Equation (4.18) and (4.17) equal to $\mathbf{0}$, we obtain the *mixed model equations* (MME for short):

$$\begin{bmatrix} \mathbf{X}^T \mathbf{R}(\boldsymbol{\varphi})^{-1} \mathbf{X} & \mathbf{X}^T \mathbf{R}(\boldsymbol{\varphi})^{-1} \mathbf{Z} \\ \mathbf{Z}^T \mathbf{R}(\boldsymbol{\varphi})^{-1} \mathbf{X} & \mathbf{Z}^T \mathbf{R}(\boldsymbol{\varphi})^{-1} \mathbf{Z} + \mathbf{G}(\boldsymbol{\varphi})^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^T \mathbf{R}(\boldsymbol{\varphi})^{-1} \mathbf{y} \\ \mathbf{Z}^T \mathbf{R}(\boldsymbol{\varphi})^{-1} \mathbf{y} \end{bmatrix}. \tag{4.19}$$

Solving the MME can be set up as an algorithm:

Algorithm 4.2 (Algorithm for solving the MME)

i. Initialize $\boldsymbol{\beta}$, e.g. by $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$.

ii. Using $\hat{\boldsymbol{\beta}}$, find $\hat{\boldsymbol{\varphi}}_{\mathrm{R}}$ by maximizing $\ell_{\mathrm{R}}(\boldsymbol{\varphi})$.

iii. Calculate an adjusted observation $\mathbf{y}(\hat{\boldsymbol{\varphi}}_{\mathrm{R}})^{\mathrm{adj}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\varphi}}_{\mathrm{R}})$, and predict $\mathbf{u}$ from a random effects model $\mathbf{y}(\hat{\boldsymbol{\varphi}}_{\mathrm{R}})^{\mathrm{adj}} = \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}$ by

$$\big(\mathbf{Z}^T \mathbf{R}(\hat{\boldsymbol{\varphi}}_{\mathrm{R}})^{-1} \mathbf{Z} + \mathbf{G}(\hat{\boldsymbol{\varphi}}_{\mathrm{R}})^{-1} \big) \mathbf{u} = \mathbf{Z}^T \mathbf{R}(\hat{\boldsymbol{\varphi}}_{\mathrm{R}})^{-1} \mathbf{y}(\hat{\boldsymbol{\varphi}}_{\mathrm{R}})^{\mathrm{adj}}.$$

iv. Recalculate the adjusted observation $\mathbf{y}(\hat{\boldsymbol{\varphi}}_{\mathrm{R}})^{\mathrm{adj}} = \mathbf{y} - \mathbf{Z}\hat{\mathbf{u}}(\hat{\boldsymbol{\varphi}}_{\mathrm{R}})$, and estimate $\boldsymbol{\beta}$ from a fixed effects model $\mathbf{y}(\hat{\boldsymbol{\varphi}}_{\mathrm{R}})^{\mathrm{adj}} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ by

$$\mathbf{X}^T \mathbf{R}(\hat{\boldsymbol{\varphi}}_{\mathrm{R}})^{-1} \mathbf{X}\boldsymbol{\beta} = \mathbf{X}^T \mathbf{R}(\hat{\boldsymbol{\varphi}}_{\mathrm{R}})^{-1} \mathbf{y}(\hat{\boldsymbol{\varphi}}_{\mathrm{R}})^{\mathrm{adj}}.$$

Repeat step ii. and iv. until convergence.

Information about the effect of the choice of initial $\hat{\boldsymbol{\beta}}$, and whether Algorithm 4.2 is garanteed to converge, is scarce. I have included a word on this in my discussion, Chapter 9.

The MLE, $\hat{\boldsymbol{\beta}}$, is asymptotically $N_p\left(\boldsymbol{\beta}, \mathrm{I}(\boldsymbol{\beta})^{-1}\right)$-distributed, where $\mathrm{I}(\boldsymbol{\beta})$ is the Fisher information matrix, i.e. $\hat{\boldsymbol{\beta}}$ is both unbiased (because $\mathrm{E}\left[\hat{\boldsymbol{\beta}}\right] = \boldsymbol{\beta}$) and efficient (because $\mathrm{Cov}\left[\hat{\boldsymbol{\beta}}\right]$ is equal to the Cramer-Rao lower bound, $\mathrm{I}(\boldsymbol{\beta})^{-1}$), and it is a consistent estimator of $\boldsymbol{\beta}$ meaning $\hat{\boldsymbol{\beta}} \xrightarrow{N\to\infty} \boldsymbol{\beta}$. This holds no matter the structure of $\mathbf{V}$ (see Section 4.3 for structures of $\mathbf{V}$).

We find $\hat{\boldsymbol{\varphi}}_{\mathrm{R}}$ by maximizing $\ell_{\mathrm{R}}(\boldsymbol{\varphi})$ wrt. $\boldsymbol{\varphi}$. But the assumptions made on the relationship between subjects and within each subject, affects what $\boldsymbol{\varphi}$ consists of. This, of course, also affects the expressions for $\hat{\boldsymbol{\beta}}$, $\hat{\mathbf{u}}$, $\mathbf{G}$, $\mathbf{R}$, and $\mathbf{V}$. In Section 4.3, I go through this in more details.

4.3 Covariance structure

This section is based on [7].

As just mentioned in the prior section, the assumptions made on the relationship between and within subjects affects the expressions for $\hat{\boldsymbol{\beta}}$, $\hat{\mathbf{u}}$, $\mathbf{G}$, $\mathbf{R}$, and $\mathbf{V}$. One assumption could be that no subjects are correlated. Another could be that there is no correlation between pairs of observation within a subject. If the assumption is, that observations within a subject are correlated, then how exactly shall we assume they are correlated? All these assumptions lead to different expressions for $\hat{\boldsymbol{\beta}}$, $\hat{\mathbf{u}}$, $\mathbf{G}$, $\mathbf{R}$, and $\mathbf{V}$. In this section, I will shortly present some of the most popular choices for covariance structure of $\mathbf{Y}$. I will proceed to calculate the expressions for $\hat{\boldsymbol{\beta}}$, $\hat{\mathbf{u}}$, $\mathbf{G}$, $\mathbf{R}$, and $\mathbf{V}$ for one of these choices (Subsection 4.3.1), and afterwards, I will show in detail how to estimate $\boldsymbol{\varphi}$ (Subsection 4.3.2). For ease of notation, I shall omit "$(\boldsymbol{\varphi})$" in the following.

The structure of $\mathrm{Cov}[\mathbf{Y}]$ can be expressed in several ways. Some choices[11] are

- *independence*: Here it is assumed that $\mathrm{Corr}\left[Y_{ij}, Y_{il}\right] = 0$, whenever $j \neq l$. That is, observations on the same subject are all uncorrelated,
- *exchangeability*: Here it is assumed that $\mathrm{Corr}\left[Y_{ij}, Y_{il}\right] = \rho$, $\forall\, j \neq l$ and $\forall i$ for some constant ρ. That is, the correlation between observations for a subject are the same no matter the time that has passed between observations,
- *autoregression*: Here it is assumed that $\mathrm{Corr}\left[Y_{ij}, Y_{il}\right] = \rho^{|j-l|}$, $|\rho| < 1$. That is, the correlation decreases as more time passes between observation, and
- *unstructured*: Here it is assumed that all observations for a subject are indeed correlated, but it is not specified exactly how they are correlated, meaning that $\mathrm{Corr}\left[Y_{ij}, Y_{il}\right] = \rho_{j,l}$. In the other cases, there is a specific structure to the correlation. That does not apply in this latter case, hence the covariance structure is termed *unstructured*.

For longitudinal data, the first-order autoregressive covariance structure is very popular, and the reason for this is obvious; it is certainly conceivable that $\mathrm{Corr}\left[Y_{i1}, Y_{i2}\right] > \mathrm{Corr}\left[Y_{i1}, Y_{in_i}\right]$, for instance.

I will focus on the case probably most relevant to my data; autoregressive structure. Before doing so, I am making a lot of assumptions, which will simplify my calculations. These assumptions may not be entirely relevant to my data, but as a way of better understanding the workings behind estimation of $\boldsymbol{\varphi}$, I find these assumptions permissible. I will be using a lot of diagonal block-matrices, where notation using direct product[12] (or Kronecker product) comes in very handy. The assumptions are:

[11] Information about the covariance structures is based on [2].
[12] See B.9.

- There is only one random factor, meaning that $\mathbf{Z} = \mathbf{I}_{m\times m} \otimes \mathbf{1}_n$.
- There is also only one fixed factor, meaning that $\mathbf{X} = \mathbf{1}_m \otimes \mathbf{I}_{n\times n}$.
- The $\mathbf{Y} \mid \mathbf{U}_i = \mathbf{u}_i$'s are independent and have the same covariance matrix, $\mathbf{R}_0$, meaning that $\mathbf{R} = \mathbf{I}_{m\times m} \otimes \mathbf{R}_0$.
- The $\mathbf{y}_i$'s all have the same variance, $\mathbf{V}_0$, meaning that $\mathbf{V} = \mathbf{I}_{m\times m} \otimes \mathbf{V}_0$.
- There is no correlation between subjects. Let $\text{Corr}\left[U_i, U_j\right] = \rho_u = 0$, $\forall\, i \neq j$, and let $\sigma_u^2 = \text{Var}\left[U_i\right]$, $\forall\, i$. Thus $\mathbf{G} = \text{Cov}\left[\mathbf{U}\right]$ has σ_u^2 on the diagonal, and $\sigma_u^2\rho_u$ on the off-diagonal. With $\rho_u = 0$, we have

$$\mathbf{G} = \sigma_u^2\left((1-\rho_u)\mathbf{I}_{m\times m} + \rho_u \mathbf{J}_{m\times m}\right) \;=\; \sigma_u^2\left((1-0)\mathbf{I}_{m\times m} + 0\cdot\mathbf{J}_{m\times m}\right) \;=\; \sigma_u^2\mathbf{I}_{m\times m}, \tag{4.20}$$

  where $\mathbf{J}_{m\times m}$ denotes an $m \times m$ matrix consisting of all 1s.

With all these assumptions, the covariance of $\mathbf{Y}$ is then

$$\begin{aligned}\mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}^T + \mathbf{R} \;&=\; (\mathbf{I}_{m\times m}\otimes\mathbf{1}_n)(\mathbf{G}\otimes 1)(\mathbf{I}_{m\times m}\otimes\mathbf{1}_n^T) + \mathbf{I}_{m\times m}\otimes\mathbf{R}_0 \;=\; \mathbf{G}\otimes\mathbf{J}_{n\times n} + \mathbf{I}_{m\times m}\otimes\mathbf{R}_0 \\ &\overset{(4.20)}{=} \sigma_u^2\mathbf{I}_{m\times m}\otimes\mathbf{J}_{n\times n} + \mathbf{I}_{m\times m}\otimes\mathbf{R}_0 \;=\; \mathbf{I}_{m\times m}\otimes\left(\sigma_u^2\mathbf{J}_{n\times n} + \mathbf{R}_0\right).\end{aligned} \tag{4.21}$$

The prediction of $\mathbf{u}$ thus becomes

$$\hat{\mathbf{u}} \overset{(4.12)}{=} \mathbf{G}\mathbf{Z}^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \overset{\text{A.10}}{=} \frac{1}{r_0 + \frac{1}{\sigma_u^2}}\left(\mathbf{I}_{m\times m}\otimes\mathbf{1}_n^T\mathbf{R}_0^{-1}\right)(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \tag{4.22}$$

with r_0 denoting the sum of all entries in $\mathbf{R}_0^{-1}$, that is, $r_0 = \mathbf{1}_n^T\mathbf{R}_0^{-1}\mathbf{1}_n$.

This gives

$$\hat{u}_i = \frac{1}{r_0 + \frac{1}{\sigma_u^2}}\mathbf{1}_n^T\mathbf{R}_0^{-1}(\mathbf{y}_i - \hat{\boldsymbol{\beta}}). \tag{4.23}$$

If the number of observations on each subject is not very large, inverting $\mathbf{R}_0$ should cause no grief.

And finally, with these assumption $\hat{\boldsymbol{\beta}}$ is actually independent of $\boldsymbol{\varphi}$:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y} \overset{\text{A.11}}{=} \begin{bmatrix}\bar{y}_{\cdot 1}\\ \vdots \\ \bar{y}_{\cdot n}\end{bmatrix}. \tag{4.24}$$

We can now proceed concentrating only on $\hat{\mathbf{u}}$.

4.3.1 Expressing $\hat{u}_i$ using autoregressive structure

In this subsection, I will calculate the expression for $\hat{u}_i$ when assuming autoregressive structure. Here, the assumption is that the correlation between observations for a subject decreases as more time passes between them. Thus

$$\mathbf{R}_0 = \sigma^2\begin{bmatrix} 1 & \rho & \rho^2 & \dots & \rho^{n-1} \\ \rho & 1 & \rho & \dots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \dots & \rho^{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \dots & 1 \end{bmatrix} \;=\; \sigma^2\mathbf{A}, \qquad |\rho| < 1$$

with

$$\mathbf{R}_0^{-1} = \frac{1}{\sigma^2}\mathbf{A}^{-1} = \frac{1}{\sigma^2(1-\rho^2)}\begin{bmatrix} 1 & -\rho & 0 & \dots & 0 & 0 \\ -\rho & 1+\rho^2 & -\rho & \dots & 0 & 0 \\ 0 & -\rho & 1+\rho^2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1+\rho^2 & -\rho \\ 0 & 0 & 0 & \dots & -\rho & 1 \end{bmatrix}.$$

This gives

$$\begin{aligned} r_0 &= \frac{1}{\sigma^2(1-\rho^2)}\left(2(1-\rho)+(n-2)(\rho^2-2\rho+1)\right) = \frac{1}{\sigma^2(1-\rho^2)}\left(2(1-\rho)+(n-2)(1-\rho)^2\right) \\ &= \frac{(1-\rho)(2+(n-2)(1-\rho))}{\sigma^2(1-\rho^2)}. \end{aligned}$$

and

$$\mathbf{1}_n^T\mathbf{R}_0^{-1}(\mathbf{y}_i-\hat{\boldsymbol{\beta}}) \overset{\text{A.12}}{=} \frac{1-\rho}{\sigma^2(1-\rho^2)}\Big((1-\rho)n(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})+\rho(y_{i1}-\bar{y}_{\cdot 1}+y_{in}-\bar{y}_{\cdot n})\Big),$$

which gives

$$\begin{aligned} \hat{u}_i &\overset{(4.23)}{=} \frac{1}{\frac{(1-\rho)(2+(n-2)(1-\rho))}{\sigma^2(1-\rho^2)}+\frac{1}{\sigma_u^2}}\cdot\frac{1-\rho}{\sigma^2(1-\rho^2)}\Big((1-\rho)n(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})+\rho(y_{i1}-\bar{y}_{\cdot 1}+y_{in}-\bar{y}_{\cdot n})\Big) \\ &\overset{\text{A.13}}{=} \frac{\sigma_u^2\Big((1-\rho)n(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})+\rho(y_{i1}-\bar{y}_{\cdot 1}+y_{in}-\bar{y}_{\cdot n})\Big)}{\sigma^2(1+\rho)+\sigma_u^2(n-\rho(n-2))}. \end{aligned}$$

Note that $\rho(y_{i1}-\bar{y}_{\cdot 1}+y_{in}-\bar{y}_{\cdot n})$ represents end-effects as the first and last diagonal elements in $\mathbf{R}_0^{-1}$ are different from the other diagonal elements. We now have an expression for $\hat{u}_i$. All we need now, is to plug in the estimates for the unknown parameters σ^2, σ_u^2 and ρ.

4.3.2 Estimating σ^2, σ_u^2 and ρ

This section is based on [7].

Retrieving the log-likelihood

$$\ell\left(\hat{\boldsymbol{\beta}},\boldsymbol{\varphi};\mathbf{y}\right) \equiv -\frac{1}{2}\log\left(|\mathbf{V}|\right)-\frac{1}{2}\left(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}\right)^T\mathbf{V}^{-1}\left(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}\right),$$

let us focus on just one of the unknown parameters in $\boldsymbol{\varphi}$. Let φ denote such a parameter. Let $S_\varphi\left(\hat{\boldsymbol{\beta}},\boldsymbol{\varphi};\mathbf{y}\right)$ be the score function wrt. φ:

$$\begin{aligned} S_\varphi\left(\hat{\boldsymbol{\beta}},\boldsymbol{\varphi};\mathbf{y}\right) = \frac{\partial}{\partial\varphi}\ell\left(\hat{\boldsymbol{\beta}},\boldsymbol{\varphi};\mathbf{y}\right) &= -\frac{1}{2}|\mathbf{V}^{-1}|\frac{\partial|\mathbf{V}|}{\partial\varphi}-\frac{1}{2}\left(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}\right)^T\frac{\partial\mathbf{V}^{-1}}{\partial\varphi}\left(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}\right) \\ &= -\frac{1}{2}\mathrm{tr}\left\{\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\varphi}\right\}-\frac{1}{2}\left(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}\right)^T\frac{\partial\mathbf{V}^{-1}}{\partial\varphi}\left(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}\right) \\ &= -\frac{1}{2}m\mathrm{tr}\left\{\mathbf{V}_0^{-1}\frac{\partial\mathbf{V}_0}{\partial\varphi}\right\}-\frac{1}{2}\left(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}\right)^T\frac{\partial\mathbf{V}^{-1}}{\partial\varphi}\left(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}\right). \end{aligned}$$

The 3rd equality comes from a rule about the derivative of the determinant of a matrix[13]. The maximum likelihood equations for φ thus becomes

$$m\text{tr}\left\{\mathbf{V}_0^{-1}\frac{\partial \mathbf{V}_0}{\partial \varphi}\right\} = -(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})^T\frac{\partial \mathbf{V}^{-1}}{\partial \varphi}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}). \tag{4.25}$$

Let $LHS(\varphi)$ denote the left-hand side of Equation (4.25), and let $RHS(\varphi)$ denote the right-hand side. The objective is now to solve

$$LHS(\varphi) = RHS(\varphi)$$

letting φ play the part of σ^2, σ_u^2 and ρ in turn.

$\boldsymbol{\varphi = \sigma_u^2}$

Let $\varphi = \sigma_u^2$. From A.14, the estimating equation for σ_u^2 is

$$\begin{aligned} LHS(\sigma_u^2) &= RHS(\sigma_u^2) \\ &\Downarrow \\ m\frac{r_0}{r_0\sigma_u^2+1} &= (\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})^T\frac{1}{(r_0\sigma_u^2+1)^2}\left(\mathbf{I}_{m\times m}\otimes\mathbf{R}_0^{-1}\mathbf{1}_n\mathbf{1}_n^T\mathbf{R}_0^{-1}\right)(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}). \end{aligned}$$

$\boldsymbol{\varphi = \sigma^2}$

Let $\varphi = \sigma^2$. Differentiating $\mathbf{V}_0 = \sigma_u^2\mathbf{J}_{n\times n} + \mathbf{R}_0 = \sigma_u^2\mathbf{J}_{n\times n} + \sigma^2\mathbf{A}$ wrt. σ^2 is just $\mathbf{A}$. The inverse of $\mathbf{V}_0$ has σ_u^2 featured in its expression meaning σ_u^2 should also be featured in $LHS(\sigma^2)$. However, according to [7], $\mathbf{V}_0^{-1} = (\sigma^2\mathbf{A})^{-1}$, and thus

$$LHS(\sigma^2) = m\text{tr}\left\{\mathbf{V}_0^{-1}\frac{\partial \mathbf{V}_0}{\partial \sigma^2}\right\} \overset{\text{A.15}}{=} \frac{1}{\sigma^2}mn.$$

I see no reason as to why $\sigma_u^2\mathbf{J}_{n\times n}$ should vanish from $\mathbf{V}_0$. I have included a word on this in my discussion, Chapter 9. Defining $\delta_{ij} = y_{ij} - \bar{y}_{\cdot j}$, we have

$$RHS(\sigma^2) \overset{\text{A.15}}{=} \frac{1}{\sigma^4(1-\rho^2)}\sum_{i=1}^{m}\left((1+\rho^2)\sum_{j=1}^{n}\delta_{ij}^2 - \rho^2(\delta_{i1}^2+\delta_{in}^2) - 2\rho\sum_{j=2}^{n}\delta_{i,j-1}\delta_{ij}\right).$$

Finally, accepting $LHS(\sigma^2)$ from [7], the estimating equation for σ^2 is

$$\begin{aligned} LHS(\sigma^2) &= RHS(\sigma^2) \\ &\Downarrow \\ \frac{1}{\sigma^2}mn &= \frac{1}{\sigma^4(1-\rho^2)}\sum_{i=1}^{m}\left((1+\rho^2)\sum_{j=1}^{n}\delta_{ij}^2 - \rho^2(\delta_{i1}^2+\delta_{in}^2) - 2\rho\sum_{j=2}^{n}\delta_{i,j-1}\delta_{ij}\right). \end{aligned}$$

$\boldsymbol{\varphi = \rho}$

Let $\varphi = \rho$. Lastly, we must now solve $LHS(\rho) = RHS(\rho)$. From A.16, the estimating equation for ρ is

$$\begin{aligned} LHS(\rho) &= RHS(\rho) \\ &\Downarrow \\ \frac{\rho}{1-\rho^2}2m(n-1) &= \frac{1}{\sigma^2(1-\rho^2)^2}\sum_{i=1}^{m}\left(4\rho\sum_{j=1}^{n}\delta_{ij}^2 - 2\rho(\delta_{i1}^2+\delta_{in}^2) - 2(1+\rho^2)\sum_{j=2}^{n}\delta_{i,j-1}\delta_{ij}\right). \end{aligned}$$

[13] See B.10.

One might have noticed that the solution to $\hat{\sigma}_u^2$ is dependent on $\mathbf{R}_0$, meaning it is dependent on σ^2 and ρ, and that the solution to $\hat{\sigma}^2$ is dependent on ρ, and vice versa. Thus, $\hat{\rho}$, $\hat{\sigma}^2$ and $\hat{\sigma}_u^2$ must be solved numerically.

I have now described how to estimate $\boldsymbol{\varphi}$ in the case of autoregressive covariance structure. Finding the best covariance structure can at times be difficult. Without having to make any assumptions on the relationship between and within subjects, one could always just use the unstructured covariance. When $\text{Corr}\left[Y_{ij}, Y_{il}\right] = \rho_{j,l}$, then $\text{Cov}\left[Y_{ij}, Y_{il}\right] = \sigma^2\rho_{j,l}$. The within-subject covariance is then

$$\mathbf{R}_0 = \sigma^2 \begin{bmatrix} 1 & \rho_{1,2} & \rho_{1,3} & \cdots & \rho_{1,n} \\ \rho_{2,1} & 1 & \rho_{2,3} & \cdots & \rho_{2,n} \\ \rho_{3,1} & \rho_{3,2} & 1 & \cdots & \rho_{3,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{n,1} & \rho_{n,2} & \rho_{n,3} & \cdots & 1 \end{bmatrix}.$$

It is worth mentioning that with this type of covariance structure, in addition to estimating σ_u^2 and σ^2, one will also have to estimate $\frac{n_i(n_i-1)}{2}$ parameters for the correlations $\rho_{i,j}$. One should be careful using the unstructured covariance, as a large number of parameters can lead to unstable estimates.

> **!** As $n_i = 3$, $\forall i$, only $\frac{3\cdot(3-1)}{2} = 3$ correlation parameters must be estimated, thus it should create no problems using the unstructured covariance wrt. my data.

4.4 Kenward-Roger approximation

This section is based on [11].

In order to find out whether there is any difference between the treatment groups, we must do inference about $\boldsymbol{\beta}$. This can be done in a few ways, for instance by using a likelihood ratio test[14] or a Wald test[15]. In the likelihood ratio test (or LRT for short), we compare a likelihood function, L, with a reduced version of the same likelihood, L_0, in which the group-term is excluded. Let $\mathbf{X}$ and $\boldsymbol{\beta}$ be the design matrix and set of coefficients in L, and let $\mathbf{X}_0$ and $\boldsymbol{\beta}_0$ be the reduced design matrix and reduced set of coefficients in L_0. We want to test the hypothesis

$$H_0^{\text{KR}_1} : \mathbf{y} = \mathbf{X}_0\boldsymbol{\beta}_0 + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}.$$

This is tested via the test statistic

$$T = 2\left(\log(L) - \log(L_0)\right),$$

which under $H_0^{\text{KR}_1}$ will be asymptotically χ_d^2-distributed.

In the Wald test, we test the hypothesis

$$H_0^{\text{KR}_2} : \mathbf{K}\boldsymbol{\beta} = \boldsymbol{\beta}_0$$

where $\mathbf{K} \in \mathbb{R}^{d\times p}$ is a matrix such that $\text{E}[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta} \wedge \mathbf{K}\boldsymbol{\beta} = \mathbf{0}$ is equivalent to $\text{E}[\mathbf{Y}] = \mathbf{X}_0\boldsymbol{\beta}_0$. It is tested via the test statistic

$$W = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T\mathbf{K}^T\left(\mathbf{K}^T(\mathbf{X}^T\mathbf{V}(\hat{\boldsymbol{\varphi}})^{-1}\mathbf{X})^{-1}\mathbf{K}\right)^{-1}\mathbf{K}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$$

with $(\mathbf{X}^T\mathbf{V}(\hat{\boldsymbol{\varphi}})^{-1}\mathbf{X})^{-1}$ being the covariance of $\hat{\boldsymbol{\beta}}$. Under $H_0^{\text{KR}_2}$, W will also be asymptotically χ_d^2-distributed.

[14] See B.11

[15] See B.12.

For small sample sizes, the approximation of the asymptotical distributions of T or W under the null-hypotheses can be poor, which may lead to misleading conclusions. It would be preferable to have a test statistic that does not have an *asymptotical* distribution, but rather *approximates* some distribution. For instance, by using a specific modification of W, we can get a test statistic that approximates an F-distribution. For some models, this new test statistic even has an exact F-distribution under $H_0^{\mathrm{KR}_2}$. In order to find an F-distributed test statistic, let $F = \frac{1}{d}W$ be a scaled test statistic with an asymptotical $\frac{1}{d}\chi^2_d$-distribution. We can modify F by approximating its distribution by an F-distribution. The idea is to match the moments of F with those of an $F(d,m)$-distributed random variable, where m is the denominator degrees of freedom. In addition to m, we also derive a scaling factor, λ, which is multiplied to F. Both λ and m are unknown. To find them, we set up a system of equations in which we match the moments of F with those of an $F(d,m)$-distributed random variable. The moments of F are approximated by a first order Taylor expansion of F. Let E^* and V^* denote these approximations for the mean and variance, respectively. Then the system of equations, we want to solve is

$$\begin{aligned}\mathrm{E}[\lambda F] &\approx \lambda E^* = \mathrm{E}[F(d,m)] = \frac{m}{m-2}\\ \mathrm{Var}[\lambda F] &\approx \lambda^2 V^* = \mathrm{Var}[F(d,m)] = \frac{2m^2(d+m-2)}{d(m-2)^2(m-4)},\end{aligned}$$

where $\mathrm{E}[F(d,m)]$ and $\mathrm{Var}[F(d,m)]$ are the moments of an $F(d,m)$-distributed random variable. Once we have $\hat{\lambda}$ and $\hat{m}$, we have approximations of the moments of F, that resemble those from an F-distribution. Hence, we have achieved our goal of finding an F-distributited test statistic. Or rather, under $H_0^{\mathrm{KR}_2}$, we have that $\lambda F \sim F(d,m)$. This approach is known as the *Kenward-Roger approximation*.

4.5 Fitted values of the LMM

This section is based on [8].

In [8] they say, that it is not really clear what the fitted values are in an LMM. In the simple classical model, Equation (4.1), if y_i is assumed Gaussian, then $Y_i \sim N(x_i\beta, \sigma^2)$. Having observed nothing else, the best prediction of y_i will be $\hat{\mu}_i = x_i\hat{\beta}$, having estimated the coefficient β. This means that the fitted value for y_i is just the individual mean $\hat{\mu}_i$. The fitted values of the LMM are a little more complicated. To see this, let us assume we have a balanced random effects model

$$\begin{aligned}&y_{ij} = \mu + u_i + \epsilon_{ij}, \quad i = 1,\dots,n\\ &U_i \overset{\text{iid.}}{\sim} N(0,\sigma_u^2),\\ &\epsilon_{ij} \overset{\text{iid.}}{\sim} N(0,\sigma^2),\\ &Y_{ij} \sim N(\mu, \sigma_u^2+\sigma^2),\end{aligned}$$

where u_i and ϵ_{ij} are independent. If we let $\mu_i = \mu + u_i$ so that $y_{ij} = \mu_i + \epsilon_{ij}$, we have the hierchical model

$$Y_{ij} \mid \mu_i \sim N(\mu_i, \sigma^2),$$

where $\mu_i \overset{\text{iid.}}{\sim} N(\mu, \sigma_u^2)$. Once μ_i have been estimated, we get the best prediction for the ith subject as $\hat{y}_{ij} = \mathrm{E}[Y_{ij} \mid \hat{\mu}_i] = \hat{\mu}_i$, where

$$\hat{\mu}_i \overset{\text{A.17}}{=} \left(1 - \frac{\sigma_u^2}{\sigma_u^2 + \frac{\sigma^2}{n}}\right)\mu + \frac{\sigma_u^2}{\sigma_u^2 + \frac{\sigma^2}{n}}\bar{y}_{i\cdot}.$$

So, the fitted value $\hat{y}_{ij}$ is not just the individual mean, like in simple model, but a weighted average between the overall mean, μ, and the group average, $\bar{y}_{i\cdot}$.

Example 1 In this example, I will set up a random effects model. The purpose of the example is to show how the fitted values differ from the sample mean. I will be using a subset, `dat`, of the data set `ant111b` from the package `DAAG`:

```
> head(dat, 8)
   site harvwt
1  DBAN  5.160
2  LFAN  2.930
3  NSAN  1.730
4  ORAN  6.790
5  OVAN  3.255
6  TEAN  2.650
7  WEAN  5.040
8  WLAN  2.020
```

On different islands in the Carribean, the harvest weight for different sites have been measured. Different types of treatment have been used, and each site is seperated into four plots. Hence, there is four measurements per site. In this example, we are only interested in the harvest weight, as we are setting up a random effects model, and will thus not be considering that the weight could be influenced by the island or the treatment, as these would be fixed effects. As such, the subset `dat` only consists of the variables `harvwt` (the harvest weight) and `site` (the site), which has 8 levels.

The random effects model have two sources of variation; σ_u^2, the variation between sites, and σ^2, the variation within sites. Figure 5 shows that there is great variation between sites, and the variations within sites are a lot smaller.

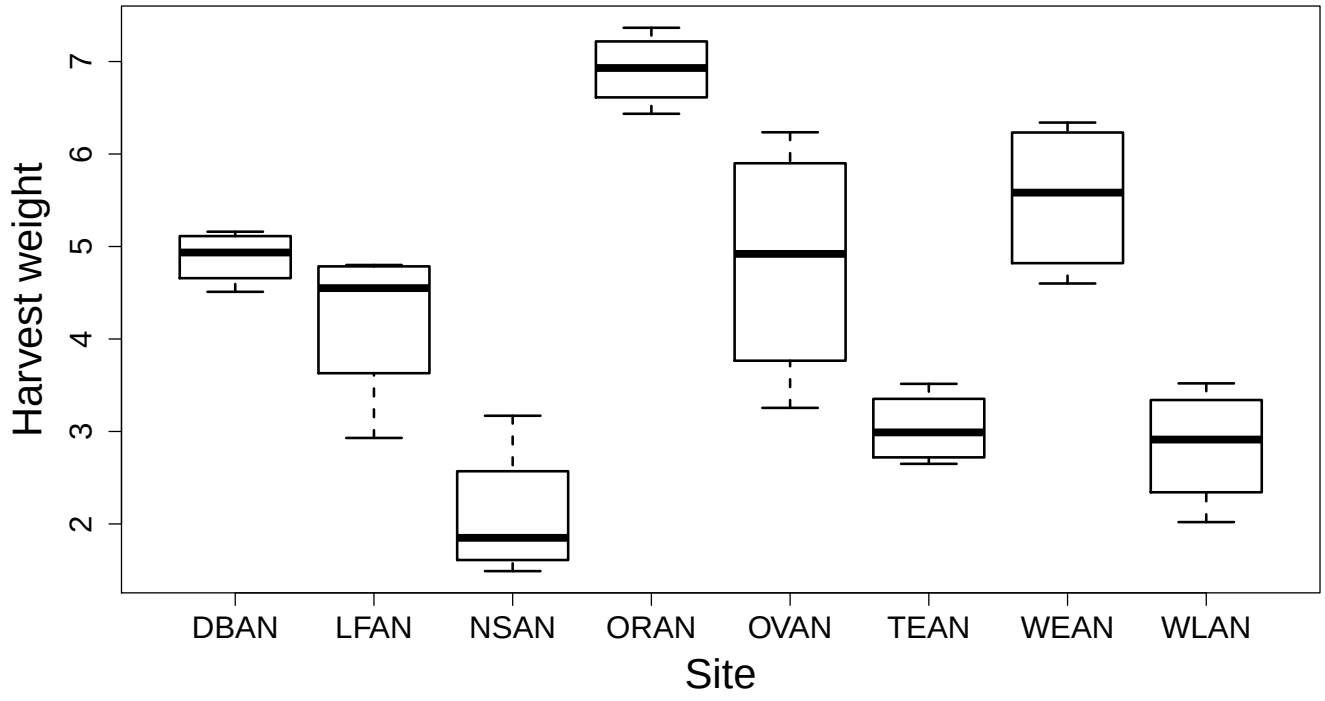

Figure 5: *Boxplot of the variation of harvest weight within each site.*

For the model, I am using the exchangeable covariance structure, hence I am assuming the correlation between plots in each site is the same. The model is:

```
Model = lme(fixed = harvwt ~ 1, random = ~1|site, correlation = corCompSymm(form = ~1|site),
    data = dat)
```

In Chapter 5, I go into a few more details about the functions in `R` for modelling an LMM. As I have no fixed effects, I only write "`~ 1`" for the `fixed`-argument. For the `random`-argument, I use `site` as this is the random effect. With `corCompSymm()`, I specify, I want to use the exchangeable covariance structure.

First, I extract $\hat{\sigma}_u^2$ and $\hat{\sigma}^2$:

```
> VarCorr(Model)
site = pdLogChol(1)
            Variance   StdDev
(Intercept) 2.3677325 1.5387438
Residual    0.5775378 0.7599591
```

We see that $\hat{\sigma}_u^2 = 2.3677325$ and $\hat{\sigma}^2 = 0.5775378$, confirming the suspicion from Figure 5, that $\hat{\sigma}_u^2 > \hat{\sigma}^2$.

Next, I look at the residuals to make sure the model is sensible. Figure 6 shows no unwanted patterns in the residuals. One thing to notice is the size of the residuals. The range of `harvwt` is $[1.490, 7.365]$. The residuals' range is $[-1.546, 1.434]$. The residuals are quite big compared to the values of `harvwt`. This, however, does not mean that `Model` is a bad model; it is very common for mixed models to have large residuals, because the residuals are a little different from those of, say, a general linear model. This is due to the more complicated nature of the fitted values of the LMM.

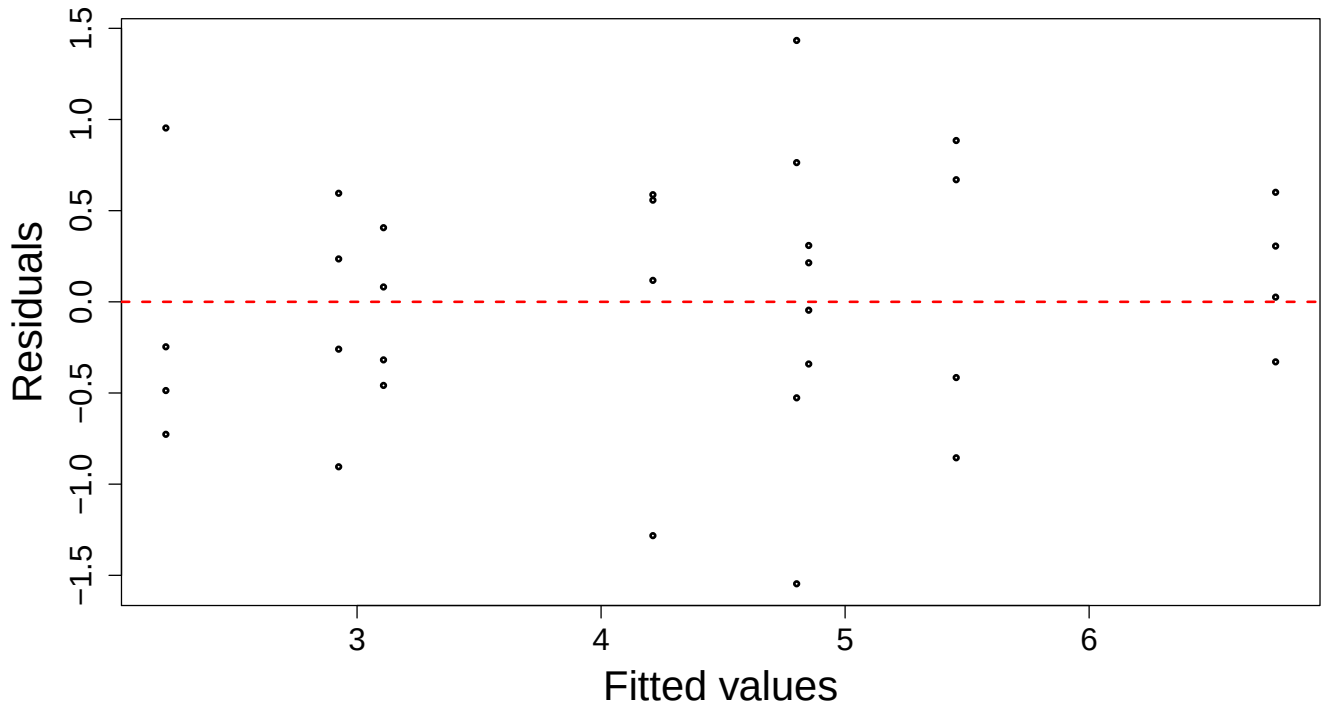

Figure 6: *Plot of the residuals vs. the fitted values of* `Model`.

Lastly, I compare the sample means and the fitted values. As we have no fixed effects in the model to distinguish between the plots, the fitted values for each plot on a site, are the same:

```
> split(Model$fitted[,2], dat$site)
$DBAN
       1        9       17       25
4.850901 4.850901 4.850901 4.850901

$LFAN
       2       10       18       26
4.21234 4.21234 4.21234 4.21234

$NSAN
       3       11       19       27
2.216544 2.216544 2.216544 2.216544

$ORAN
       4       12       20       28
6.764226 6.764226 6.764226 6.764226

$OVAN
       5       13       21       29
4.801418 4.801418 4.801418 4.801418

$TEAN
       6       14       22       30
3.108408 3.108408 3.108408 3.108408

$WEAN
       7       15       23       31
5.455295 5.455295 5.455295 5.455295

$WLAN
       8       16       24       32
2.924616 2.924616 2.924616 2.924616
```

Hence, when we take the mean of the fitted values for each site, the mean is exactly the same as the fitted value. Below are the fitted values for each site (`fit`) and the means of `harvwt` at each site (`means`). The data frame shows the difference in the sample means and the fitted values:

```
> fit = sapply(split(Model$fitted[,2], dat$site), mean)
> means = sapply(split(dat$harvwt, dat$site), mean)
> data.frame(mean = means, fitted = fit)
        mean    fitted
DBAN 4.88500 4.850901
LFAN 4.20750 4.212340
NSAN 2.09000 2.216544
ORAN 6.91500 6.764226
OVAN 4.83250 4.801418
TEAN 3.03625 3.108408
WEAN 5.52625 5.455295
WLAN 2.84125 2.924616
```

So, the fitted values in an LMM is not just the sample means, and hence, as [8] hinted at, they are more difficult to understand.

□

4.6 The LMM with random intercepts and slopes

This section is based on [2] and [10].

The LMM with random intercepts is a special LMM. The idea of the *random intercepts model* (RIM for short) is, that it is not always sensible to assume, that subjects have the same initial levels. The RIM allows for the influence of each subject on their repeated measurements. From the vocabulary-test example, we know that the most simple[16] LMM is given as

$$y_{ij} = \mu + x_{ij}\beta + u_i + \epsilon_{ij}.$$

Obviously, here we have a common initial level for all subjects, given by μ. Inserting $\mu_i = \mu + u_i$ back into the equation, we get the RIM

$$y_{ij} = \mu_i + x_{ij}\beta + \epsilon_{ij}, \tag{4.26}$$

where y_{ij} is the jth observation for the ith subject. The u_i is, as before, the intercept-deviation for subject i from the average intercept μ. Equation (4.26) indicates that subject i's response at the jth timepoint is influenced by the subject's initial level. We still have the assumption that $U_i \sim N\left(0, \sigma_u^2\right)$ and $\epsilon_{ij} \sim N\left(0, \sigma^2\right)$, and that u_i and ϵ_{ij} are independent.

In the RIM, the variance of each measurement is

$$\text{Var}\left[Y_{ij}\right] = \text{Var}\left[U_i + \epsilon_{ij}\right] \;=\; \sigma_u^2 + \sigma^2.$$

This is also called the *total residual variance*, as $u_i + \epsilon_{ij}$ is called the *total residuals*. The covariance between the total residuals at two timepoints, j and k, from the same subject is

$$\text{Cov}\left[Y_{ij}, Y_{ik}\right] = \text{Cov}\left[U_i + \epsilon_{ij}, U_i + \epsilon_{ik}\right] \overset{\text{A.18}}{=} \sigma_u^2, \quad j \neq k.$$

The correlation is then

$$\text{Corr}\left[Y_{ij}, Y_{ik}\right] = \text{Corr}\left[U_i + \epsilon_{ij}, U_i + \epsilon_{ik}\right] \;=\; \frac{\sigma_u^2}{\sigma_u^2 + \sigma^2}, \quad j \neq k.$$

This correlation is known as *compound symmetry*. It gives the ratio of the individual variance, σ_u^2, to the total variance, $\sigma_u^2 + \sigma^2$. Because the correlation is constant, we say that the RIM has exchangeable correlation structure.

As visible by $\text{Var}\left[Y_{ij}\right]$ and $\text{Cov}\left[Y_{ij}, Y_{ik}\right]$, the RIM constrains the variance of each measurement to be the same and the covariance between any pair of measurements (for the same subject, of course) to be equal. There are certainly cases where this constraint is unrealistic. For longitudinal data, one can easily imagine that measurements taken further apart will be less correlated than those taken close to each other in time. Consequently, $\text{Cov}\left[Y_{ij}, Y_{ik}\right] = \sigma_u^2$ will not match the observed covariance pattern. A way to deal with this is to allow for heterogeneous slopes. With the slope being just β, we have that in the RIM it is assumed that the rate of change across time is the same for all subjects. This may not always be a fair assumption. We can extend the RIM by also allowing for subject-specific slopes. This is the *random intercepts and slopes model* (RISM for short), and it is given by

$$y_{ij} = \mu_i + x_{ij}\beta + x_{ij}b_i + \epsilon_{ij} \;=\; \mu_i + x_{ij}\tilde{b}_i + \epsilon_{ij} \tag{4.27}$$

[16] "Simple" as in there is only one fixed and one random effect. The RIM, however, is not restricted to having only one fixed effect, but it must have only one random.

with $\tilde{b}_i = \beta + b_i$. The b_i is the slope-deviation for subject i from the average slope β. It is assumed that $b_i \sim N(0, \sigma_b^2)$, and that also b_i and ϵ_{ij} are independent. Let σ_{ub} denote the covariance of u_i and b_i. The random-effects covariance matrix is

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_u^2 & \sigma_{ub} \\ \sigma_{ub} & \sigma_b^2 \end{bmatrix}.$$

Being a covariance matrix, one must make sure $\mathbf{\Sigma}$ is positive semi-definit. The total residual is now $u_i + x_{ij} b_i + \epsilon_{ij}$, and

$$\begin{aligned} \text{Var}\left[Y_{ij}\right] &= \text{Var}\left[U_i + x_{ij} b_i + \epsilon_{ij}\right] \\ &= \text{Var}\left[U_i\right] + \text{Var}\left[x_{ij} b_i\right] + \text{Var}\left[\epsilon_{ij}\right] + 2\text{Cov}\left[U_i, x_{ij} b_i\right] + 2\text{Cov}\left[U_i, \epsilon_{ij}\right] + 2\text{Cov}\left[x_{ij} b_i, \epsilon_{ij}\right] \\ &= \sigma_u^2 + x_{ij}^2 \sigma_b^2 + \sigma^2 + 2 x_{ij} \sigma_{ub}. \end{aligned}$$

We see that the variance of a measurement now depends on the timepoint at which it was taken. This is apparent by the subscribt j. The covariance becomes

$$\text{Cov}\left[Y_{ij}, Y_{ik}\right] = \text{Cov}\left[U_i + x_{ij} b_i + \epsilon_{ij}, U_i + x_{ik} b_i + \epsilon_{ik}\right] \overset{\text{A.19}}{=} \sigma_u^2 + x_{ij} x_{ik} \sigma_b^2 + (x_{ij} + x_{ik}) \sigma_{ub}.$$

The covariance is no longer the same for any pair of measurements on a subject.

The variance terms indicate how much heterogeneity there is in the population; σ_u^2 indicates the spread around the population intercept, and σ_b^2 indicates the spread around the population slope, and σ_{ub} indicate to what degree u_i and b_i co-vary. For instance, $\sigma_{ub} > 0$ suggests that subjects with higher initial values (than the population average) also have higher slopes, and the opposite if $\sigma_{ub} < 0$.

> **!** If there indeed is a difference in how well patients recover according to the type of treatment, then it certainly makes sense to use a RIM or RISM; it would seem likely that patients in group 3 would have a higher intercept and/or a steeper positive slope than patients in, say, group 1.

I will now move on to performing an analysis on my data using mixed models.

5 Results of modelling data with LMMs

In this chapter, I go through the results of modelling my respons variables with LMMs as described in Chapter 4. In Chapter 3, I could not definetly determine the distribution of the samples as being Gaussian. I have investigated this further, and by visual inspection, I find the Gaussian distribution the most fitting (despite the Shapiro-Wilk test rejecting all samples as being Gaussian in Chapter 3). Hence, I will continue under the assumption, that my data is in fact Gaussian. I will model the three types of movement seperately. I will start the analysis by modelling the pro/sup-movement. I will do this in as many details, I feel is needed. Afterwards, I will perform the analysis of the uln/rad- and ex/flex-movements, but in much less details. I will be using different types of covariance structures, and I will compare the results to see which covariance structure is most realistic for the data.

5.1 Analysis using `correlation = corCAR1()`

First thing I must decide is which function in `R` to use. The packages `nmle` and `lme4` both have functions for mixed models. I find the function `lme()` from the `nlme`-package the best of all my choices[17].

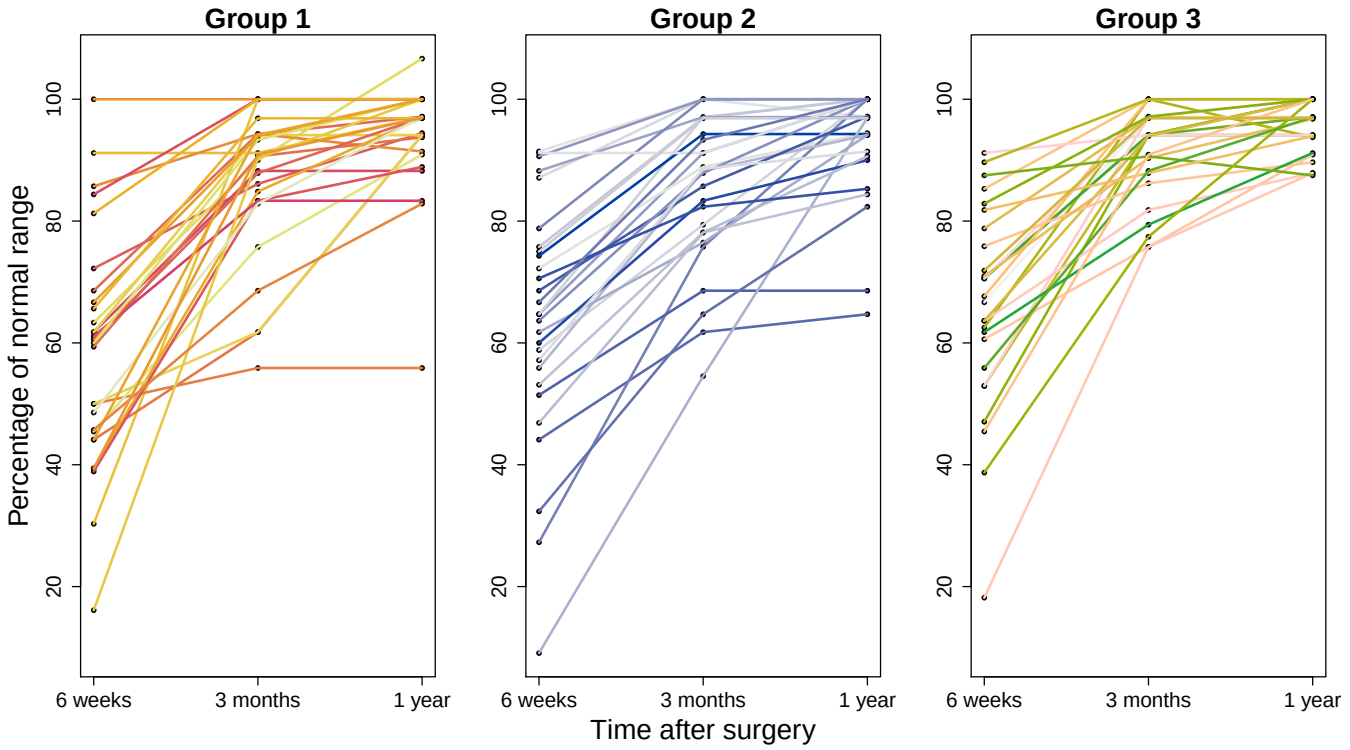

Figure 7: *In this plot only the measurements of pro/sup in percentages are used. Each graph represents a group. Each line in each graph represents the time series for a patient in that group.*

Next thing I must decide is whether I should use a RIM or a RISM. This is easily decided by simply plotting the time series of `ps` for every patient. Figure 7 shows that patients do indeed have very different intercepts, and the lines also do not seem to prescribe to one common slope. The figure also shows that the intercepts for group 3 are not higher than those of the other two groups, and the slopes of group 3 does not look steeper than the slopes of the other two groups. Hence, I do not see what I might expect

[17] Other functions in the packages either could not fit models with random effects, only allowed for the exchangeable structure, or were for a non-linear fit. Hence, I found `lme()` the best of my choices as it does not have these restrictions.

to see, if there really was a difference in the treatment groups. Either way, I will set up both a RIM and a RISM, and use `anova()` to see which model, I should use. Remember, `anova()` tests the hypothesis, that there is no significant difference in the models given as input.

In setting up the models, I have to specify the covariance structure through the argument `correlation` in `lme()`. In accordance with the four types of covariance structure discussed in Section 4.3, I have listed how they are specified in `lme()`:

- Independence: `correlation` defaults to the independence structure when nothing is specified.
- Exchangeable: As explained in Section 4.6, the exchangeable structure is also known as compound symmetry, hence `correlation = corCompSymm()` is for the exchangeable structure.
- Autoregression: There are two choices for an autoregressive structure of order 1; `correlation = corAR1()` and `correlation = corCAR1()`. Only with `corCAR1()` can I account for having non-equidistant timepoints.
- Unstructured: With `correlation = corSymm()`, no structure is imposed on the correlation matrix.

As I find it reasonable to use the autoregressive structure for my data, I will start my analysis with `corCAR1()`. I will set up three models with this autoregressive structure, and I will refer to these models as the "ar"-models and will be adding ".ar" to their names in R. As I am setting up *linear* models, I must make sure the relationship between the dependent and the independent variables is linear. I already investigated this in Chapter 1, and came to the conclusion, that I needed a squared term of `time`. Below, I have set up two models; a RIM and a RISM:

```
1 ps.rim.ar = lme(ps ~ time + I(time^2) + group, random = ~1|subject, data = Wrist, method = "REML
      ", correlation = corCAR1(form = ~ time|subject), control = lmeControl(opt = "optim",
      msMaxIter = 60))
2 ps.rism.ar = update(ps.rim, random = ~1+time|subject)
```

Adding `time` to the `random`-argument is what creates the random slopes. For the `correlation`-argument, I need to specify that `time` should be used when calculating the correlations and that observations with different levels in `subject` are uncorrelated. Hence, `corCAR1(form = ~time|subject)`. With this covariance structure, I needed a few extra iterations than the default 50 to reach convergence. Hence, the `control`-argument. Below the result of the test is shown:

```
1 > anova(ps.rim.ar, ps.rism.ar)
2           Model df      AIC      BIC    logLik   Test  L.Ratio p-value
3 ps.rim.ar     1  8 1973.976 2001.953 -978.9880
4 ps.rism.ar    2 10 1936.519 1971.491 -958.2597 1 vs 2 41.45667  <.0001
```

The `anova`-function takes either one or multiple inputs. In Appendix B, I have explained how the function works and what it tests in those situations. The test above shows that the RISM is prefered with a p-value below 0.0001. Thus I continue my analysis of the pro/sup-movement with a RISM.

Going forward, I will start by taking a look at the estimated coefficients. I will confirm that $\mathbf{V}_0$, $\mathbf{R}_0$ and $\boldsymbol{\Sigma}$ are positive semi-definit. I will test whether `group` is relevant to the model, and then end the analysis by checking that the model assumptions are satisfied.

The RISM for my data looks like

$$y_{ij} = \mu + \text{time}_j \beta_{\text{time}} + \text{time}_j^2 \beta_{\text{time}^2} + \mathbb{1}[\text{group } 2]\beta_{\text{gr2}} + \mathbb{1}[\text{group } 3]\beta_{\text{gr3}} + \text{time}_j b_i + u_i + \epsilon_{ij},$$

where $\text{time}_j \in \{0, 7, 46\}$ is the timepoints and

$$\mathbb{1}[\text{group } m] = \begin{cases} 1, & \text{patients belongs to group } m \\ 0, & \text{otherwise} \end{cases} \qquad \text{for } m = 2, 3.$$

Their corresponding coefficients are β_{time}, β_{time^2}, β_{gr2} and β_{gr3}, respectively. This is just an extention of the RISM, Equation (4.27), in Section 4.6. The fixed effects are extracted form `ps.rism.ar` by:

```
> ps.rism.ar$coefficients$fixed
(Intercept)        time    I(time^2)      group2      group3
  61.893866    4.145904   -0.075069   -0.286157    2.704198
```

This means that

$$\mu = 61.89387, \quad \beta_{\text{time}} = 4.14590, \quad \beta_{\text{time}^2} = -0.075070, \quad \beta_{\text{gr2}} = -0.28616, \quad \beta_{\text{gr3}} = 2.70420.$$

The random effects are extracted by (here only the eight first patients):

```
> ps.rism.ar$coefficients$random$subject
      (Intercept)        time
id01    4.252543 -0.0753924
id02   14.964132 -0.2611401
id03    6.033762 -0.0995685
id04   -8.695947  0.1725030
id05    2.738471 -0.0227380
id06   -0.096404 -0.0089380
id07  -20.615433  0.2857262
id08    2.673071 -0.0452410
```

The "`(Intercept)`"-column is the u_i's and the "`time`"-column is the b_i's. The fitted value for the 1st patient (who belongs to group 1) at the 2nd time point is then

$$\begin{aligned} \hat{y}_{12} &= 61.89387 + 7 \cdot 4.14590 + 7^2 \cdot (-0.07507) + 0 \cdot (-0.28616) + 0 \cdot 2.70420 + 7 \cdot (-0.075392) + 4.25254 \\ &= 90.96162. \end{aligned}$$

We can check this result against the result from `R`:

```
> fitted(ps.rism.ar)[2]
    id01
90.96162
```

I must now check that all the covariance matrices, $\mathbf{V}_0$, $\mathbf{R}_0$ and $\boldsymbol{\Sigma}$, are positive semi-definite. They are extracted through the function `getVarCov()`, which is implemented in my function `superLMM()`. The result for `ps.rism.ar` is:

```
> superLMM(ps.rism.ar)$vcov
        V0   R0 Sigma
[1,] TRUE TRUE  TRUE
```

They are all positive semi-definit.

Finally, is `group` significant to the model? I will answer this via three tests. The first test is `anova()` with the model as input. It tests whether each of the terms in the formula of the model are significant. The second test is also `anova()`, but with two inputs; the model and the corresponding null-model. It compares the models via an LRT. When I use the term "null-model" in this chapter, I mean a model without the `group`-term. The last test is performed by the function `intervals()`, which gives the 95%-confidence intervals for the parameter estimates. If 0 is contained in an interval, then we can accept that $\beta_i = 0$. The results of the three tests, in the order mentioned, for `ps.rism.ar` are:

```
> superLMM(ps.rism.ar)$group
$anova1
             p-value
(Intercept) 0.00000
time        0.00000
I(time^2)   0.00000
group       0.36011

$anova2
[1] 0.28728

$intervals
                lower      est.     upper
(Intercept) 57.65469 61.89387 66.13304
time         3.71112  4.14590  4.58068
I(time^2)   -0.08393 -0.07507 -0.06621
group2      -4.79709 -0.28616  4.22478
group3      -1.84831  2.70420  7.25671
```

We see that both of the `anova`-tests say that `group` is insignificant to the model, and that we can set $\beta_{\text{gr.2}} = \beta_{\text{gr.3}} = 0$.

Checking model assumptions

Before these results can be accepted, I must check that the model assumptions are satisfied. The assumptions were that

i. the random effects are independent, homoscedastic and Gaussian,

ii. the errors are independent, homoscedastic and Gaussian, and

iii. explanatory variables are related linearly to the response.

Wrt. to the 3rd assumption, I believe, I have already have the explanatory variables in their correct form, as I have included a squared term of `time`. I cannot test if `group` should be transformed as it is a factor and it would make no sense to, say, take the logarithm of `group`. Hence, I will concern myself with only the first two assumptions.

To test assumption i. and ii., I have to supply `superLMM()` with the argument `ass`, which is either `1` or `2`, respectively. When setting `ass = 1`, if the input model is a RIM, nothing else needs to be put into `superLMM()`. The function then makes three plots; a plot of the random intercepts vs. patient index, a Q-Q plot and a histogram of the intercepts. It also outputs the p-values from performing a Bartlett test and Levene's test for homoscedasticity and the p-value from performing a Shapiro-Wilk test for normality. If the input model is a RISM, `superLMM()` also needs the argument `is` specified, which is either `"I"` or

`"S"` for random intercepts or slopes, respectively. If `is = "I"`, the function makes the same plots and tests of the intercepts as just described. If `is = "S"`, the function makes the same plots and tests as just described, but for the random slopes instead. Below are the results from `ps.rism.ar` when setting `is = "I"`, and Figure 8 shows the three plots:

```
> superLMM(ps.rism.ar, ass = 1, is = "I")
            Bart     Leve     Shap
p-value 0.17719 0.20025 0.11581
```

All the p-values are above 0.05, so according to the tests, assumption i. is satisfied for the random intercepts. The Q-Q plots in Figure 8 agrees with the Shapiro-Wilk test. The histograms, however, do not look as good as expected from the nice Q-Q plots. The top plot in Figure 8 shows no unwanted patterns in the intercepts, which agrees with the tests.

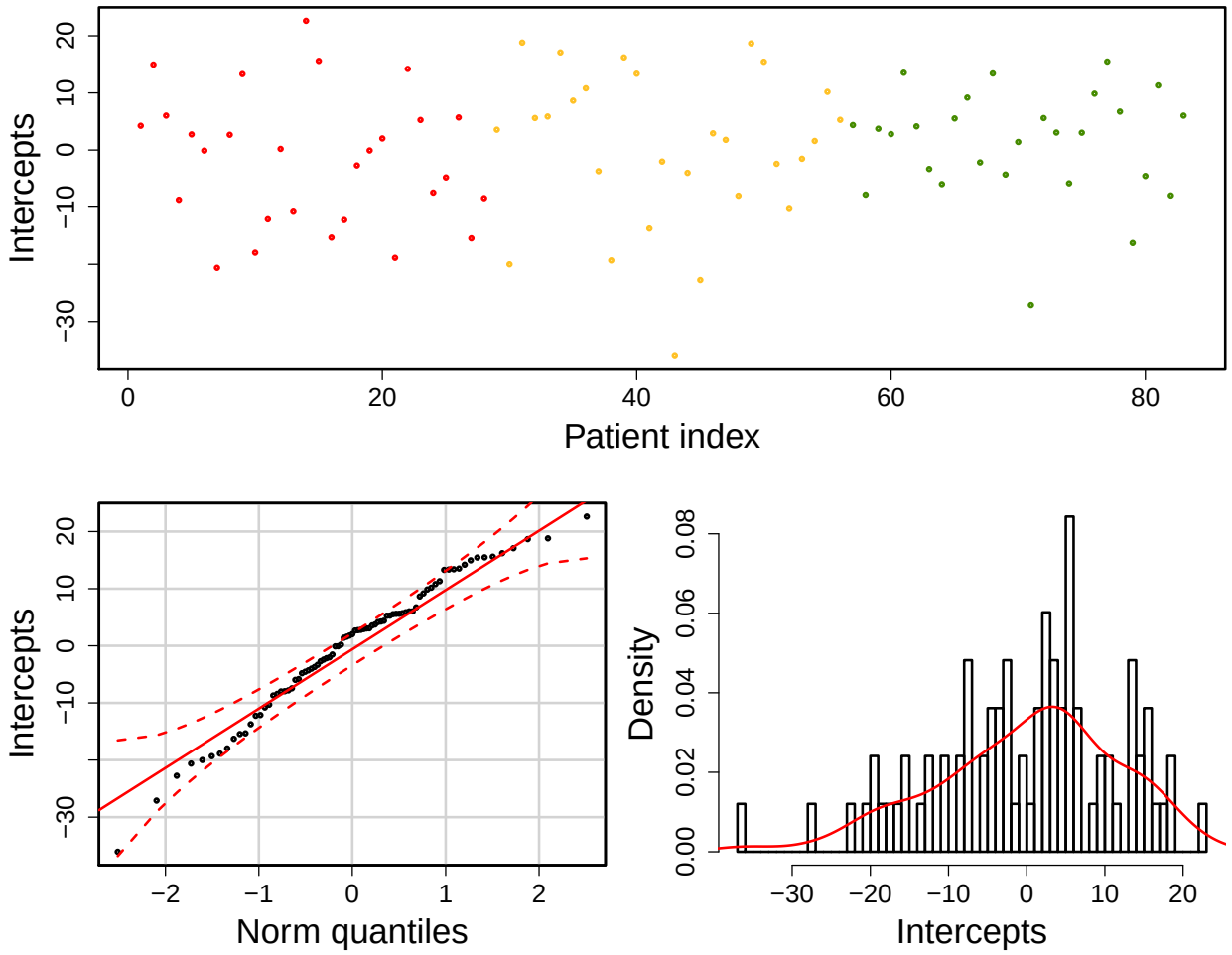

Figure 8: *1st row: Intercepts vs. patient index. The red, yellow and green dots are for groups 1, 2 and 3, respectively. 2nd row: Q-Q plot and histogram with density curve of the random intercepts.*

Next, we must test the random slopes. The results are:

```
> superLMM(ps.rism.ar, ass = 1, is = "S")
            Bart     Leve     Shap
p-value 0.23366 0.30335 0.12843
```

Assumption i. is satisfied for the random slopes, according to the tests. Figure 9 looks just about the same as Figure 8; no patterns in the top plot, a Q-Q plot, that agrees with the Shapiro-Wilk test and a histogram, that does not look as good as expected.

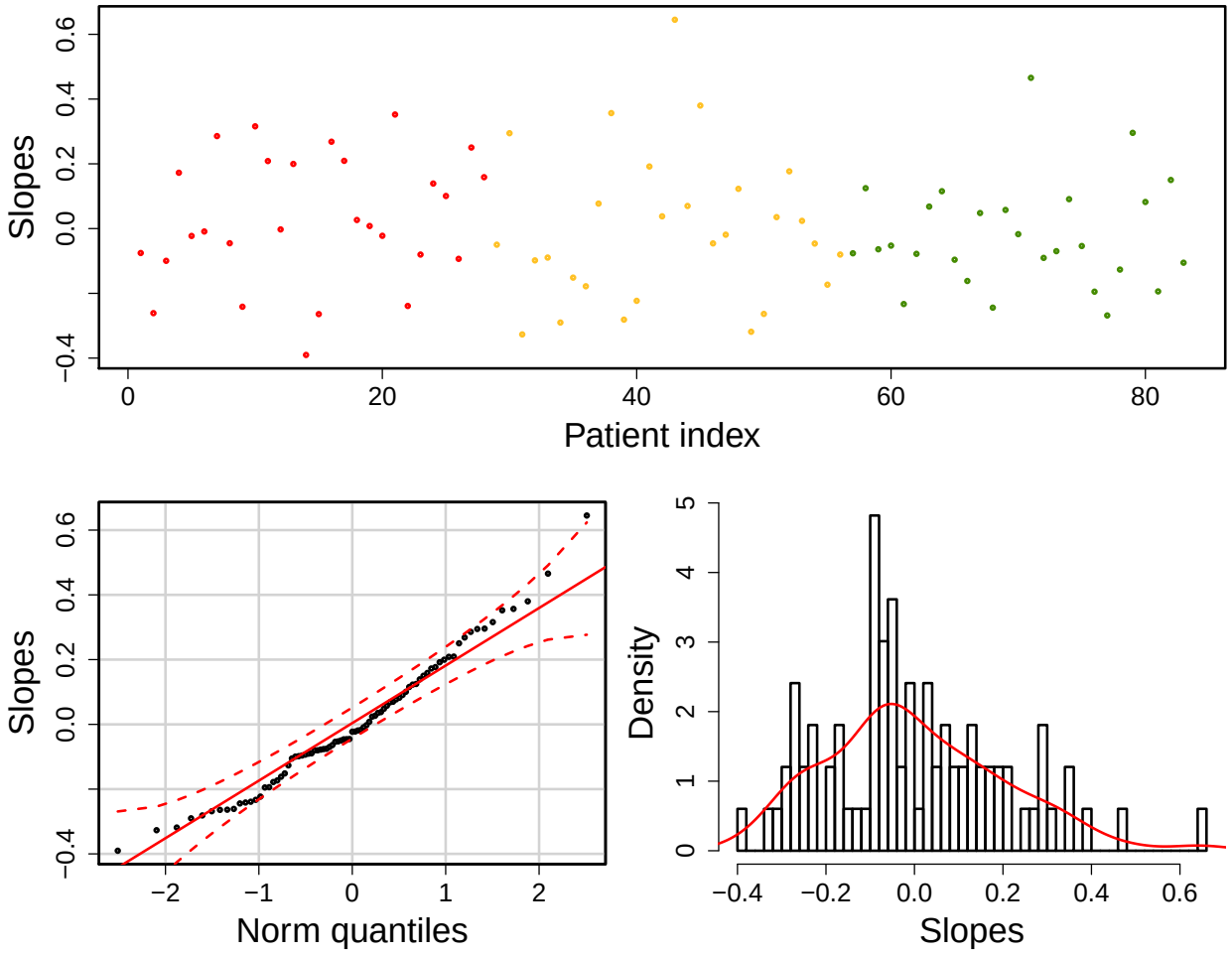

Figure 9: *1st row: Slopes vs. patient index. The red, yellow and green dots are for groups 1, 2 and 3, respectively. 2nd row: Q-Q plot and histogram with density curve of the random slopes.*

When setting `ass = 2`, `superLMM()` makes the three plots in Figure 10, and it gives the p-values from performing a Bartlett test and Levene's test on the residuals to test for homoscedasticity, and the p-value from performing a Shapiro-Wilk test for normality. When performing a Bartlett test and Levene's test to confirm the homoscedastic variance of the residuals, I get conflicting results:

```
> superLMM(ps.rism.ar, ass = 2)
             Bart     Leve      Shap
p-value 0.00031 0.82783 7.498e-07
```

To investigate these results further, I use the top plot in Figure 10, which deserves an explanation: it is intended for detecting patterns in the residuals for each patient. The vector of residuals has $3 \cdot 83$ entries, that is, three residuals per patient (one for each timepoint). These are connected by a colored line (red, yellow or green for group 1, 2 or 3, respectively). An unwanted pattern that could show heteroscedasticity could be if the sets of residuals for each patient were increasing or forming a "U"-shape. In Figure 10, I see both increasing sets and "U"-shapes, but I also see many other shapes and, most importantly,

no obvious reacurring patterns. Hence, I agree with the Levene's test. The Shapiro-Wilk test says that the normality assumption has been violated. The Levene's test is not as sensitive as the Bartlett test wrt. violations of normality.

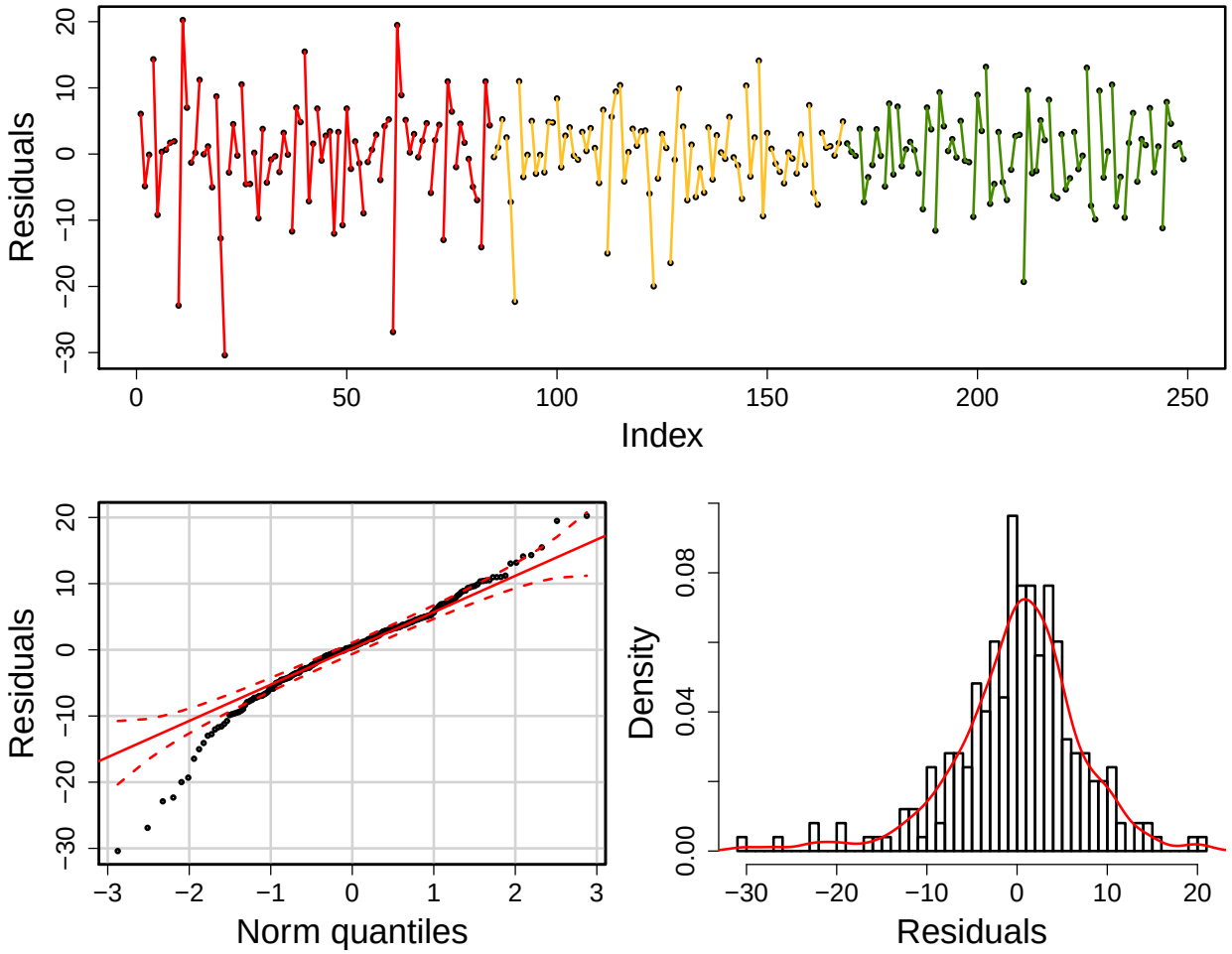

Figure 10: *1st row: Residuals plotted against their index. Each colored line represent the residuals for a patient (red, yellow or green for group 1, 2 or 3, respectively). 2nd row: Q-Q plot and histogram with density curve of the residuals.*

The result of the Shapiro-Wilk test is a little surprising. As both the intercepts and the slopes were accepted as being Gaussian despite the not so pretty histograms, I find it strange that the residuals are not Gaussian, when their histogram (see Figure 10) much more looks like something from a Gaussian sample (in fact, if you simulate a Gaussian sample, it is not difficult to get a result that easily resembles both the Q-Q plot and histogram in Figure 10). Perhaps these confusing results are due to of the small sample size.

I am sufficiently satisfied, that assumption i. and ii. have been satisfied. I end the analysis of the pro/sup-movement by concluding that the treatment plays no role in the patients' recovery of the injured hand wrt. the pro/sup-movement.

The analysis of the ex/flex- and uln/rad-movements will not be as extensive as the analysis of the pro/-sup movement. I will, in much less details,

- choose a model,

- determine whether $\mathbf{V}_0$, $\mathbf{R}_0$ and $\mathbf{\Sigma}$ (or $\mathbf{G}$ in the case of a RIM) are positive semi-definit,
- test whether `group` is relevant to the model, and
- test model assumptions i. and ii.

Analysing `ur`

As before, I start by testing whether I need a RIM, `ur.rim.ar`, or a RISM, `ur.rism.ar`, both set up exactly like the models for `ps`. In this case `anova()` gives a p-value indicating, that the RIM is prefered:

```
1 > anova(ur.rim.ar, ur.rism.ar)$"p-value"[2]
2 [1] 0.41452
```

Since the RIM is prefered, if there indeed is a difference in the treatment groups, I would expect the intercepts for group 3 to be higher than the intercepts for the other two groups. Figure 11 shows that this is not the case. It also shows that we do have random intercepts, but it is not entirely evident that we do not need random slopes. Out of curiosity, I checked to see how small the estimates of the random slopes were, and they were not as small as I had thought they would be.

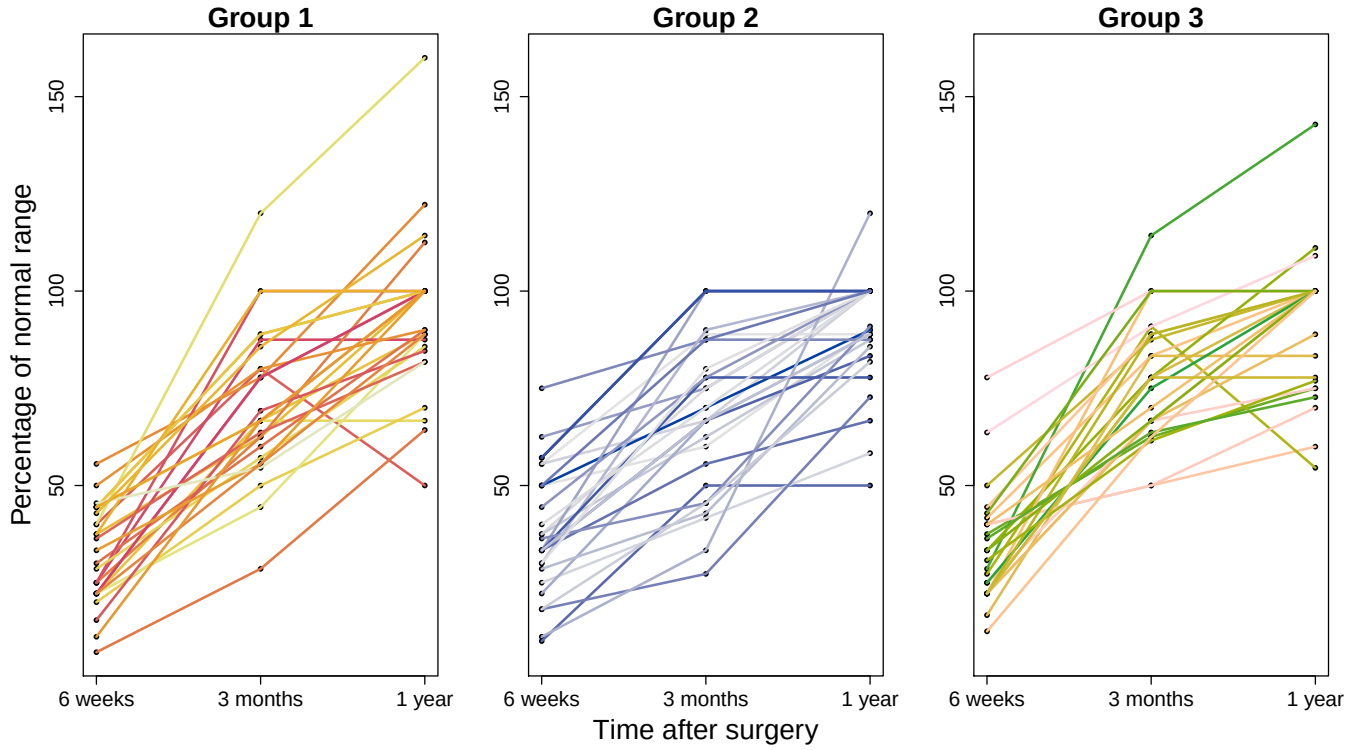

Figure 11: *In this plot only the measurements of uln/rad in percentages are used. Each graph represents a group. Each line in each graph represents the time series for a patient in that group.*

To see if I really should use `ur.rim.ar` over `ur.rism.ar`, I have compared the estimated slopes with those from `ps.rism.ar`. Let b_i^{ur} and b_i^{ps} denote these estimated slopes. Figure 12 shows b_i^{ps} and b_i^{ur}, and we do see that the range of b_i^{ps} is bigger than that of b_i^{ur}, but I am not convinced that I should accept that $b_i^{\text{ps}} \approx 0$. Therefore, I will be using `ur.rism.ar` in my analysis. If b_i^{ps} really is just zero, then I will be adding a zero-term to the model, which will make no difference. Hence, I can proceed with `ur.rism.ar`.

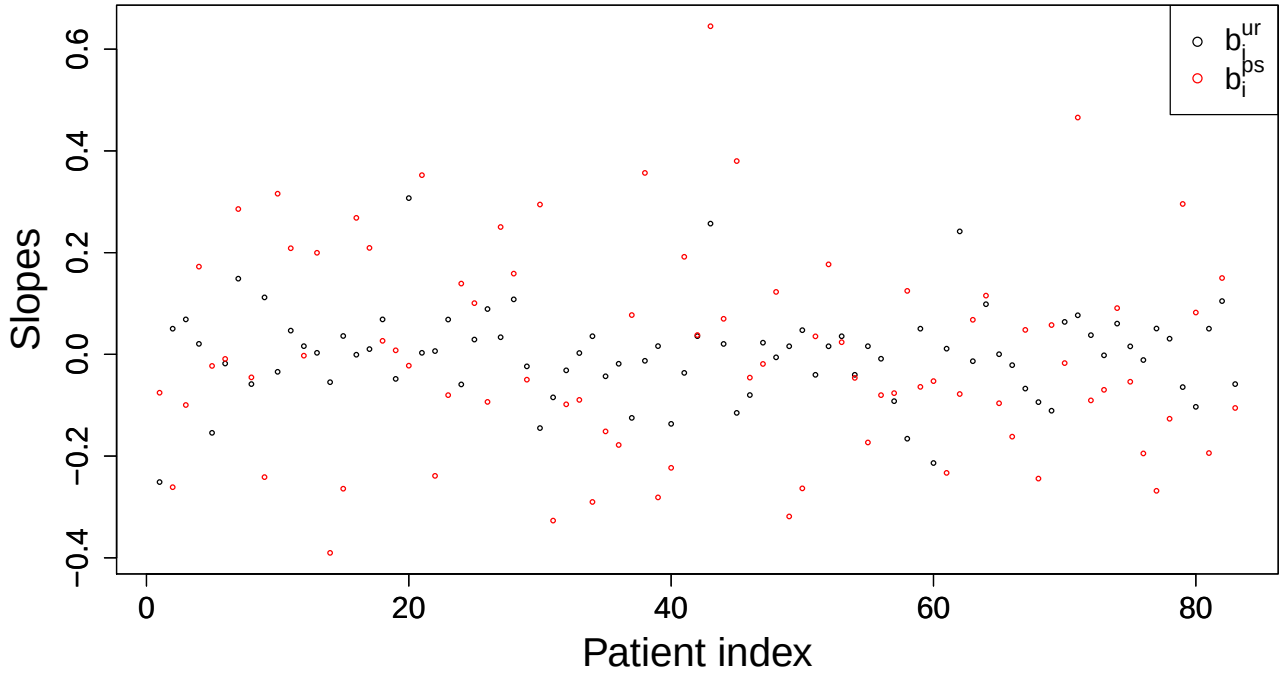

Figure 12: *The estimates slopes in* `ur.rism.ar` *and* `ps.rism.ar`.

Below are the results from `superLMM()` (I will not be showing any of the plots from `superLMM()`):

```
> superLMM(ur.rism.ar)
$vcov
        V0    R0 Sigma
[1,] TRUE TRUE  TRUE

$group$anova1
             p-value
(Intercept) 0.00000
time        0.00000
I(time^2)   0.00000
group       0.72217

$group$anova2
[1] 0.7101

$group$intervals
                lower      est.     upper
(Intercept) 28.45686  33.81199  39.16713
time         5.46868   6.10641   6.74413
I(time^2)   -0.11872  -0.10566  -0.09260
group2      -6.90004   0.07710   7.05424
group3      -4.51892   2.52253   9.56398

> superLMM(ur.rism.ar, ass = 1, is = "I")
            Bart     Leve    Shap
p-value 0.57259  0.49011 0.82730

> superLMM(ur.rism.ar, ass = 1, is = "S")
            Bart     Leve     Shap
p-value 0.40876  0.50730  0.00340
```

```
31
32 > superLMM(ur.rism.ar, ass = 2)
33              Bart      Leve      Shap
34 p-value 0.32225 0.99941 0.30036
```

All of the covariance matrices are positive semi-definit, and `group` is irrelevant to the model. All assumptions are satisfied, apart from the estimated random slopes being Gaussian (the Q-Q plot and histogram of the slopes looks very similar to those in Figure 10). I am not too concerned with violations of the normality assumption as long as the homoscedasticity assumption is not violated.

In summation: Modelling the uln/rad-movement with a RISM gives positive semi-definite covariance matrices, all tests agree that `group` is irrelevant to the model, and the model assumptions are satisfied. I conclude that the treatment plays no role in the patients' recovery of the injured hand wrt. the uln/rad-movement.

Analysing `ef`

Just like in the analysis of the uln/rad-movement, `anova()` tells me I should use a RIM for the ex/flex-movement. Once again, I see no indication in Figure 13 that there is a difference in the intercepts wrt. the groups. I have investigated the range of the estimated slopes, and once again I am not convinced that the slopes are all the same. I have also compared the slopes with those from `ps.rism.ar`, and I find that b_i^{ef} resembles b_i^{ur} in Figure 12. Hence, I will perform my analysis with a RISM, `ef.rism.ar`.

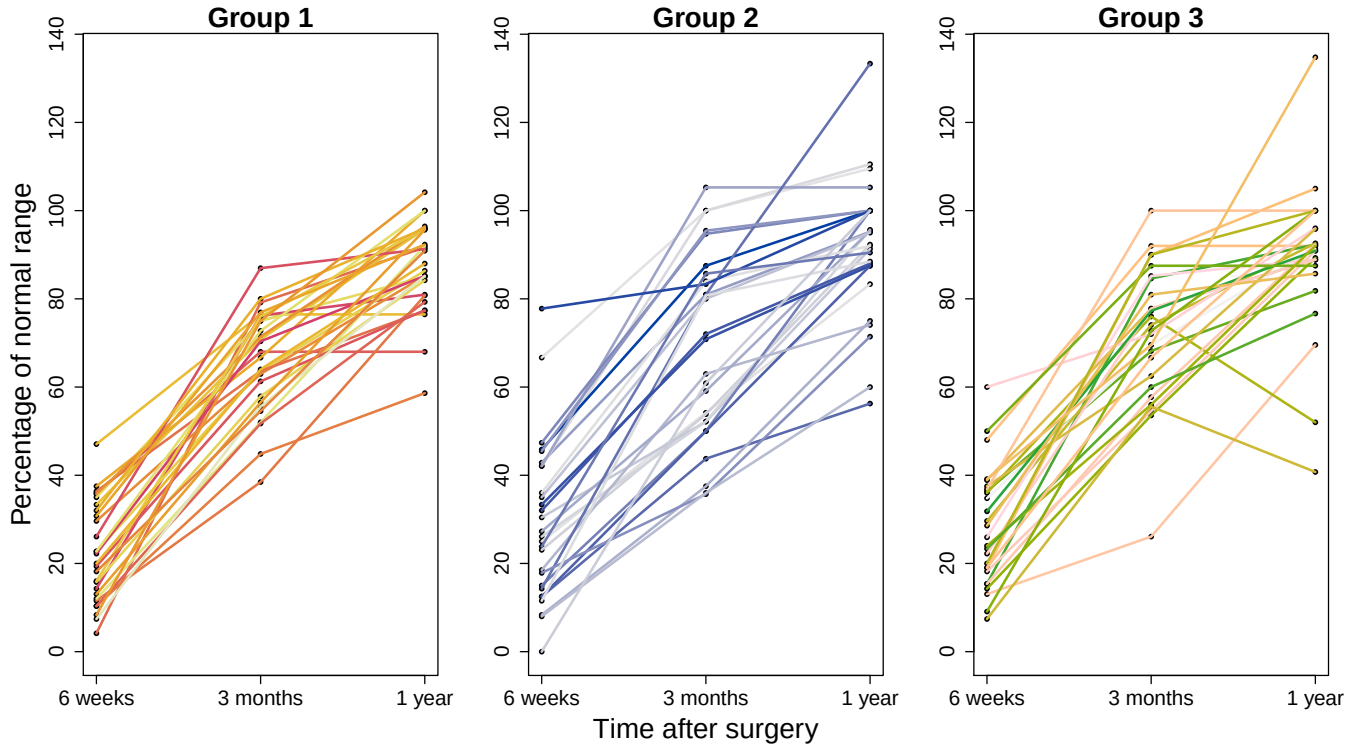

Figure 13: *In this plot only the measurements of ex/flex in percentages are used. Each graph represents a group. Each line in each graph represents the time series for a patient in that group.*

Below are the results from `superLMM()` (I will not be showing any of the plots from `superLMM()`):

```
> superLMM(ef.rism.ar)
$vcov
       V0    R0 Sigma
[1,] TRUE TRUE  TRUE

$group$anova1

            p-value
(Intercept) 0.00000
time        0.00000
I(time^2)   0.00000
group       0.24374

$group$anova2
[1] 0.23174

$group$intervals
                lower      est.     upper
(Intercept) 18.50728 23.45852 28.40975
time         6.41496  6.92821  7.44147
I(time^2)   -0.13132 -0.12083 -0.11033
group2      -1.29168  5.06211 11.41590
group3      -2.20206  4.21030 10.62265

> superLMM(ef.rism.ar, ass = 1, is = "I")
            Bart     Leve     Shap
p-value 0.00326 0.00161 0.87110

> superLMM(ef.rism.ar, ass = 1, is = "S")
            Bart     Leve     Shap
p-value 0.32826 0.41292 0.95484

> superLMM(ef.rism.ar, ass = 2)
            Bart     Leve     Shap
p-value 0.12205 0.99584 0.01414
```

All of the covariance matrices are positive semi-definit, and `group` is irrelevant to the model. All assumptions are satisfied, apart from intercepts being homoscedastic and the residuals being Gaussian. I have checked the Q-Q plot and histogram of the residuals (not shown here), and I see no reason not to assume the residuals are (asymptotically) Gaussian. The intercepts, according to both the Bartlett test and Levene's test, are heteroscedastic. The top plot in Figure 14 shows no particular patterns in the intercepts indicating heteroscedasticity, but we do see a bigger spread in the intercepts for the patients in group 2, than in the other two groups. This is confirmed by the bottom plot. Perhaps this is what results in the low p-values. As I see no unwanted patterns in the top plot, I am causious in accepting the results from the Bartlett and Levene's test.

In summation: Modelling the ex/flex-movement with a RISM gives positive semi-definite covariance matrices, all tests agree that `group` is irrelevant to the model, and the model assumptions were (almost all) satisfied. I conclude that the treatment plays no role in the patients' recovery of the injured hand wrt. any of the three types of movement.

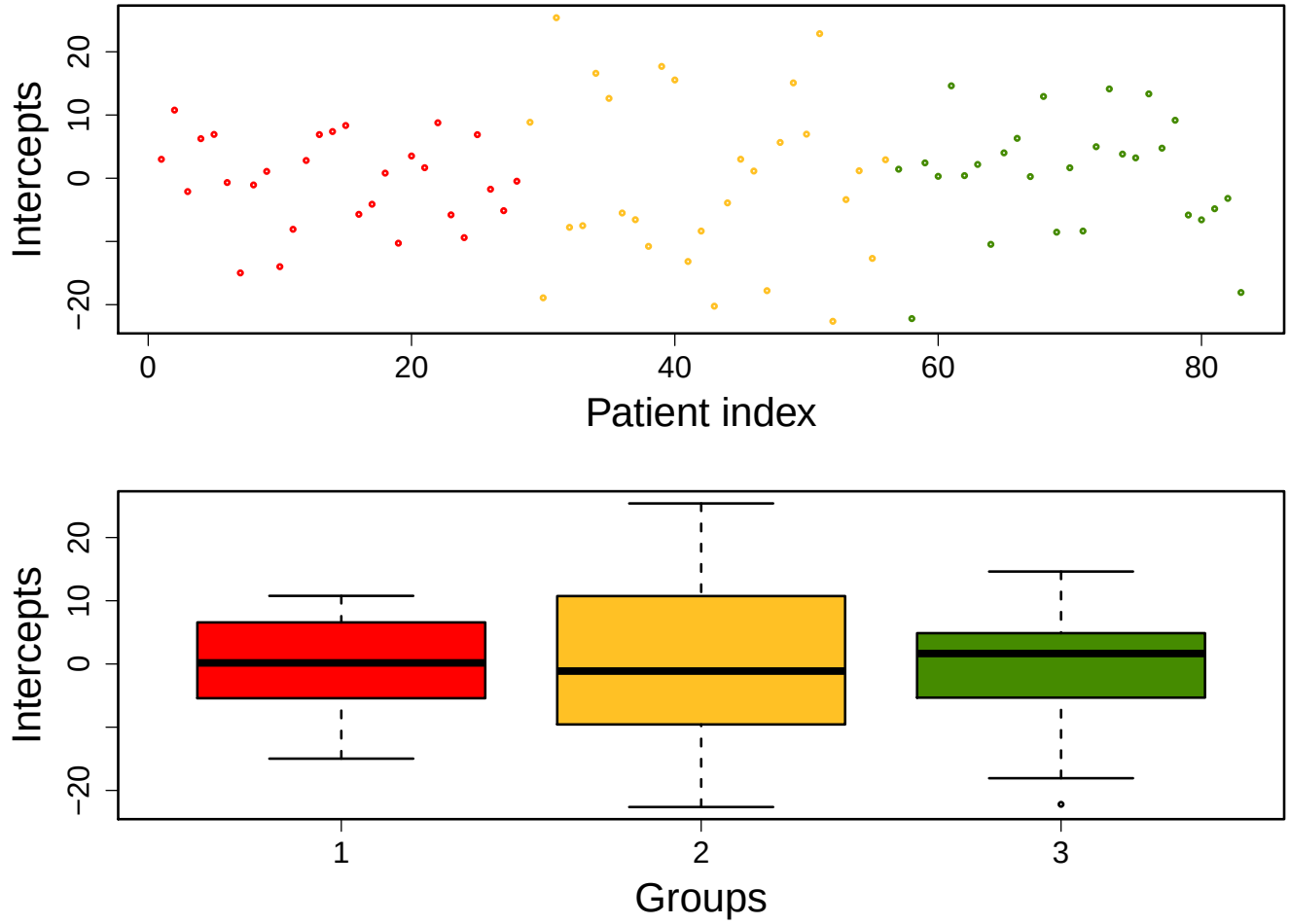

Figure 14: *1st row: Intercepts vs. patient index. The red, yellow and green dots are for groups 1, 2 and 3, respectively. 2nd row: Boxplots of the intercepts for each group.*

5.2 Analysis using `correlation = corSymm()`

So far I have assumed an autoregressive covariance structure would be the most fitting. Of my choices, the only other realistic covariance structure, is the unstructured one, I believe. In this section, I will set up the models with `correlation = corSymm()`, which gives the unstructured covariance, to see if the models with this structure fits the data better. The models are:

```
1 ps.rism.un = lme(ps ~ time + I(time^2) + group, random = ~1+time|subject, data = Wrist, method =
      "REML", correlation = corSymm(form = ~ 1|subject), control = lmeControl(opt = "optim",
    msMaxIter = 60))
2 ur.rim.un = lme(ur ~ time + I(time^2) + group, random = ~1|subject, data = Wrist, method = "REML
    ", correlation = corSymm(form = ~ 1|subject))
3 ef.rim.un = lme(ef ~ time + I(time^2) + group, random = ~1|subject, data = Wrist, method = "REML
    ", correlation = corSymm(form = ~ 1|subject))
```

I have added "un" to the names of the models for "*un*structured". This time I am relying on the results when `anova()` tells me I need a RIM for the uln/rad- and ef/flex-movements. Figure 15 shows that the estimated slopes for the RISMs of the "un"-models for the uln/rad- and ef/flex-movements are all very close to zero.

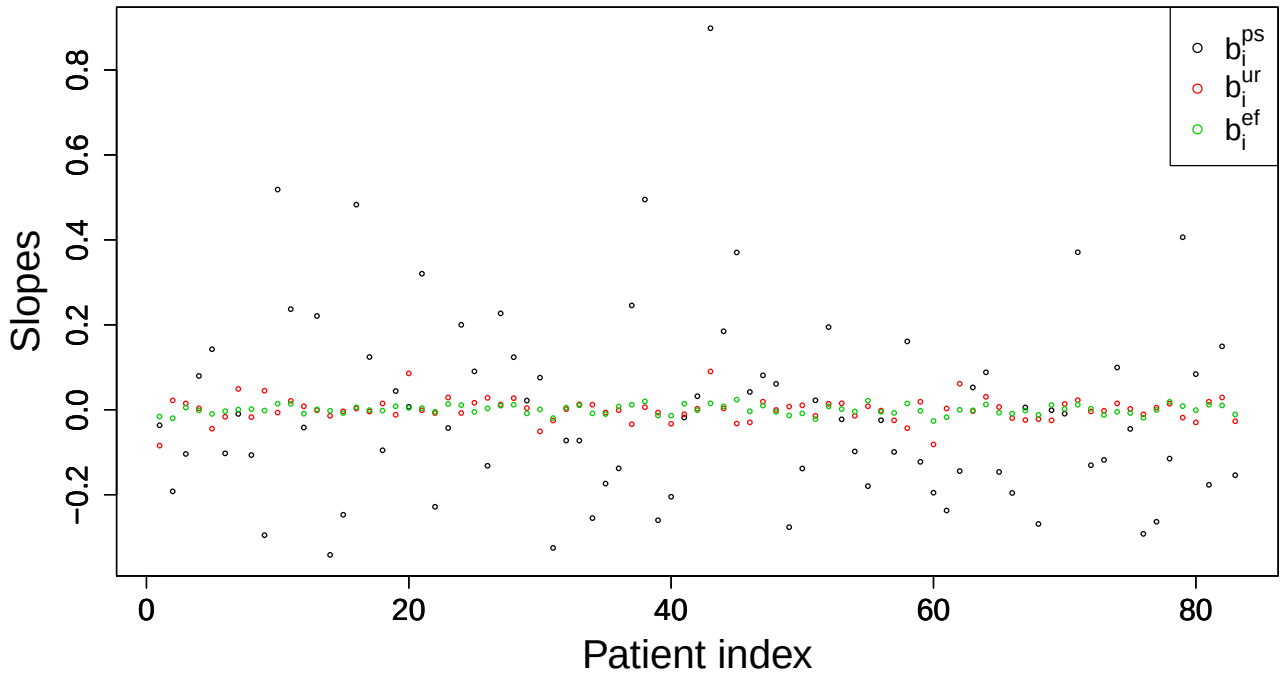

Figure 15: *The estimated slopes in the RISMs of pro/sup, uln/rad and ex/flex.*

This first thing to do is to make sure it even makes sense to use `correlation = corSymm()`, by

- determining whether $\mathbf{V}_0$, $\mathbf{R}_0$ and $\boldsymbol{\Sigma}$ (or $\mathbf{G}$) are positive semi-definite,
- testing whether `group` is relevant to the model, and
- testing model assumptions i. and ii..

The results are gathered in Table 1, 2 and 3. Table 1 shows that all covariance matrices are positive semi-definit.

Table 1: *This table contains the results of using `superLMM()$vcov` on the "un"-models.*

	`superLMM()$vcov`		
	`ps.rism.un`	`ur.rim.un`	`ef.rim.un`
$\mathbf{V}_0$	TRUE	TRUE	TRUE
$\mathbf{R}_0$	TRUE	TRUE	TRUE
$\boldsymbol{\Sigma}$	TRUE	TRUE	TRUE

In Table 2, we see that `group` is irrelevant to all the "un"-models. I have not included the results for μ, β_{time} and β_{time^2} as they are of less interest (they are significant to the models, though).

Table 2: *The row for `$anova1` are the p-values for `group`. The row for `$anova2` are the p-values from the comparison of the model and the null-model. The rows for `$intervals` are the 95%-confidence intervals for $\beta_{gr.2}$ and $\beta_{gr.3}$.*

	`superLMM()$group`		
	`ps.rism.un`	`ur.rim.un`	`ef.rim.un`
`$anova1`	0.58401	0.89186	0.19036
`$anova2`	0.58044	0.89146	0.17924
`$intervals`	$\beta_{\text{gr.2}} \in [-3.12, 5.68]$ $\beta_{\text{gr.3}} \in [-2.13, 6.76]$	$\beta_{\text{gr.2}} \in [-6.55, 6.99]$ $\beta_{\text{gr.3}} \in [-5.31, 8.36]$	$\beta_{\text{gr.2}} \in [-0.73, 11.61]$ $\beta_{\text{gr.3}} \in [-1.99, 10.46]$

In Table 3, we see that almost all model assumptions are satisfied. When testing for normality of $\hat{u}_i$, $\hat{b}_i$ and ϵ_{ij} in `ps.rism.un`, I get that none of them are Gaussian according to the Shapiro-Wilk test. Both the Q-Q plots and histograms of $\hat{u}_i$ and $\hat{b}_i$ (not shown here) seem to agree with this, but I see no reason not assume that ϵ_{ij} is Gaussian (plots also not shown here) or homoscedastic. Once again, I am contributing the low p-values from the Bartlett test to violations of normality.

Table 3: *Whether it is `ass=1,is="I"`, or `ass=1,is="S"` or `ass=2`, there are three rows. The first row are the p-values from the Bartlett test of the either the intercepts, slopes or residuals. The second row are the p-values from the Levene's test, and the third row are the p-values from the Shapiro-Wilk test.*

	`superLMM(ass = c(1, 2), is = c("I", "S"))`		
	`ps.rism.un`	`ur.rim.un`	`ef.rim.un`
`ass = 1, is = "I"`	0.19678	0.80862	0.00403
	0.44824	0.87102	0.01261
	0.00183	0.69126	0.29960
`ass = 1, is = "S"`	0.18322		
	0.59453		
	0.00016		
`ass = 2`	0.00000	0.21954	0.20516
	0.88786	0.99646	0.99925
	0.00001	0.16510	0.12081

Just like with `ef.rism.ar`, I also have problem wrt. homoscedasticity of the intercepts for `ef.rim.un`. The plots of the intercepts for `ef.rim.un` looks just about the same as Figure 14, which leaves me with no final conclusion as to whether I should accept the p-values from the Bartlett and Levenes's tests.

All in all, if I accept the results from `superLMM()$group` despite some problems with both the normality and homoscedasticity assumptions, I find that the treatment plays no role in the patients' recovery of the injured hand wrt. any of the three types of movement.

5.3 Analysis using `correlation = NULL`

I want to test my assumption that measurements on a patient must be correlated. For longitudinal data, it rarely makes sense to use the independence structure (for this, one uses `correlation = NULL`) as there will usually be some kind of correlation between measurements on the same subject. I have set up three models with independence structure ("in"-models); `ps.rism.in`, `ur.rism.in` and `ef.rism.in`, that is, three RISMs. I will not go through the analysis of these model here, but I have checked that they satisfy almost all model assumptions (with a few problems wrt. the normality of the residuals, and the homoscedasticity assumption for the intercepts of `ef.rism.in`), that the covariance matrices are positive semi-definit and that the conclusion about `group` is the same (i.e. `group` is irrelevant to the model). Having set up these models, I am now able to test whether my assumption about correlated measurements is correct.

5.4 Comparison of the "ar"-, "un"- and the "in"-models

To find out which covariance structure fits data best, I will now compare the three sets of models. In order to compare models with different covariance structure, they must both be either RIMs or RISMs.

If one is a RIM and the other is a RISM, `anova()` cannot perform an LRT and give a p-value, but only give the AIC and the value of the maximized (restricted) log-likelihoods. Of course, one could compare the AICs, and then conclude that the model with the smallest AIC is the better one. But, as mentioned in Appendix B, how big does the difference in AIC have to be before we can say that the models are different? To answer this, we will need the p-value from the test `anova()` performs. As it will not make a worse model when adding the slopes to the RIM, I will instead be comparing `ur.rism.ar` with `ur.rism.un` (I have, of course, checked that this RISM satisfy all model assumptions, that the covariance matrices are positive semi-definit and that the conclusion about `group` is the same). The same goes for the models for the ex/flex-movement. The p-values from comparing the sets of models with `anova()` are gathered in the matrix, `modcomp.anova`, below:

```
> modcomp.anova
                    ps       ur       ef
ar vs. in 0.41816 0.96648 0.91340
ar vs. un 0.00000 0.06602 0.12450
in vs. un 0.00000 0.14244 0.24263
```

The columns represent the different types of movement, and the row indicate which type of models have been compared. The results show that there is no significant difference in the three types of models. Only for the pro/sup-movement is there a slight advantage of the "un"-models. This means that for the most part, the "in"-models are just as suitable for the data as the "ar"- and "un"-models.

I have not looked at the correlations, $\hat{\rho}$, yet, but I must assume they are very close to zero. First, I will extract $\hat{\rho}$ from the "ar"-models:

```
> superLMM(ps.rism.ar)$cor.param
   Phi
0.053

> superLMM(ur.rism.ar)$cor.param
   Phi
0.197

> superLMM(ef.rism.ar)$cor.param
   Phi
0.198
```

As $|\rho| \in [0,1]$, these $\hat{\rho}$'s are sufficiently low, that the results from `modcomp.anova` seem reasonable.

Next, I will extract $\hat{\rho}_{1,2}$, $\hat{\rho}_{1,3}$ and $\hat{\rho}_{2,3}$ from the "un"-models:

```
> superLMM(ps.rism.un)$cor.param
 Correlation:
   1      2
2 -0.533
3 -0.107  0.897

> superLMM(ur.rim.un)$cor.param
 Correlation:
   1      2
2  0.047
3 -0.326  0.147
```

```
12
13 > superLMM(ef.rim.un)$cor.param
14   Correlation:
15     1       2
16 2   0.096
17 3 -0.483 -0.038
```

For `ps.rism.un`, we have that $\hat{\rho}_{1,2} = -0.533$ and $\hat{\rho}_{2,3} = 0.897$. These correlations agrees with `anova()` saying this model is prefered over `ps.rism.in`. For `ur.rim.un`, $\hat{\rho}_{1,3} = -326$ and for `ef.rim.un`, $\hat{\rho}_{1,3} = -0.483$. These correlations may be a little high compared to the results in `modcomp.anova`. I have included a note on this in my discussion, Chapter 9.

Kenward-Roger approximation

So far my conclusions about the relevance of the group-term, is that `group` is insignificant to the models. I want to do a final test about the relevance. So far I have used `anova()` for comparison of the full models and their corresponding null-models. The function `anova()` bases its comparison on an LRT. As mentioned in Section 4.4, the test statistic for the LRT has an asymptotic distribution under the null-hypothesis, but for small sample sizes the approximation of this distribution may be poor. Hence, we would rather use a test, where the test statistic is approximately F-distributed under the null-hypothesis. The solution is to use the function `KRmodcomp()` from the `pbkrtest`-package. This function performs an F-test with the Kenward-Roger approximation. In order to use `KRmodcomp()`, the models must be fitted with the `lmer`-function, which only has the exchangeable covariance structure implemented. Hence, in order to use the Kenward-Roger approximation, I must make sure that a model with exchangeable covariance structure is realistic for my data. I will not show any of the results from using exchangeable covariance structure in `lme()`, but I have set up models with this structure (the "ex"-models), and I found that they resemble all the previous models. Hence, the "ex"-models are just as good any of the other models. I can now move on to setting up the models with `lmer()`. The models are:

```
1 ps.lmer.rism = lmer(ps ~ group + time + I(time^2) + (time|subject), data = Wrist, REML = T)
2 ur.lmer.rim = lmer(ur ~ group + time + I(time^2) + (1|subject), data = Wrist, REML = T)
3 ef.lmer.rim = lmer(ef ~ group + time + I(time^2) + (1|subject), data = Wrist, REML = T)
```

For the random effects, I specify either "`(time|subject)`" or "`(1|subject)`" for a RISM or a RIM, respectively. I will test the relevance of `group` both with `anova()` and `KRmodcomp()` to see how big the differences are between the results of the two functions. The results are:

```
1 > anova(update(ps.lmer.rism, .~.-group), ps.lmer.rism)$"Pr(>Chisq)"[2]
2 [1] 0.34885
3 > KRmodcomp(largeModel=ps.lmer.rism, smallModel=update(ps.lmer.rism, .~.-group))$test$p.value[1]
4 [1] 0.36675
5
6 > anova(update(ur.lmer.rim, .~.-group), ur.lmer.rim)$"Pr(>Chisq)"[2]
7 [1] 0.72841
8 > KRmodcomp(largeModel=ur.lmer.rim, smallModel=update(ur.lmer.rim, .~.-group))$test$p.value[1]
9 [1] 0.7368
10
11 > anova(update(ef.lmer.rim, .~.-group), ef.lmer.rim)$"Pr(>Chisq)"[2]
12 [1] 0.21381
13 > KRmodcomp(largeModel=ef.lmer.rim, smallModel=update(ef.lmer.rim, .~.-group))$test$p.value[1]
14 [1] 0.22607
```

All tests agree, that `group` is insignificant to the models. Furthermore, notice that the p-values from `anova()` are very similar to the p-values from `KRmodcomp()`. This tells us, that we can trust the results from `anova()` even though they are based on asymptotic results.

5.5 Conclusion

No matter the covariance structure, there were problems wrt. the normality assumption of either the random intercepts, slopes or the residuals, and some problems wrt. the homoscedasticity assumption for the models of the ex/flex-movement. Assuming these problems could be solved from having a much larger sample to analyse, I end this chapter with the conclusion that there is no evidence to reject H_0^{main}. But perhaps it would be better to use a model that does not require having to satisfy any assumptions. In the following chapter, I present a way to do this by using *generalized estimating equations*, and in Chapter 7, the results of this approach are presented.

5.6 Source code: `superLMM()`

```
superLMM = function(mod, ass = NULL, is = NULL){
  par(mar = c(4.5,4.5,2,1), lwd = 2, cex.lab = 2, cex.axis = 1.5, cex.main = 2)

  ## Positive semi-definit
  V0 = getVarCov(mod, type = "marginal", individual = 1)$id01
  R0 = getVarCov(mod, type = "conditional", individual = 1)$id01
  Sigma = getVarCov(mod, type = "random.effects")
  clist = list(V0, R0, Sigma)
  psd = c()
  for(i in 1:3){
    if(matrixcalc::is.symmetric.matrix(clist[[i]])){ psd[i] = matrixcalc::is.positive.semi.
    definite(clist[[i]]) }
    else if(!matrixcalc::is.symmetric.matrix(clist[[i]])){
      X = (t(clist[[i]]) + clist[[i]]) / 2
      psd[i] = matrixcalc::is.positive.semi.definite(X)
    }
  }
  m.vcov = matrix(psd, ncol = 3); colnames(m.vcov) = c("V0", "R0", "Sigma")

  ## Relevance of "group"
  a1 = matrix(c(anova(mod)$"p-value"), ncol = 1, nrow = 4)
  rownames(a1) = rownames(anova(mod)); colnames(a1) = "p-value"
  a2 = anova(update(mod, .~.-group, method = "ML"), update(mod, method = "ML"))$"p-value"[2]
  int = intervals(mod, which = "fixed")[1]$fixed[,1:3]

  ## Association parameters
  cor.param = mod$modelStruct$corStruct

  ## Assumptions
  ass.matrix = matrix(c(0,0,0), ncol = 3)
  colnames(ass.matrix) = c("Bart", "Leve", "Shap"); rownames(ass.matrix) = "p-value"
  rism = dim(mod$coefficients$random$subject)[2]
  if(!is.null(ass)){
    if(ass == 1){
      if((rism == 1) || (rism == 2 && is == "I")){
        layout(matrix(c(1,1,2,3), 2, 2, byrow = T))
        col = c(rep(2,28), rep("goldenrod1", 28), rep("chartreuse4", 27))
```

```
37          plot(mod$coefficients$random$subject[,1], col = col, cex = 0.5, ylab = "Intercepts",
        xlab = "Patient␣index")
38          car::qqp(mod$coefficients$random$subject[,1], ylab = "Intercepts", xlab = "Norm␣
        quantiles", main = "␣", cex = 0.5)
39          hist(mod$coefficients$random$subject[,1], probability = T, 50, xlab = "Intercepts", main
         = "␣")
40          lines(density(mod$coefficients$random$subject[,1]), lwd = lwd, col = 2)
41
42          ass.matrix[1,] = c(bartlett.test(mod$coefficients$random$subject[,1], Wrist$group[which(
        Wrist$time==0)])$p.value,
43                             car::leveneTest(mod$coefficients$random$subject[,1], Wrist$group[
        which(Wrist$time==0)])$"Pr(>F)"[1],
44                             shapiro.test(mod$coefficients$random$subject[,1])$p.value)
45        }
46        if(rism == 2){
47          if(is == "S"){
48            layout(matrix(c(1,1,2,3), 2, 2, byrow = T))
49            col = c(rep(2,28), rep("goldenrod1", 28), rep("chartreuse4", 27))
50            plot(mod$coefficients$random$subject[,2], col = col, cex = 0.5, ylab = "Slopes", xlab
        = "Patient␣index")
51            car::qqp(mod$coefficients$random$subject[,2], ylab = "Slopes", xlab = "Norm␣quantiles"
        , main = "␣", cex = 0.5)
52            hist(mod$coefficients$random$subject[,2], probability = T, 50, xlab = "Slopes", main =
         "␣")
53            lines(density(mod$coefficients$random$subject[,2]), lwd = lwd, col = 2)
54
55            ass.matrix[1,] = c(bartlett.test(mod$coefficients$random$subject[,2], Wrist$group[
        which(Wrist$time==0)])$p.value,
56                               car::leveneTest(mod$coefficients$random$subject[,2], Wrist$group[
        which(Wrist$time==0)])$"Pr(>F)"[1],
57                               shapiro.test(mod$coefficients$random$subject[,2])$p.value)
58          }
59        }
60      }
61
62      if(ass == 2){
63        layout(matrix(c(1,1,2,3), 2, 2, byrow = T))
64        col2 = c(rep(2,3*28), rep("goldenrod1",3*28), rep("chartreuse4",3*27))
65        plot(mod$residuals[,2], ylab = "Residuals", main = "␣", cex = 0.5)
66        for(i in seq(1,249,3)){ lines(c(i:(i+2)),mod$residuals[i:(i+2),2], col = col2[i]) }
67        car::qqp(mod$residuals[,2], ylab = "Residuals", xlab = "Norm␣quantiles", main = "␣", cex =
         0.5)
68        hist(mod$residuals[,2], probability = T, 50, xlab = "Residuals", main = "␣")
69        lines(density(mod$residuals[,2]), col = 2)
70
71        ass.matrix[1,] = c(bartlett.test(mod$residuals[,2], Wrist$subject)$p.value,
72                           car::leveneTest(mod$residuals[,2], Wrist$subject)$"Pr(>F)"[1],
73                           shapiro.test(mod$residuals[,2])$p.value)
74      }
75    }
76
77    if(is.null(ass)){ out = list(vcov = m.vcov, group = list(anova1 = a1, anova2 = a2, intervals =
        int), cor.param = cor.param); out }
78    else
79      print(ass.matrix)
80  }
```

6 | Generalized Estimating Equations

Generalized estimating equations models (GEE models for short) are an extension of the generalized linear models (GLMs for short). GLMs are fixed effects models, which assumes all observations are independent of each other. Hence, GLMs are not suited for longitudinal data in most cases. Extending the GLM by specifying a correlation structure among observations on the same subject, we get the GEE model. The correlation structure is equivalent to $\mathbf{R}_0$ in the LMM. In Chapter 2, when testing the assumptions on the data, I had problems identifying the data as Gaussian. In Chapter 4, I worked under the assumption, that the data was Gaussian as it seemed the most fitting distribution. Perhaps it would be better not to make any such assumptions. An advantage of the GEE model over the LMM is that it uses quasi-likelihood. This means that the full likelihood of the data is not specified, hence no assumptions on the distribution of the data are necessary.

To understand the setup in the GEE model, it is useful to review GLMs. I will therefore start this chapter off with a quick run-through of the GLM before presenting the GEE model.

6.1 Generalized linear models

This section is based on [2], [7] and [12].

The GLM is an extension of the general linear model (LM for short). For a specific *link function* and *variance function*, the GLM reduces to an LM. The assumptions in the LM are, that the response is Gaussian, the relationship between the response and the covariates is linear, and the variance of the response does not depend on the mean. In the GLM, the response does not have to be Gaussian. It just has to have a distribution from the exponential family. The most popular distributions in the exponential family are the Gaussian, Binomial, Poisson and Gamma distributions. And the variance is actually a function of the mean. No matter the distribution, the density of the response can be written in the following form:

Definition 6.1 (Density for the exponential (dispersion) family) Let $\mathbf{Y} = \begin{bmatrix} Y_1 & \cdots & Y_N \end{bmatrix}^T$ be a random vector, where the Y_i's are iid. with a distribution from the exponential family. The marginal densities are then

$$f(y_i;\theta_i,\phi) = h\left(y_i,\phi\right)\exp\left(\frac{y_i\theta_i - b(\theta_i)}{\phi}\right), \tag{6.1}$$

where ϕ is called the *dispersion parameter*, which may be known or unknown, θ is called the *canonical parameter*, and $h(\cdot)$ and $b(\cdot)$ are functions that depend on the distribution.

Example 2 shows how the density from a Binomially distributed random variable can be written in the form (6.1).

Example 2 In this example, I will show how a probability mass function (the discrete equivalent of the density function) can be written in the form (6.1). Let X be a Binomially distributed random variable, i.e. $X \sim Bin(n,\pi)$. The probability mass function for X is defined as

$$p(x) = \binom{n}{x}\pi^x(1-\pi)^{n-x}, \quad x = 0,1,\ldots,n.$$

We can ignore the subscripts in Equation (6.1), as X is a univariate random variable.

Rewritting $p(x)$, it becomes

$$\begin{aligned} p(x) &= \binom{n}{x} \exp\left(\log\left(\pi^x (1-\pi)^{n-x}\right)\right) \;=\; \binom{n}{x} \exp\left(x\log\pi + (n-x)\log(1-\pi)\right) \\ &= \binom{n}{x} \exp\left(x\left(\log(\pi) - \log(1-\pi)\right) + n\log(1-\pi)\right) \;=\; \binom{n}{x} \exp\left(x\log\left(\frac{\pi}{1-\pi}\right) + n\log(1-\pi)\right) \\ &= h(x)\exp\left(x\theta - b(\theta)\right), \end{aligned}$$

where $h(x) = \binom{n}{x}$, $\theta = \log\left(\frac{\pi}{1-\pi}\right)$, $b(\theta) = -n\log(1-\pi)$, and where ϕ is ignored as $\phi = 1$ for this distribution.

□

Linear predictor, link function and variance function

Three things are needed in specifying the GLM; the *linear predictor*, the *link function* and the *variance function*. The linear predictor, η_i, is a linear combination of the covariates

$$\eta_i = \mathbf{x}_i^T \boldsymbol{\beta},$$

where $\mathbf{x}_i$ is the vector of explanatory variables for subject i. The linear predictor is independent of the distribution of the data. Let $\mu_i = \mathrm{E}[Y_i]$. The link function, $g(\cdot)$, is

$$g(\mu_i) = \eta_i.$$

Thus, the mean value is not a linear combination of the covariates (like we have with the general linear model), but instead it is assumed that the mean value is a *function* of the linear predictor, i.e.

$$\mu_i = g(\eta_i)^{-1} = g(\mathbf{x}_i^T \boldsymbol{\beta})^{-1}.$$

Both μ_i and η_i are functions of $\boldsymbol{\beta}$, and one ought to write $\mu_i(\boldsymbol{\beta})$ and $\eta_i(\boldsymbol{\beta})$ to emphasize this. For the ease of notation, I will not do that. Each distribution has its own canonical link function shown below:

Distribution	Gaussian	Binomial	Poisson	Gamma
$g(\mu_i)$	μ_i	$\mathrm{logit}(\pi)$	$\log(\mu_i)$	$\frac{1}{\mu_i}$

The link function specifies how the covariates relate to the mean of the response. When working with, say, Gaussian data one is not restricted to only using $g(\mu_i) = \mu_i$. One can choose which ever link function that seems suitable for the data. Hence, the link function cannot be thought of as being dependent on the distribution of the data.

To find the variance function, we will need the mean and variance of Y_i:

$$\mathrm{E}[Y_i] = \mu_i \overset{\text{A.20}}{=} b'(\theta_i) \quad \text{and} \quad \mathrm{Var}[Y_i] \overset{\text{A.20}}{=} \phi b''(\theta_i).$$

Isolating θ_i in the expression of $\mathrm{E}[Y_i]$, we get

$$\theta_i = b'(\mu_i)^{-1}. \tag{6.2}$$

Inserting this into $\mathrm{Var}[Y_i]$, we see how the variance becomes a function of the mean:

$$\mathrm{Var}[Y_i] = \phi b''\left(b'(\mu_i)^{-1}\right) = \phi V(\mu_i), \tag{6.3}$$

where $V(\mu_i)$ is the variance function. Each distribution has its own variance function shown below:

Distribution	Gaussian	Binomial	Poisson	Gamma
$V(\mu_i)$	1	$\pi(1-\pi)$	μ_i	μ_i^2

The choice of variance function specifies the distribution of the data. This means that, unlike $g(\cdot)$, one does not have the freedom of choice. The GLM reduces to an LM if $V(\mu_i) = 1$ and $g(\mu_i) = \mu_i$. The dispersion parameter accounts for any excess variation not included by $V(\cdot)$. For instance, for a random variable $X \sim N(\mu, \sigma^2)$, as $V(\mu) = 1$, this means that ϕ must be σ^2 to account for the variation in X.

Example 3 In this example, I will show, that $\mathrm{E}[Y_i] = b'(\theta_i)$ and $\mathrm{Var}[Y_i] = \phi b''(\theta_i)$ agrees with the mean and variance of the random variable from Example 2. For $X \sim Bin(n,\pi)$, the mean and variance are $\mathrm{E}[X] = n\pi$ and $\mathrm{Var}[X] = n\pi(1-\pi)$. In Example 2, $b(\theta) = -n\log(1-\pi)$ and $\theta = \log\left(\frac{\pi}{1-\pi}\right)$. In order to take the derivative of $b(\cdot)$ wrt. θ, I will need to write $b(\cdot)$ with θ featured in its expression. I will do this by isolation π in the expression for θ:

$$\begin{aligned}\theta = \log\left(\frac{\pi}{1-\pi}\right) &\Leftrightarrow \frac{\pi}{1-\pi} = \exp(\theta) \Leftrightarrow \pi = \exp(\theta) - \pi\exp(\theta) \Leftrightarrow \pi\left(1+\exp(\theta)\right) = \exp(\theta) \\ &\Leftrightarrow \pi = \frac{\exp(\theta)}{1+\exp(\theta)}.\end{aligned}$$

I will now show that $b'(\theta) = \mathrm{E}[X]$:

$$\begin{aligned}b'(\theta) = -\frac{d}{d\theta} n\log\left(1 - \frac{\exp(\theta)}{1+\exp(\theta)}\right) &= -\frac{d}{d\theta} n\log\left(\frac{1}{1+\exp(\theta)}\right) = \frac{d}{d\theta} n\log\left(1+\exp(\theta)\right) \\ = n\frac{1}{1+\exp(\theta)}\exp(\theta) &= n\pi = \mathrm{E}[X].\end{aligned}$$

Next, I will show that $b''(\theta) = \mathrm{Var}[X]$:

$$\begin{aligned}b''(\theta) = \frac{d}{d\theta} b'(\theta) &= \frac{d}{d\theta} n\frac{\exp(\theta)}{1+\exp(\theta)} = n\frac{\exp(\theta)\left(1+\exp(\theta)\right) - \exp(\theta)^2}{\left(1+\exp(\theta)\right)^2} \\ = n\frac{\exp(\theta)}{1+\exp(\theta)}\left(\frac{1+\exp(\theta)}{1+\exp(\theta)} - \frac{\exp(\theta)}{1+\exp(\theta)}\right) &= n\pi(1-\pi) = \mathrm{Var}[X].\end{aligned}$$

□

Estimating equations

Having specified the linear predictor, the link function, and the variance function, one can estimate the coefficients, $\boldsymbol{\beta}$, by solving an estimating equation. Let ℓ_i and S_i be the log-likelihood and score function, respectively, of Equation (6.1) for the ith subject. By the chain rule, we have that

$$S_i = \frac{\partial \ell_i}{\partial \mu_i} = \frac{\partial \ell_i}{\partial \theta_i} \times \frac{\partial \theta_i}{\partial \mu_i} \overset{A.21}{=} \frac{y_i - b'(\theta_i)}{\phi V(\mu_i)}. \tag{6.4}$$

Thus

$$\frac{\partial \ell_i}{\partial \beta_j} = \frac{\partial \ell_i}{\partial \mu_i} \times \frac{\partial \mu_i}{\partial \beta_j} \overset{(6.4)}{=} \frac{y_i - b'(\theta_i)}{\phi V(\mu_i)} \times \frac{\partial \mu_i}{\partial \beta_j}.$$

The estimating equations thus becomes

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{N} \frac{\partial \ell_i}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{N} \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}}\right)^T \frac{y_i - b'(\theta_i)}{\phi V(\mu_i)} = \sum_{i=1}^{N} \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}}\right)^T \mathrm{Var}[Y_i]^{-1}(y_i - \mu_i) = \mathbf{0}. \tag{6.5}$$

Note that $S(\boldsymbol{\beta})$ was derived using the density of the data. If the distribition is unknown, the density cannot be defined, and if the distribution does not belong to the exponential family, the likelihood function cannot be defined by Equation (6.1). Thus, the GLM fails as a possible approach for modelling the data.

I will now present a different approach of estimating $\boldsymbol{\beta}$, that does not need the density or knowledge of the distribution of the data. Instead of S_i, we will define a function that satisfy all the same properties as S_i, i.e.

$$\mathrm{E}[S_i] = 0 \quad \text{and} \quad \mathrm{Var}[S_i] = -\mathrm{E}\left[\frac{\partial S_i}{\partial \mu_i}\right].$$

Let that function be

$$q_i = q(\mu_i, y_i) = \frac{y_i - \mu_i}{\phi V(\mu_i)}.$$

It is shown in A.22, that q_i satisfy the properties. As for the expression for q_i, it is chosen because it is equal to the expression for S_i (Equation (6.4)) with the important distinction that $b'(\theta_i)$ requires knowledge of the distribution of y_i in order to be calculated, whereas μ_i in q_i does not. As q_i mimics a proper score function, the integral of q_i will mimic a proper log-likelihood. The *log quasi-likelihood function* for subject i is thus defined as

$$Q_i = Q(\mu_i, y_i) = \int_{y_i}^{\mu_i} \frac{y_i - t_i}{\phi V(t_i)}\, dt_i.$$

The log quasi-likelihood for all subjects is

$$Q = Q(\boldsymbol{\mu}, \mathbf{y}) = \sum_{i=1}^{N} Q_i = \sum_{i=1}^{N} \int_{y_i}^{\mu_i} \frac{y_i - t_i}{\phi V(t_i)}\, dt_i.$$

Differentiating Q_i wrt. μ_i of course just gives q_i. As $g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}$ does not depend on a specific distribution, we can differentiate Q_i wrt. $\boldsymbol{\beta}$:

$$\frac{\partial Q_i}{\partial \boldsymbol{\beta}} = \frac{\partial Q_i}{\partial \mu_i} \times \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} = \frac{y_i - \mu_i}{\phi V(\mu_i)} \times \frac{\partial \mu_i}{\partial \boldsymbol{\beta}}.$$

We now get new estimating equations (also called *maximum quasi-likelihood equations*):

$$S(\boldsymbol{\beta}) = \frac{\partial Q}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{N} \frac{\partial Q_i}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{N} \frac{y_i - \mu_i}{\phi V(\mu_i)} \times \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{N} \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}}\right)^T \mathrm{Var}[Y_i]^{-1}(y_i - \mu_i) = \mathbf{0}. \tag{6.6}$$

Equations (6.5) and (6.6) may look identical, but they are not in the sense that they are derived in different ways. Equation (6.5) can only be derived when the density is know and the distribution belongs to the exponential family. Equation (6.6) can be derived regardless of the distribution of data or whether the density is known; we only need to specify how the mean depend on the covariates through the link function, and how the variance variates as a function of the mean, i.e. the variance function, in order to estimate $\boldsymbol{\beta}$. Solution of Equation (6.6) are called *quasi-likelihood* estimates.

6.2 GEE models

This section is based on [2], [12] and [13].

In a data set with more than one observation per subject, it would be very likely that observations on the same subject are somehow correlated. Nothing in the specification of the GLM takes this into account. Hence, a GLM with multiple responses per subject will only be a sensible model in the case, that no observations on the same subject are correlated. The GEE model[18] is an extension of the GLM that *does* take into account observations on the same subject possibly being correlated. The GEE model uses a quasi-likelihood approach and thus does not require a specification of the underlying distribution of the response. The setup resembles the GLM, only now we have a multidimensional response per subject instead of a univariate response, and in addition to the linear predictor, the link function and the variance function, we also have to specify a correlation structure.

Let $\mathbf{y}_i$ be the vector of responses from the ith subject, where the response, y_{ij}, is taken at timepoint j. The linear predictor, link function and variance function are

$$\begin{aligned} \eta_{ij} &= \mathbf{x}_{ij}^T \boldsymbol{\beta}, \\ g(\mu_{ij}) &= \eta_{ij}, \\ \mathrm{Var}\left[Y_{ij}\right] &= \phi V(\mu_{ij}), \end{aligned} \tag{6.7}$$

where $\mathbf{x}_{ij}$ is the vector of covariates for subject i at timepoint j. Additionally, the correlation structure of the repeated measurements must be specified. It is assumed that the number of timepoints is fixed. Let that number be n. Subjects do not need to have been measured at all n timepoints, but for the sake of simplicity (and because it is the case with my data), I will assume they were. The correlation matrix for subject i is denoted by $\mathbf{R}_i$, and it is assumed it depends on a vector of association parameters, $\mathbf{a}$, hence we write $\mathbf{R}_i(\mathbf{a})$. Just like in the LMM, we have that all subjects have the same covariance matrix, hence we can write $\mathbf{R}_i(\mathbf{a}) = \mathbf{R}_0(\mathbf{a})$. The association parameters are the same for all subjects. The correlation structure can be expressed in several ways that the reader may recognize from Section 4.3:

- Independence: $\mathbf{R}_0(\mathbf{a}) = \mathbf{I}_{n\times n}$. This rarely makes sense for longitudinal data, however.
- Exchangeable: $\mathbf{R}_0(\mathbf{a})_{jk} = \rho_{jk} = \begin{cases} 1, & j = k \\ a, & j \neq k \end{cases}$.
- Autoregression: $\mathbf{R}_0(\mathbf{a})_{jk} = \rho_{jk} = a^{|j-k|}$, $a \in]0,1[$. This is usually the most popular choice for longitudinal data.
- Unstructured: $\mathbf{R}_0(\mathbf{a})_{jk} = \rho_{jk} = a_{jk}$. This is the most efficient choice when n is a small number.

Other choices are also available. Note, that if $\mathbf{R}_0(\mathbf{a}) = \mathbf{I}_{n\times n}$, the GEE models assumes that there is no within-subject correlation and thus reduces to a GLM.

6.2.1 GEE estimation of parameters

Let $\mathbf{A}_i = \mathrm{diag}\left\{\sqrt{V(\mu_{ij})}\right\}_{j=1,\ldots,n}$. The covariance matrix for $\mathbf{y}_i$ is defined as

$$\mathbf{V}_i(\mathbf{a}) = \phi \mathbf{A}_i \mathbf{R}_0(\mathbf{a}) \mathbf{A}_i. \tag{6.8}$$

[18] The term "model" is used a little loosely when speaking of the GEE model. As it uses a quasi-likelihood and not a proper likelihood, we do not actually have any model. Nevertheless, for simplicity, I will call it a model.

The coefficients $\boldsymbol{\beta}$ are estimated by solving the estimating equations

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{N} \frac{\partial \boldsymbol{\mu}_i^T}{\partial \boldsymbol{\beta}} \mathbf{V}_i(\mathbf{a})^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i) = \mathbf{0}. \tag{6.9}$$

In order to solve Equation (6.9), we must find $\mathbf{V}_i(\mathbf{a})$, meaning we must find $\mathbf{a}$ and ϕ. These parameters are unknown and must be estimated.

Finding a

Let

$$r_{jk} = \frac{(Y_{ij} - \mu_{ij})(Y_{ik} - \mu_{ik})}{\phi\sqrt{V(\mu_{ij})V(\mu_{ik})}}, \quad \forall i$$

which is a function of $\boldsymbol{\beta}$ and has mean

$$\mathrm{E}\left[r_{jk}\right] \overset{\text{A.23}}{=} \rho_{jk} \;=\; \mathrm{Corr}\left[Y_{ij}, Y_{ik}\right].$$

Gathered in one vector, we have $\mathbf{r} = \left[r_{12}, r_{13}, \ldots, r_{n-1,n}\right] \in \mathbb{R}^{\frac{n(n-1)}{2} \times 1}$. Let

$$\boldsymbol{\rho}(\mathbf{a}) = \mathrm{E}\left[\mathbf{r}\right] \;=\; \left[\rho_{12}, \rho_{13}, \ldots, \rho_{n-1,n}\right],$$

which is a function of $\mathbf{a}$ because ρ_{jk} is a function of a. Then $\mathbf{a}$ is estimated by the estimating equations

$$\frac{\partial \boldsymbol{\rho}(\mathbf{a})}{\partial \mathbf{a}} \mathbf{W}^{-1}\left(\mathbf{r} - \boldsymbol{\rho}(\mathbf{a})\right) = \mathbf{0}, \tag{6.10}$$

where $\mathbf{W}$ is the covariance matrix for $\mathbf{r}$, typically specified as $\mathrm{diag}\left\{\mathrm{Var}\left[r_{jk}\right]\right\}$. Notice how the form of this estimating equation is the same as in Equation (6.9); a product between the derivative of the mean, the covariance function and the raw residuals.

Finding ϕ

Isolating ϕ in Equation (6.7), we get

$$\phi \overset{(6.7)}{=} \frac{\mathrm{Var}\left[Y_{ij}\right]}{V(\mu_{ij})} \;=\; \frac{\mathrm{E}\left[(Y_{ij} - \mu_{ij})^2\right]}{V(\mu_{ij})}.$$

When μ_{ij} is estimated, we can insert it to get an estimate of ϕ. But, the estimate for ϕ must hold for all observations. Hence, if we calculate the Pearson residuals

$$\varepsilon_{ij} = \frac{y_{ij} - \hat{\mu}_{ij}}{\sqrt{\mathbf{V}_i(\hat{\mathbf{a}})_{jj}}},$$

and sum over these corrected for the degrees of freedom, we get the estimate of ϕ

$$\hat{\phi} = \frac{1}{N-p} \sum_{i=1}^{N} \sum_{j=1}^{n} \varepsilon_{ij}^2. \tag{6.11}$$

Solving Equation (6.9) wrt. $\boldsymbol{\beta}$ means that $\hat{\boldsymbol{\beta}}$ is dependent on the current estimate of $\mathbf{a}$. As $\mathbf{r}$ is a function of $\boldsymbol{\beta}$, $\hat{\mathbf{a}}$ is dependent on the current estimate of $\boldsymbol{\beta}$. This means that the GEE estimator of $\boldsymbol{\beta}$ is found through an iterative algorithm:

Algorithm 6.2 (Algorithm for the GEE estimator)

i. Initialize $\hat{\boldsymbol{\beta}}$, e.g. as the coefficients from a GLM.

ii. Using $\hat{\boldsymbol{\beta}}$, estimate $\mathbf{R}_0(\mathbf{a})$ by inserting $\hat{\mathbf{a}}$ estimated from solving Equation (6.10), and estimate ϕ by Equation (6.11).

iii. Using $\hat{\mathbf{a}}$ and $\hat{\phi}$, estimate $\mathbf{V}_i(\mathbf{a})$ by Equation (6.8).

iv. Using $\mathbf{V}_i(\hat{\mathbf{a}})$, estimate $\boldsymbol{\beta}$ by solving Equation (6.9).

Repeat step ii.-iv. until convergence.

I want to know whether or not it is a garantee, that Algorithm 6.2 converges, and if different initial $\hat{\boldsymbol{\beta}}$'s will lead to the same final GEE estimator. Unfortunately, I have found no sources that says anything about this matter, so perhaps it is simply unknown. The function `geeglm()` from the `geepack`-package is one of the functions one might use for correlated GLM-type data, such as `Wrist`. I have looked at the source code for `geeglm()` thinking that if non-convergence was a possibility, then it must be implemented in the function, that the iterations should stop when reaching a certain limit. No such thing was implemented. This leads me to think, that convergence is always garanteed, but I cannot be sure. In `geeglm()`, it is implemented that the initial $\hat{\boldsymbol{\beta}}$ is the coefficients from modelling the data with a GLM, just like step i. in Algorithm 6.2. I believe that the initial $\hat{\boldsymbol{\beta}}$ is chosen like this as it will lower the number of iterations, because the coefficients from the GLM usually are not that far from the GEE estimator. Coincidentally, unrelated to this report, I have coded a function called `my_gee()` in `R` based on this algorithm. The function, `my_gee()`, was intended as a homemade version of `geeglm()`. I will now, through an example, argue that `my_gee()`, and thus Algorithm 6.2, always returns the same GEE estimator no matter the initial $\hat{\boldsymbol{\beta}}$.

Example 4 The data set I will be using is `Orthodont` from the `nlme`-package. It consists of 27 boys and girls. At the ages 8, 10, 12 and 14, the change in an orthodontic measurement (distance from pituitary to pterygomaxillary fissure) have been recorded. The data consists of four variables; `distance`, `age`, `Sex` and `Subject`. Below, the first eight rows of the data are shown:

```
> head(Orthodont, 8)
   distance age Subject  Sex
1      26.0   8     M01 Male
2      25.0  10     M01 Male
3      29.0  12     M01 Male
4      31.0  14     M01 Male
5      21.5   8     M02 Male
6      22.5  10     M02 Male
7      23.0  12     M02 Male
8      26.5  14     M02 Male
```

I will now model the response, `distance`, with both `geeglm()` and `my_gee()`. Through the argument `init.beta` in `my_gee()`, I specify nine different initial $\hat{\boldsymbol{\beta}}$'s to see what GEE estimator it will result in and how many iterations are used. The initial $\hat{\boldsymbol{\beta}}$'s will be vectors where all entries are equal to `init.beta`. In `my_gee()` only $g(\mu_i) = \mu_i$ and $V(\mu_i) = 1$ is implemented, and `init.beta` is per default the coefficients from a GLM, just like in `geeglm()`. The nine values of `init.beta` are

```
ib = c(NULL, -5000, -777, -0.3, 0, 0.00005, 4, 100, 1234, 5678),
```

where NULL is the default setting. I have made a for-loop that runs through these values and gathers the GEE estimators in a matrix, `coefs`, together with the number of iterations used. The model with `geeglm()` and the models with `my_gee()` are shown below, where `ib` takes the nine values one at a time:

```
m.geeglm = geeglm(distance ~ age + Sex, id = Subject, data = Orthodont, family = "gaussian",
    corstr = "unstructured")
m.my_gee = my_gee(distance ~ age + Sex, id = Orthodont$Subject, data = Orthodont, corstr = "
    unstructured", init.beta = ib[i])
```

Below, in `coefs`, the GEE estimators are listed in the order mentioned above. This means that the first row in `coefs` is the GEE estimator, we get when initializing $\hat{\boldsymbol{\beta}}$ with the coefficients from a GLM. This estimator is not perfectly equal to that of `m.geeglm`, but the differences are very small. We see that even with very big differences in the initial $\hat{\boldsymbol{\beta}}$'s, the GEE estimator is always the same. Only the number of iterations change. Thus, I believe that Algorithm 6.2 always returns the same GEE estimator no matter the initial $\hat{\boldsymbol{\beta}}$.

```
> m.geeglm$coefficients
(Intercept)         age   SexFemale
  17.696016   0.6597990 -2.2232260

> coefs
              ib (Intercept)       age SexFemale iterations
 [1,]       NULL    17.59809 0.6691337 -2.255985         13
 [2,] -777.00000    17.59809 0.6691337 -2.255985         13
 [3,]   -0.30000    17.59809 0.6691337 -2.255985         12
 [4,]    0.00000    17.59809 0.6691337 -2.255985         12
 [5,]    0.00005    17.59809 0.6691337 -2.255985         12
 [6,]    4.00000    17.59809 0.6691337 -2.255985         12
 [7,]  100.00000    17.59809 0.6691337 -2.255985         13
 [8,] 1234.00000    17.59809 0.6691337 -2.255985         13
 [9,] 5678.00000    17.59809 0.6691337 -2.255985         13
```

□

Much like the maximum likelihood estimator from Chapter 4, the GEE estimator, $\hat{\boldsymbol{\beta}}$, is asymptotically $N_p\left(\boldsymbol{\beta}, \mathrm{I}(\boldsymbol{\beta})^{-1}\right)$-distributed, where

$$\mathrm{I}(\boldsymbol{\beta})^{-1} = -\mathrm{E}\left[\frac{\partial}{\partial \boldsymbol{\beta}^T} S(\boldsymbol{\beta})\right] = \sum_{i=1}^{N} \frac{\partial \boldsymbol{\mu}_i^T}{\partial \boldsymbol{\beta}} \mathbf{V}_i(\mathbf{a})^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}^T}. \tag{6.12}$$

Upon convergence, we would like to interpret on the effects related to the groups. Ideally, we would set up some models and then do comparisons using LRTs. But as $\hat{\boldsymbol{\beta}}$ are estimates obtained using quasi-likelihood, and not the usual likelihood, no LRTs can be performed. Luckily, we can still do a Wald test to determine whether $\hat{\beta}_i = 0$. Thus we rely on inference about $\boldsymbol{\beta}$. Regardless of whether $\mathbf{V}_i(\mathbf{a})$ is specified correctly, $\hat{\boldsymbol{\beta}}$ will be a consistent estimator of $\boldsymbol{\beta}$. Hence, whether or not $\hat{\boldsymbol{\beta}}$ is consistent cannot be used as an argument to choose one correlation structure over an other. Instead, we can look at the standard errors for $\hat{\boldsymbol{\beta}}$. For this, we will need the covariance of $\hat{\boldsymbol{\beta}}$. The standard errors are the squared diagonal elements in the covariance matrix. There are two ways to estimate $\mathrm{Cov}\left[\hat{\boldsymbol{\beta}}\right]$:

- The model based estimate:

$$\mathrm{Cov}\left[\hat{\boldsymbol{\beta}}\right]_{\mathrm{M}} \overset{(6.12)}{=} \mathrm{I}(\boldsymbol{\beta}).$$

If the mean value and correlation structure are specified correctly, then $\text{Cov}\left[\hat{\boldsymbol{\beta}}\right]_{\text{M}}$ is a consistent estimator of $\text{Cov}\left[\hat{\boldsymbol{\beta}}\right]$. The model based estimate yields a close approximation of the standard errors of $\hat{\boldsymbol{\beta}}$ provided $\mathbf{V}_i(\mathbf{a})$ is a close approximation of the true underlying covariance.

- The empirical estimate:

$$\text{Cov}\left[\hat{\boldsymbol{\beta}}\right]_{\text{E}} = \text{I}(\boldsymbol{\beta})\left(\sum_{i=1}^{N}\frac{\partial\boldsymbol{\mu}_i^T}{\partial\boldsymbol{\beta}}\mathbf{V}_i(\hat{\mathbf{a}})^{-1}\text{Cov}\left[\mathbf{Y}_i\right]\mathbf{V}_i(\hat{\mathbf{a}})^{-1}\frac{\partial\boldsymbol{\mu}_i}{\partial\boldsymbol{\beta}^T}\right)\text{I}(\boldsymbol{\beta}).$$

  Even if the correlation structure is misspecified, $\text{Cov}\left[\hat{\boldsymbol{\beta}}\right]_{\text{E}}$ is a consistent estimator of $\text{Cov}\left[\hat{\boldsymbol{\beta}}\right]$. The empirical estimate is also known as the "robust" estimate because it is always consistent and because it yields correct standard errors of $\hat{\boldsymbol{\beta}}$ for large sample sizes, despite $\mathbf{V}_i(\mathbf{a})$ being misspecified.

Clearly, if $\text{Cov}\left[\mathbf{Y}_i\right]$ is specified correctly, so that $\text{Cov}\left[\mathbf{Y}_i\right] = \mathbf{V}_i(\mathbf{a})$, then $\text{Cov}\left[\hat{\boldsymbol{\beta}}\right]_{\text{E}} = \text{Cov}\left[\hat{\boldsymbol{\beta}}\right]_{\text{M}}$. This means, that when testing different covariance structures, the one that fits data best, will be the one that reduces the difference between $\text{Cov}\left[\hat{\boldsymbol{\beta}}\right]_{\text{E}}$ and $\text{Cov}\left[\hat{\boldsymbol{\beta}}\right]_{\text{M}}$. In practice, $\text{Cov}\left[\mathbf{Y}_i\right]$ is replaced by $(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_i)(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_i)^T$.

The GLM does not take into account that observations may be correlated. But how much exactly does that influence the model? And how big does the difference between $\text{Cov}\left[\hat{\boldsymbol{\beta}}\right]_{\text{M}}$ and $\text{Cov}\left[\hat{\boldsymbol{\beta}}\right]_{\text{E}}$ have to be for it to be obvious, that the chosen covariance structure is unrealistic for the data? I will try to answer these questions in Example 5.

Example 5 The example is based on the data set `respitory` from the package `HSAUR2`. I will be using a subset of `respitory` in which I have made a few changes. The new data set is called `resp` and consists of 444 observations on 111 patients (4 observations per patient). The patients were observed for their respitory status, which was either "poor" or "good". Below are the first 10 rows of the data:

```
> head(resp, 10)
    centre     group gender age month subject baseline status
112      1   placebo female  46     1       1     poor      0
223      1   placebo female  46     2       1     poor      0
334      1   placebo female  46     3       1     poor      0
445      1   placebo female  46     4       1     poor      0
113      1   placebo female  28     1       2     poor      0
224      1   placebo female  28     2       2     poor      0
335      1   placebo female  28     3       2     poor      0
446      1   placebo female  28     4       2     poor      0
114      1 treatment female  23     1       3     good      1
225      1 treatment female  23     2       3     good      1
```

The data consists of 8 variables:

- `baseline`: the patients' status at the start of the study,
- `month`: once a month in the next four following months, their respitory status was registered. This variable indicates which month, we are at,
- `status`: the patients' current status coded as "0" for "poor" and "1" for "good",
- `centre`: the centre they were treated at, either 1 or 2,
- `group`: the patients were randomized into two treatment groups and were given either a placebo or treatment drug,

- `gender`: selfexplanatory,
- `age`: their age at the time the observation was registered ranging from 11 to 68 years, and
- `subject`: a factor identifying each patient.

The response variable is binary, so I will be using the link and variance function from the Binomial distribution. I will set up a GLM and two GEE models; one with independence and one with unstructured correlation structure. I will fit a logistic regression model to the data using `glm()`, and I will set up the GEE models with the function `geeglm()` from the package `geepack`. The setup of the `geeglm`-function is fairly selfexplanatory as it resembles `glm()`. Only a few extra arguments are specified:

```
1 resp.glm = glm(status ~ centre + group + gender + baseline + age, family = "binomial", data =
      resp)
2 resp.gee1 = geeglm(status ~ centre + group + gender + baseline + age, family = "binomial",
      corstr = "independence", scale.fix = T, id = subject, data = resp)
3 resp.gee2 = geeglm(status ~ centre + group + gender + baseline + age, family = "binomial",
      corstr = "unstructured", scale.fix = T, id = subject, data = resp)
```

Through the argument `corstr`, the correlation structure is specified, and the argument `id` is needed to seperate the observations according to subjects when estimating $\mathbf{V}_i(\mathbf{a})$, for instance. Through `scale.fix`, I tell the function whether ϕ should be estimated or fixed at the value 1. Only when there might be a problem with overdispersion, do I want to estimate ϕ. That is not the case here, though:

```
1 > tapply(resp$status, interaction(resp$group, resp$month), mean)
2   placebo.1 treatment.1   placebo.2 treatment.2   placebo.3 treatment.3   placebo.4 treatment.4
3     0.49123     0.68519     0.38596     0.70370     0.45614     0.72222     0.43860     0.62963
4
5 > tapply(resp$status, interaction(resp$group, resp$month), var)
6   placebo.1 treatment.1   placebo.2 treatment.2   placebo.3 treatment.3   placebo.4 treatment.4
7     0.25439     0.21978     0.24123     0.21244     0.25251     0.20440     0.25063     0.23760
```

The very first column

placebo.1
0.49123

shows that the mean for the placebo-group at the 1st timepoint is 0.49123. In these outputs, we see that there is not much difference between the means and variances of `status` at each of the timepoints for the treatment groups. Hence, I fix ϕ at 1.

I have extracted the standard errors from all three models and gathered them in the matrix below for easy comparison:

```
1 > SEmatrix
2                    glm   gee1.M   gee1.E    diff.1   gee2.M   gee2.E    diff.2
3 (Intercept)    0.33765  0.33728  0.46033 -0.12305  0.47994  0.46128  0.01866
4 centre2        0.23957  0.23930  0.35682 -0.11752  0.33960  0.35480 -0.01520
5 grouptreatment 0.23684  0.23658  0.35078 -0.11420  0.33596  0.34945 -0.01349
6 gendermale     0.29467  0.29435  0.44320 -0.14886  0.41794  0.44365 -0.02571
7 baselinegood   0.24129  0.24102  0.35005 -0.10903  0.34337  0.34804 -0.00467
8 age            0.00886  0.00885  0.01300 -0.00415  0.01258  0.01291 -0.00033
```

First column is the standard errors of `resp.glm`. The next three columns are the model based standard errors, the empirical standard errors and the difference between these, respectively, for `resp.gee1`. The last three columns are the same, but for `resp.gee2`.

As mentioned earlier, the GLM is just a GEE model with independence correlation structure, so it may seem strange that `glm` and `gee1.M` are not identical. This is due to the algorithm `geeglm()` uses during estimation. Had I used the function `gee()` from the package `gee`, I would have got identical results. However, `glm` and `gee1.M` are close enough, that we may proceed with `geeglm()`. Looking only at `glm`, there is no way to tell whether the model is a good fit for the data. It is a bit easier for the GEE models as we can compare the model based and the empirical standard errors, in order to conclude whether the model is any good. In the columns `diff.1` and `diff.2`, it is clear that unstructured correlations structure is more realistic for the data, as `diff.2` < `diff.1` by about a factor of 10 per variable in the models. To further underline `resp.gee2` being the better model, the estimated association parameters, $\hat{\mathbf{a}}$, show that there indeed is a correlation between observations on the same subject[19]:

```
> resp.gee2$geese$alpha
alpha.1:2 alpha.1:3 alpha.1:4 alpha.2:3 alpha.2:4 alpha.3:4
    0.328     0.208     0.298     0.439     0.363     0.399
```

The correlations are not extremely high, but they are certainly not 0 either. In conclusion, in this example, we have seen that the GLM is an insufficient model, when the responses are multidimensionel. And we have seen how to use the standard errors to choose the more fitting correlation structure for the data.

□

6.3 GEE vs. LMM

This section is based on [13].

In the GEE model, we estimate $\boldsymbol{\beta}$, so that we can say something about the population-averaged effect. In the LMM, we furthermore estimate the u_i's, which enables us to "dig a little deeper" and say something about the subject-specific effects.

> **!** We want to know the effect of the treatment on each of the patients, and we want to use this to determine whether there is a difference in the treatment groups. In that sense, we should use an LMM as we are interested in the subject-specific effects. But we could also formulate the goal of this report a little differently, so that it makes sense to use a GEE model. Stated in a different way, we want to know how well the patients react to Ibuprofen wrt. wrist mobility. We want to know the population-averaged effect of the treatment groups. This allows us to compare the groups.

The GEE model can only take into account one source of clustering. If the vocabulary test-example from Chapter 4 had been extended to including several schools with several classes, we would also have to take into account the within-school and within-class correlation. This is no problem for the LMM as it would account for these extra sources of variation by having a random effect for each source. We cannot do that in the GEE model as no random effects are specified in the model. It is only through the correlation, $\mathbf{R}_0(\mathbf{a})$, where the within-subject association among repeated measurements is incorporated, that we account for the population-averaged effects. So, a disadvantage of the GEE model is that we cannot use it if we have more than one source of random effects. And, obviously, if we are interested in the subject-specific effects, the GEE model is no good.

[19] Some literature uses $\boldsymbol{\alpha}$ instead of $\mathbf{a}$, which is why "`alpha`" is in the output from `R`.

An advantage of the GEE model is that, as mentioned, even if the correlation is misspecified, both $\hat{\boldsymbol{\beta}}$ and $\text{Cov}\left[\hat{\boldsymbol{\beta}}\right]_{\text{E}}$ are still consistent estimators. The fact that the GEE model provides two covariance matrices for $\hat{\boldsymbol{\beta}}$ is helpful when choosing the better model according to covariance structure. Another advantage of the GEE model is, as mentioned also, that no assumptions on the distribution of the data is needed. Wrt. the residuals, there is no assumption of them being Gaussian with zero mean, but it does improve efficiency of $\hat{\boldsymbol{\beta}}$ if this holds. Actually, if the mean and variance of the response is correctly specified, the residuals will be Gaussian with zero mean.

7 | Results of modelling data with GEEs

In this chapter, I go through the results of modelling the response variables with GEE models as described in Chapter 6. The setup of this chapter is the same as in Chapter 5; first I will choose a function to set up models. Then I will perform the analysis with the autoregressive, the unstructured and the independence correlation structure, respectively, and I will end the analysis with a comparison of the three sets of models to see which correlations structure is most realistic for the data. I will end the chapter with a conclusion of whether there is any significant difference between the treatment groups.

Choosing a function and a model

The packages `geepack`, `gee` and `geeM` offer the functions `geeglm()`, `gee()` and `geem`, respectively, for creating a GEE model. The setup in the functions is nearly identical, and having tested these functions on my data, I find I get all the same results. The only differences are what is displayed in some of the outputs, for instance, the output of `summary()` is a little different for each of the functions. Hence, it makes no difference which function I choose, and as such, I will choose `geeglm()`.

Next, I must decide whether ϕ should be estimated freely or fixed at 1. For this, I find the means and variances across the groups at the three timepoints for each type of movement:

```
> tapply(Wrist$ps, interaction(Wrist$group, Wrist$time), mean)
    1.0     2.0     3.0     1.7     2.7     3.7    1.46    2.46    3.46
58.416  63.620  66.118  86.940  85.728  91.516  94.210  93.554  95.914

> tapply(Wrist$ps, interaction(Wrist$group, Wrist$time), var)
    1.0     2.0     3.0     1.7     2.7     3.7    1.46    2.46    3.46
365.89  390.05  279.78  144.25  154.90   57.68   84.36   84.44   18.32

> tapply(Wrist$ur, interaction(Wrist$group, Wrist$time), mean)
    1.0     2.0     3.0     1.7     2.7     3.7    1.46    2.46    3.46
31.952  38.166  33.828  71.238  68.222  77.403  93.860  89.605  92.491

> tapply(Wrist$ur, interaction(Wrist$group, Wrist$time), var)
    1.0     2.0     3.0     1.7     2.7     3.7    1.46    2.46    3.46
152.65  253.14  202.52  362.75  406.50  258.38  407.92  212.50  327.23

> tapply(Wrist$ef, interaction(Wrist$group, Wrist$time), mean)
    1.0     2.0     3.0     1.7     2.7     3.7    1.46    2.46    3.46
22.426  29.311  27.920  66.028  69.634  71.771  87.150  92.058  89.478

> tapply(Wrist$ef, interaction(Wrist$group, Wrist$time), var)
    1.0     2.0     3.0     1.7     2.7     3.7    1.46    2.46    3.46
127.50  319.69  170.19  132.36  443.08  245.07  102.44  238.74  280.13
```

The output of `tapply()` needs a quick explanation. The first column in the first `tapply()`

1.0

58.416

tells us, that the mean of `ps` for group 1 at the first timepoint (coded as 0) is 58.416. Hence, the first row in the output refers to the groups (either 1, 2 or 3) at the timepoints (either 0, 7, or 46), and the second row is the mean for that group at that timepoint. Clearly, there is a difference between the means and the

variances, which means that overdispersion could be a problem, and hence ϕ must be estimated freely. This is done by setting `scale.fix = FALSE`, which is the default setting in `geeglm()`, and as such, I will just omit this argument in my models.

In Chapter 5, I reasoned that I should include a squared term of `time` in the model in order to achieve linearity between the response and the covariates. This time I will also include a squared term in the models, which means the link function is just the identity, i.e. $g(\mu_i) = \mu_i = \mathbf{x}_i^T \boldsymbol{\beta}$. In `geeglm()` this is expressed by `family = gaussian(link = "identity")`, in which I also specify the variance function as belonging to the Gaussian family[20]. Like in Chapter 5, I will begin my analysis using autoregressive covariance structure. The models are set up as follows (here only the model for `ps` is shown):

```
1 ps.gee.ar = geeglm(ps ~ time + I(time^2) + group, id = subject, data = Wrist, corstr = "ar1",
      family = gaussian(link = "identity"))
```

The models for the uln/rad- and ex/flex-movements, `ur.gee.ar` and `ef.gee.ar`, are set up the exact same way only with `ur` and `ef` instead of `ps`. The autoregressive covariance structure is specified through `corstr = "ar1"`. In the GEE model no random effects are specified, although one could view the argument `id` as being somewhat equivalent to the `random`-argument in `lme()`.

7.1 Analysis using `corstr = "ar1"`

I will start off by making sure the models (henceforth termed the "ar"-models) are not overdispersed. They should not be, as ϕ have not been set to any fixed value. To check for overdispersion, we compare the estimated dispersion (extracted from `summary()`) with what it ought to be according to the formula in Algorithm 6.2:

$$\frac{1}{N-p}\sum_{i=1}^{N}\sum_{j=1}^{n}\varepsilon_{ij}^2.$$

These two dispersions can be extracted from my function `superGEE()` by `superGEE()$dispersion`. For easy comparison of all the models, I have gathered these two values of $\hat{\phi}$ in a simple matrix, where the rows refer to the estimates form either the summary or the formula:

```
1 > phi.ar
2           ps.gee.ar ur.gee.ar ef.gee.ar
3 summary      171.80    284.11    221.46
4 formula      175.32    289.93    226.00
```

I deem these differences small enough that none of the "ar"-models are overdispersed.

Only two things need to be tested; whether `group` is relevant to the models and whether the residuals are homoscedastic. As it would improve efficiency in the sense, that the mean and variance of the response will have been correctly specified, if the residuals are Gaussian with zero mean, I will also be testing this. To test the relevance of the treatment groups, I am using three Wald tests, all of which is implemented in `superGEE()`. The first Wald test is performed by `anova()` with just one input (the model, we want to test). This test tests the hypotheses

$$H_{0,1}: \hat{\beta}_{\text{time}} = 0, \qquad H_{0,2}: \hat{\beta}_{\text{time}^2} = 0, \qquad H_{0,3}: \hat{\beta}_{\text{group}} = 0.$$

The second test is also performed by `anova()`, but with two inputs; the model, we want to test, and the

[20] I have tested other choices for the variance function, but only the one from the Gaussian family gave sensible results.

corresponding null-model. Just like in Chapter 5, I am using the term "null-model" to mean a model without the `group`-term and otherwise identical to the corresponding full model. This test is just the usual comparison of two models, only compared using a Wald statistic instead of AIC. The third test is the result of using the function `esticon()` on the model. This function also performs a Wald test, but tests the hypotheses that

$$H_{0,1}: \hat{\beta}_{\text{time}} = 0, \quad H_{0,2}: \hat{\beta}_{\text{time}^2} = 0, \quad H_{0,3}: \hat{\beta}_{\text{gr.2}} = 0, \quad H_{0,4}: \hat{\beta}_{\text{gr.3}} = 0,$$

and it also gives the 95%-confidence intervals for the estimated coefficients. Remember, if zero is contained in the interval, we can accept that the corresponding coefficient is zero. The results of the three Wald tests are:

```
> superGEE(ps.gee.ar)$group
$anova1
             p–value
time         0.00000
I(time^2)    0.00000
group        0.14380

$anova2
[1] 0.14380

$esticon
              p–value     lower     upper
(Intercept)   0.00000  55.62940  65.63713
time          0.00000   3.62623   4.66557
I(time^2)     0.00000  −0.08473  −0.06541
group2        0.60576  −4.33137   7.42813
group3        0.05838  −0.16581   9.51960
```

The third p-value from `$anova1` and the p-value from `$anova2` seem to be the same. Had I included more decimals, we would see, that they are very close, but not equal[21] (this goes for all the models in this chapter). The output shows that `group` is irrelevant to `ps.gee.ar`. The same goes for the models `ur.gee.ar` and `ef.gee.ar` as seen in Table 4. In the table, I am including the p-values and confidence intervals for `group` only, as these results are of most interest.

Table 4: *Results of using `superGEE()$group` on the "ar"-models. The row for `$anova1` are the p-values for `group`. The row for `$anova2` are the p-values from the comparison of the model and the corresponding null-model. The rows for `$esticon` are the 95%-confidence intervals for $\hat{\beta}_{gr.2}$ and $\hat{\beta}_{gr.3}$, and the p-values for these coefficients.*

	`superGEE()$group`	
	`ur.gee.ar`	`ef.gee.ar`
`$anova1`	0.88083	0.09464
`$anova2`	0.88083	0.09464
`$esticon`	$\hat{\beta}_{\text{gr.2}} \in [-6.64, 6.93]$, 0.96712 $\hat{\beta}_{\text{gr.3}} \in [-5.03, 8.02]$, 0.65431	$\hat{\beta}_{\text{gr.2}} \in [-0.61, 11.53]$, 0.07793 $\hat{\beta}_{\text{gr.3}} \in [-0.72, 9.24]$, 0.09379

The results from `superGEE()$group` are unanimous; `group` is irrelevant to the models.

[21] I thought perhaps the p-values from `$anova1` was a result of `anova()` one by one removing `time`, `I(time)^2` and `group` from the model, and then comparing the reduced model with `ps.gee.ar`. Upon investigation, I found that that is not so.

Testing `ps.gee.ar` for homoscedasticity, I get the plots in Figure 16 and the following p-values:

```
> superGEE(ps.gee.ar, test = "homo")
    Bart     Leve
0.00020  0.98368
```

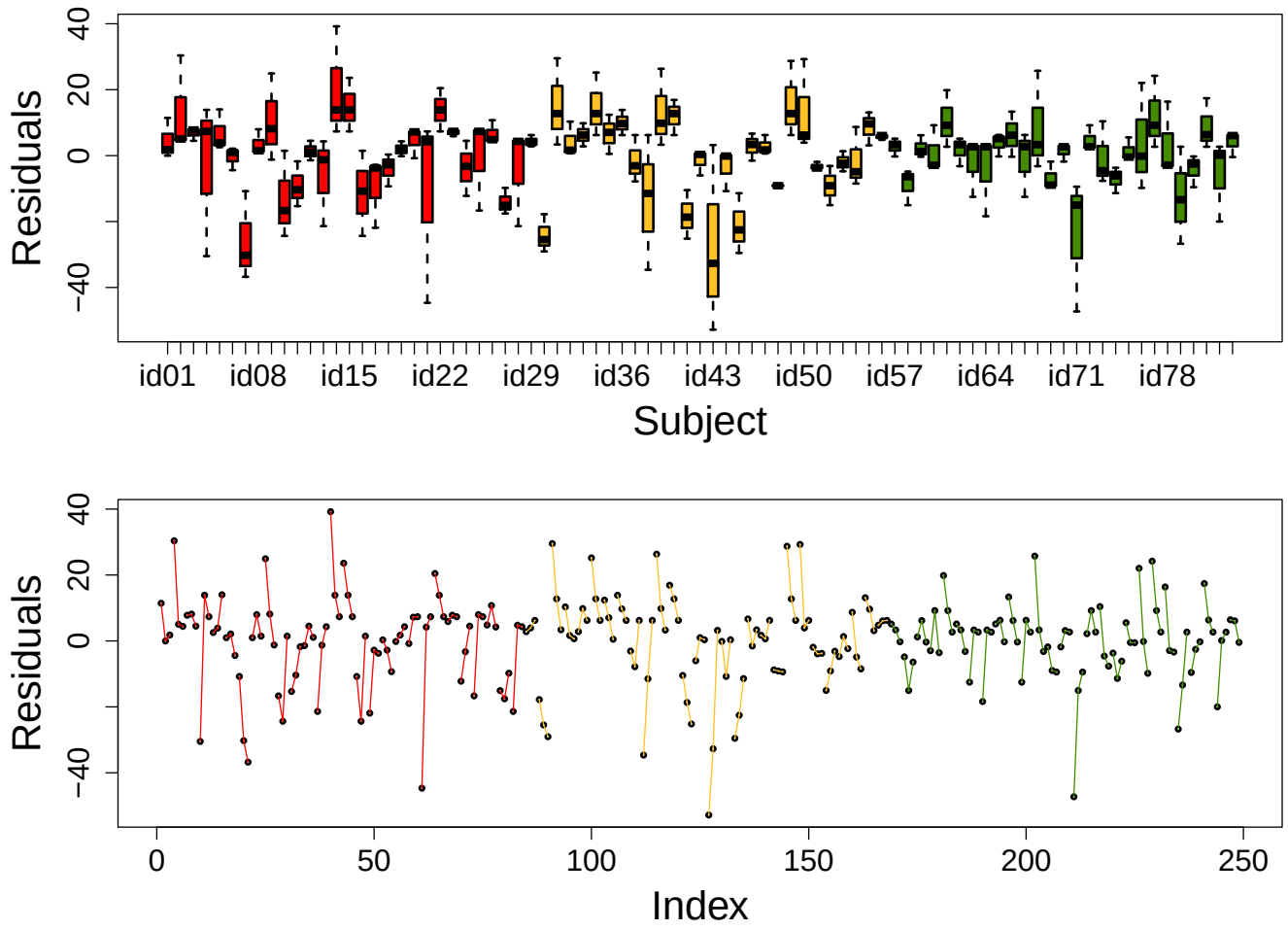

Figure 16: *Top: Boxplots of each patients' residuals. Bottom: Residuals vs. their index numbers. In both plots, red, yellow and green represent group 1, 2 and 3, respectively.*

The top plot in Figure 16 is intended for detecting patterns in the residuals between the groups. It shows that there is a slightly smaller spread in the residuals from group 3, but otherwise no obvious patterns that seperate the groups. An unwanted pattern that could show non-homoscedasticity could perhaps be, that the boxplots for, say, group 1 all tended to have maximum values far from the median. The bottom plot is intended for detecting patterns in the residuals for each patient. This plot was explained in Chapter 5 (just before Figure 10), where an unwanted pattern could perhaps be each patient's residuals forming a "U"-shape. The plot shows no obvious trend in each patient's residuals. I see no reason not to assume homoscedasticity, eventhough the Bartlett test disagrees with the Levene's test. This is likely do to the residuals not being Gaussian, which we see when testing for normality by

```
> superGEE(ps.gee.ar, test = "norm")
     Shap
1.463e-08
```

Figure 17 shows that, although being centered around zero, the residuals are not Gaussian according to

the Q-Q plot. The histogram agrees less with the Shapiro-Wilk test. We cannot rule out that they may be asymptotically Gaussian. For `ur.gee.ar` and `ef.gee.ar`, the residuals are homoscedastic, and the plots from are similar to Figure 16 (although for `ef.gee.ar`, it is groups 1 that has the smallest spread). Neither of the models have Gaussian residuals. The Q-Q plots for the models look a lot better than the one in Figure 17, but the histograms are not nearly as good as the one for `ps.gee.ar`. The problems with normality could be due to the autoregressive structure not being the most fitting for the data.

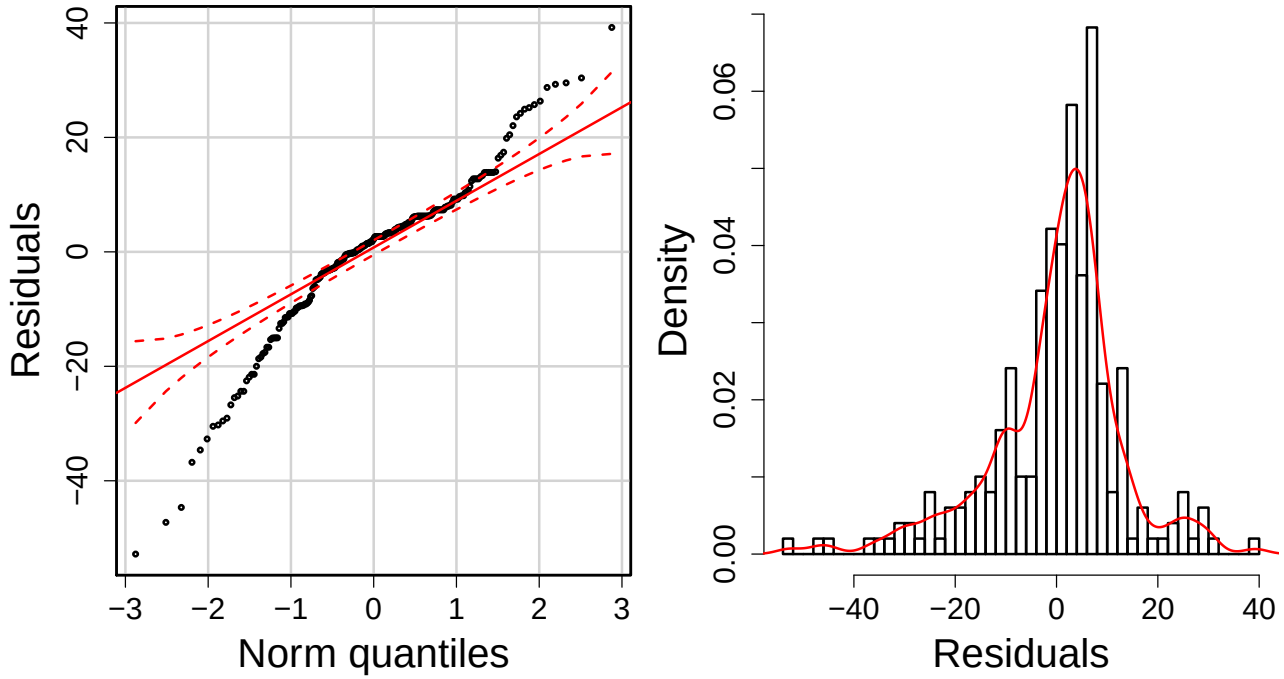

Figure 17: *Left: Q-Q plot of the residuals. Right: Histogram of the residuals with density curve (red line).*

7.2 Analysis using `corstr = "unstructured"`

The models `ps.gee.un`, `ur.gee.un` and `ef.gee.un` are set up exactly like the "ar"-models, only with `corstr = "unstructured"` used instead. I will refer to these models as the "un"-models. Once again, I am starting off by confirming that they are not overdispersed. Next, I move on to testing the relevance of `group`. The results from `superGEE()$group` are shown in Table 5. These results show that `group` is irrelevant to the "un"-models.

Table 5: *Results of using `superGEE()$group` on the "un"-models. The row for `$anova1` are the p-values for `group`. The row for `$anova2` are the p-values from the comparison of the model and the corresponding null-model. The rows for `$esticon` are the 95%-confidence intervals for $\hat{\beta}_{gr.2}$ and $\hat{\beta}_{gr.3}$, and the p-values for these coefficients.*

	`superGEE()$group`		
	`ur.gee.un`	`ur.gee.un`	`ef.gee.un`
`$anova1`	0.14861	0.91882	0.09131
`$anova2`	0.14861	0.91882	0.09131
`$esticon`	$\hat{\beta}_{gr.2} \in [-4.20, 7.19]$, 0.60772 $\hat{\beta}_{gr.3} \in [-0.21, 9.13]$, 0.06102	$\hat{\beta}_{gr.2} \in [-6.16, 7.21]$, 0.87798 $\hat{\beta}_{gr.3} \in [-5.09, 7.45]$, 0.68496	$\hat{\beta}_{gr.2} \in [-0.39, 11.54]$, 0.06688 $\hat{\beta}_{gr.3} \in [-0.87, 9.14]$, 0.10514

I will not show the results from using `superGEE(test = c("homo", "norm"))` on the "un"-models as they are identical to the results from the "ar"-models.

7.3 Analysis using `corstr = "independence"`

The last set of models are the "in"-models, i.e. models with `corstr = "independence"`. These models are also not overdispersed, and Table 6 shows that `group` is not relevant to these models either.

Table 6: *Results of using `superGEE()$group` on the "in"-models. The row for `$anova1` are the p-values for `group`. The row for `$anova2` are the p-values from the comparison of the model and the corresponding null-model. The rows for `$esticon` are the 95%-confidence intervals for $\hat{\beta}_{gr.2}$ and $\hat{\beta}_{gr.3}$, and the p-values for these coefficients.*

	GsuperGEE()$group		
	ur.gee.in	ur.gee.in	ef.gee.in
$anova1	0.11750	0.70523	0.10997
$anova2	0.11750	0.70523	0.10997
$esticon	$\hat{\beta}_{\text{gr.2}} \in [-4.72, 6.94]$, 0.70844 $\hat{\beta}_{\text{gr.3}} \in [-0.06, 9.38]$, 0.05312	$\hat{\beta}_{\text{gr.2}} \in [-7.42, 6.71]$, 0.92204 $\hat{\beta}_{\text{gr.3}} \in [-4.46, 8.90]$, 0.51407	$\hat{\beta}_{\text{gr.2}} \in [-1.25, 11.51]$, 0.11476 $\hat{\beta}_{\text{gr.3}} \in [-0.54, 9.58]$, 0.08008

I will not show the results from using `superGEE(test = c("homo", "norm"))` on the "in"-models as they are, just like the "un"-models, identical to the results from the "ar"-models. I have included a comment on this in my discussion as I find it a bit odd.

I will now move on to comparing the models in order to find which set of models has the most appropriate covariance structure.

7.4 Comparison of the "ar"-, "un"- and the "in"-models

When comparing models with different covariance structure, we must look at the standard errors of $\hat{\boldsymbol{\beta}}$. The empirical standard errors are the squared diagonal elements in $\text{Cov}\left[\hat{\boldsymbol{\beta}}\right]_{\text{E}}$, and the model based standard errors are the squared diagonal elements in $\text{Cov}\left[\hat{\boldsymbol{\beta}}\right]_{\text{M}}$. After extracting the covariances, `superGEE()` calculates the empirical and the model based standard errors. Then these are subtracted from each other. Let $\boldsymbol{\delta}$ denote these differences. Remember, the model with the smallest $\boldsymbol{\delta}$, is the one with the most fitting covariance structure. With `superGEE()$delta`, the standard errors are extracted. For easy comparison, I have gathered all the $\boldsymbol{\delta}$'s for each type of model in three matrices, where the rows show $\boldsymbol{\delta}$ for the "ar"-, "un"- and "in"-models, respectively. Below are the matrices for `ps`, `ur` and `ef`:

```
> modcomp.delta.ps
   (Intercept)      time  I(time^2)   group1   group2
ar     0.57015   0.02071  6.701e-05  0.04145  0.17369
un     0.57073   0.07712  1.129e-03  0.00975  0.13253
in     0.57171  -0.07375 -2.024e-03  0.39562  0.53857

> modcomp.delta.ur
   (Intercept)      time  I(time^2)    group1    group2
ar    -0.28777   0.01279  3.298e-04 -0.00979 -0.02823
un    -0.29281  -0.01028 -1.526e-06 -0.03167 -0.03596
in    -0.27818  -0.10745 -2.333e-03  0.53881  0.53854

> modcomp.delta.ef
   (Intercept)      time  I(time^2)    group1   group2
ar    -0.08670  -0.00123  7.735e-05 -0.31712  0.21123
un    -0.08676   0.01566  4.467e-04 -0.31835  0.19314
in    -0.08632  -0.12311 -2.607e-03  0.22063  0.78670
```

In general, the "ar"- and "un"-models have the smallest $\boldsymbol{\delta}$'s, but the differences between the three sets of models for each type of movement are not at all as obvious as they were in Example 5. Hence, there is no clear indication of either of the covariance structures being more fitting than the others. This very much agrees with the results from the LMMs from Chapter 5.

With these results, it is interesting to see what the association parameters, **a**, have been estimated to be. I would expect $\hat{\mathbf{a}}$ to be small so as to resemble the independence structure as, according to the standard errors, the "in"-models are about as fitting as the "ar"- and "un"-models. Below is $\hat{a}$ for the "ar"-models:

```
> superGEE(ps.gee.ar)$ass.param
alpha
0.474

> superGEE(ur.gee.ar)$ass.param
alpha
0.471

> superGEE(ef.gee.ar)$ass.param
alpha
0.528
```

For the "un"-models, $\hat{\mathbf{a}}$ is

```
> superGEE(ps.gee.un)$ass.param
alpha.1:2 alpha.1:3 alpha.2:3
    0.689     0.166     0.315

> superGEE(ur.gee.un)$ass.param
alpha.1:2 alpha.1:3 alpha.2:3
    0.388     0.180     0.594

> superGEE(ef.gee.un)$ass.param
alpha.1:2 alpha.1:3 alpha.2:3
    0.585     0.237     0.514
```

These estimated values of the association parameter are larger than what I would have expected, when comparing to $\hat{\mathbf{a}}$ from Example 5, where there was a clear difference between the "in"- and the "un"-model. These intra-subject correlations seem to contradict the results from comparing the standard errors. I have included a comment on this in my discussion, Chapter 9.

7.5 Conclusion

Whether an autoregressive, unstructured or independence correlation structure is used, `group` is not significant to the model for either of the three types of movement. And there seem to only be little difference in the models with these correlation structures with a very slight advantage to the "ar"- and "un"-models.

With that, I end this chapter by seeing no reason to reject H_0^{main}, just like in Chapter 5.

7.6 Source code: `superGEE()`

```
superGEE = function(mod, test = NULL){
  cex.legend = 1.5; lwd = 2; cex.lab = 2; cex.axis = 1.5; cex.main = 2; mar = c(4.5,4.5,2,1)

  ## Residuals
  if(!is.null(test)){
    if(test == "homo"){
      par(mar=mar, mfrow=c(2,1), lwd=lwd, cex.lab=cex.lab, cex.axis=cex.axis, cex.main=cex.main)
      col = c(rep(2,28), rep("goldenrod1",28), rep("chartreuse4",27))
      plot(mod$residuals ~ Wrist$subject, col=col, ylab="Residuals", xlab="Subject", cex=0.5)

      col2 = c(rep(2,3*28), rep("goldenrod1",3*28), rep("chartreuse4",3*27))
      plot(mod$residuals, ylab = "Residuals", cex = 0.5)
      for(i in seq(1,3*83,3)){ lines(c(i:(i+2)), mod$residuals[i:(i+2)], col = col2[i]) }
      res.test = c("Bart"=bartlett.test(mod$residuals, Wrist$subject)$p.value,
                   "Leve"=car::leveneTest(as.vector(mod$residuals), Wrist$subject)$"Pr(>F)"[1])
    }

    if(test == "norm"){
      par(mar=mar, mfrow=c(1,2), lwd=lwd, cex.lab=cex.lab, cex.axis=cex.axis, cex.main=cex.main)
      car::qqp(mod$residuals, ylab="Residuals", xlab="Norm␣quantiles", main="␣", cex=0.5)
      hist(mod$residuals, probability = T, 50, xlab = "Residuals", main = "␣")
      lines(density(mod$residuals), col = 2)
      res.test = c("Shap" = shapiro.test(mod$residuals)$p.value)
    }
  }

  ## Relevance of "group"
  w1 = matrix(anova(mod)$"P(>|Chi|)", ncol=1, nrow=3)
  rownames(w1) = rownames(anova(mod)); colnames(w1) = "p-value"
  w2 = anova(update(mod, .~.-group), mod)$"P(>|Chi|)"
  w3 = matrix(c(doBy::esticon.geeglm(mod, diag(5))$"Pr(>|X^2|)",
                doBy::esticon.geeglm(mod, diag(5))$"Lower",
                doBy::esticon.geeglm(mod, diag(5))$"Upper"),
              ncol = 3, nrow = length(mod$coefficients))
  rownames(w3) = names(mod$coefficients); colnames(w3) = c("p-value", "lower", "upper")

  ## Standard errors
  se.e = sqrt(diag(mod$geese$vbeta)); se.m = sqrt(diag(mod$geese$vbeta.naiv))
  m = matrix(0, ncol = length(se.e), nrow = 1); colnames(m) = names(mod$coefficients)
  m[1,] = se.e - se.m

  ## Association parameter
  ass.param = mod$geese$alpha

  ## Dispersion
  phi.summary = summary(mod)$dispersion$"Estimate"
  phi.formula = sum(residuals(mod)^2)/df.residual(mod)

  if(is.null(test)){
    out = list(group = list(anova1 = w1, anova2 = w2, esticon = w3), delta = m,
               dispersion = list(phi.summary = phi.summary, phi.formula = phi.formula),
               ass.param = ass.param); out
  }
  else if(!is.null(test)){ print(res.test) }
}
```

8 A different setup

It is not unthinkable, that a patients ability to, say, rotate their wrist affects their ability to bend the wrist. Instead of creating a model for each type of movement (henceforth termed a *longitudinal model*), perhaps a better model would be a model for each timepoint, where the response for each patient is a vector with the measurements for each type of movement at the specific timepoint (henceforth termed a *clustered model*). In this chapter, I go through the analysis of my data with clustered models. In Chapters 5 and 7, there were problem wrt. homoscedasticity and normality of either the random effects or the residuals. Perhaps the clustered models will give less problems with the model assumptions of the LMMs.

Some rearranging of `Wrist` is needed for the clustered models. The results of which are shown below:

```
> head(Wrist2,9)
   subject        t1        t2        t3 tom group
1     id01  72.222222  86.11111  94.44444  ps     1
2     id01  36.000000  68.00000  68.00000  ef     1
3     id01  40.000000  80.00000  50.00000  ur     1
4     id02  91.176471  91.17647  97.05882  ps     1
5     id02  47.058824  76.47059  76.47059  ef     1
6     id02  37.500000  62.50000 100.00000  ur     1
7     id03  68.571429  94.28571  97.14286  ps     1
8     id03   4.166667  79.16667  91.66667  ef     1
9     id03  22.222222  77.77778 100.00000  ur     1
```

The data set is called `Wrist2`. Just like `Wrist`, it has a column with id's identifying each patient, and a column indicating which group the patient belongs to. Instead of a column with measurements per type of movement, `Wrist2` has a column with measurements per timepoint (columns `t1`, `t2` and `t3` for the 1st, 2nd and 3rd timepoint). The column `tom` indicates what type of movement is measured in each row.

8.1 Clustered LMMs

Before I can set up my LMMs, I must make sure the new response variables, `t1`, `t2` and `t3`, are Gaussian and linear in `group` and `tom`. Figure 18 shows that we may accept `t1` and `t2` being Gaussian as sufficiently many points fall within the 95%-confidence interval. Wrt. `t3`, I have investigated its distribution further, and the most fitting is the Gaussian. Because of the problems with the distribution of `t3`, in the next section, Section 8.2, I will set up clustered GEE models.

The relationship between the new response variables and `group` and `tom` is only linear for some patients. Neither `group` nor `tom` can be transformed as they are both factors, and after having tested several tranformations of the responses, I found no single transformation for either of the responses that resulted in a linear relationship for all patients. Here one has the choice of either continuing under the assumption, that the relationships are all sufficiently linear, or one can work with a non-linear mixed model (NLMM for short). In the NLMM, we have that $\mathrm{E}[\mathbf{Y}_i] = f(\mathbf{X}_i, \boldsymbol{\beta})$, which means that the mean of the set of responses for subject i is now a *function* of $\mathbf{X}_i$ and $\boldsymbol{\beta}$. In order to use a NLMM, we have to know exactly how this function is given. With no apparent guesses as to what $f(\cdot,\cdot)$ may be, I will not attempt to work with the NLMM. Instead, I will assume the relationships are all sufficiently linear, and then try to see if using clustered models results in fewer problems, wrt. meeting model assumptions, than the longitudinal models.

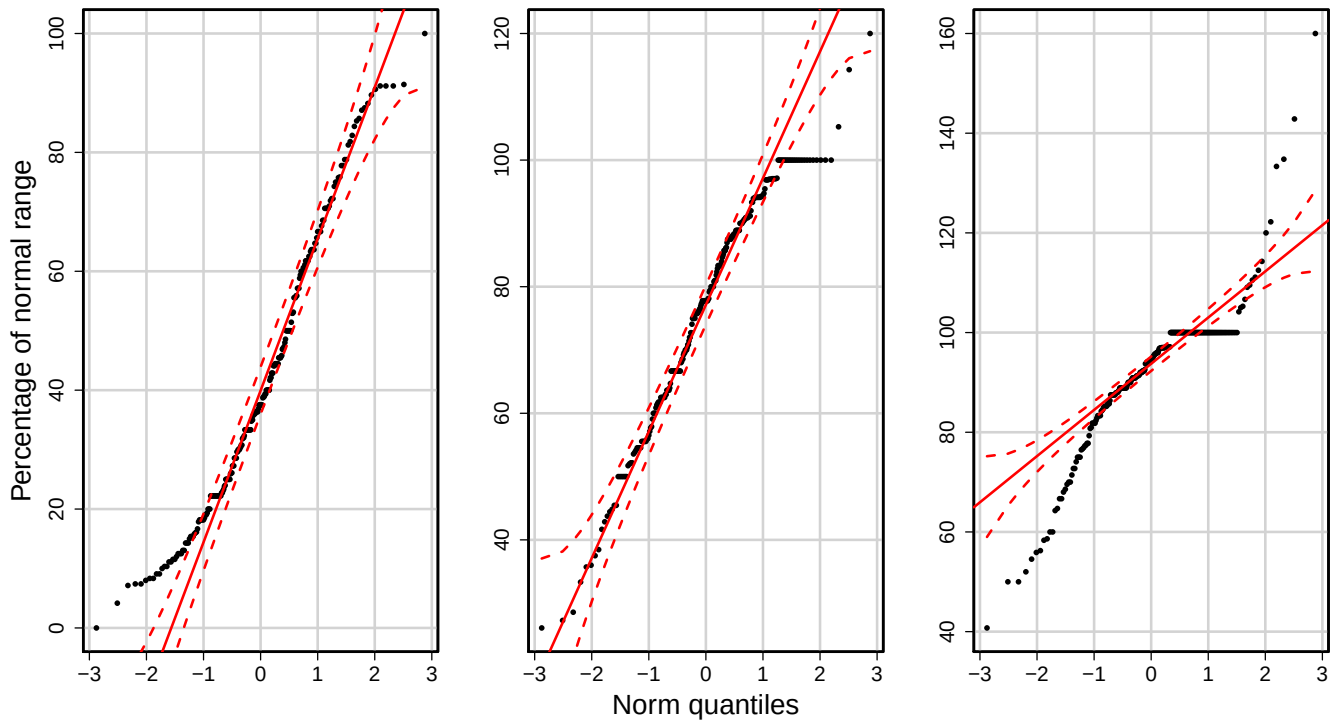

Figure 18: *Q-Q plots of* `t1`*,* `t2` *and* `t3`*, respectively, with 95%-confidence intervals (red punctures lines).*

I will only use models with unstructured covariances in the main analysis. Later, I will set up models with independence structure to see, if we may assume no correlation between types of movement. The models are:

```
1 t1.lmm = lme(t1 ~ tom + group, random = ~1|subject, data = Wrist2, method = "REML", correlation
      = corSymm(form = ~ 1|subject))
2 t2.lmm = lme(t2 ~ tom + group, random = ~1|subject, data = Wrist2, method = "REML", correlation
      = corSymm(form = ~ 1|subject))
3 t3.lmm = lme(t3 ~ tom + group, random = ~1|subject, data = Wrist2, method = "REML", correlation
      = corSymm(form = ~ 1|subject))
```

In order to use `superLMM()`, a few changes are needed, such as no longer needing the `is`-argument, and other small changes. These changes are implemented in a new function, `superLMM2()`, which is identical to the original function, apart from these small changes. Table 7 shows that all covariance matrices are positive semi-definit.

Table 7: *This table contains the results of using* `superLMM2()$vcov` *on the clustered models.*

	`superLMM2()$vcov`		
	`t1.lmm`	`t2.lmm`	`t3.lmm`
$\mathbf{V}_0$	TRUE	TRUE	TRUE
$\mathbf{R}_0$	TRUE	TRUE	TRUE
$\boldsymbol{\Sigma}$	TRUE	TRUE	TRUE

Table 8 shows that group is irrelevant to the models.

Table 8: *The row for $anova1 are the p-values for group. The row for $anova2 are the p-values from the comparison of the model and the null-model. The rows for $intervals are the 95%-confidence intervals for $\beta_{gr.2}$ and $\beta_{gr.3}$.*

	superLMM2()$group		
	t1.lmm	t2.lmm	t3.lmm
$anova1	0.18362	0.16467	0.99949
$anova2	0.17245	0.15515	0.88540
$intervals	$\beta_{gr.2} \in [-0.94, 12.91]$ $\beta_{gr.3} \in [-1.83, 12.15]$	$\beta_{gr.2} \in [-7.69, 6.17]$ $\beta_{gr.3} \in [-1.55, 12.44]$	$\beta_{gr.2} \in [-5.85, 4.59]$ $\beta_{gr.3} \in [-4.61, 5.93]$

Table 9 shows that all model assumptions are satisfied apart from the intercepts and residuals of t3.lmm being Gaussian.

Table 9: *For both ass=1 (the intercepts) and ass=2 (the residuals), there are three rows. The first row are the p-values from the Bartlett test. The second row are the p-values from the Levene's test, and the third row are the p-values from the Shapiro-Wilk test.*

	superLMM2(ass = c(1, 2))		
	t1.lmm	t2.lmm	t3.lmm
ass = 1	0.29629	0.26388	0.77327
	0.14325	0.15739	0.54648
	0.97867	0.17024	0.00057
ass = 2	0.35055	0.60650	0.00000
	0.99783	0.99993	0.93671
	0.52707	0.40914	0.00000

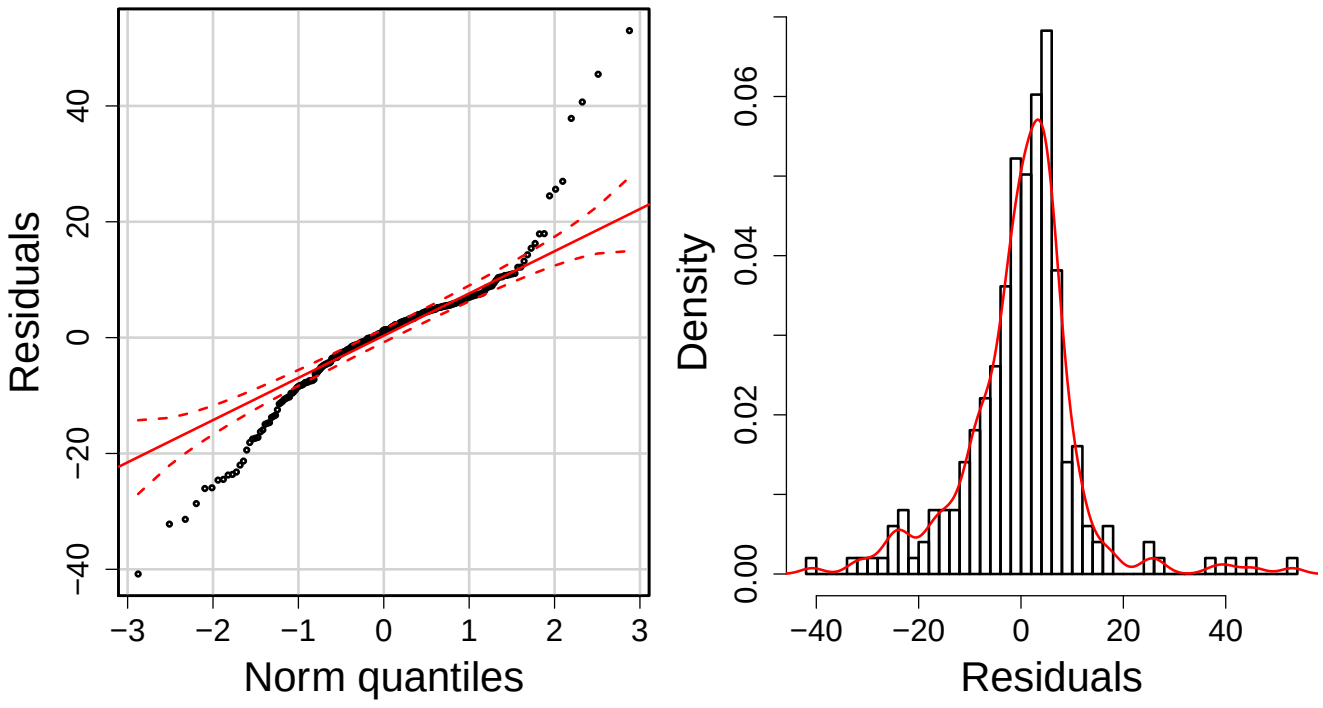

Figure 19: *Q-Q plot and histogram with density curve of the residuals of t3.lmm.*

Once again, the Q-Q plot and histogram of the residuals, seem to disagree on the distribution (see Figure 19). From the histogram, I would consider the residuals Gaussian, but not from the Q-Q plot. But, most importantly, there are no violations of the homoscedasticity assumptions. Likewise for the intercepts.

I have now confirmed that the clustered models give less problems wrt. satisfying model assumptions, than the longitudinal models. Table 8 shows that the conclusion about the relevance of the treatment groups have, as suspected, not changed. I will now confirm this using Kenward-Rogers approximation. As before, I set up models using `lmer()`. I am allowed to use this function despite it only having the exchangeable structure implemented. Just like with the longitudinal models, I have also here, in addition to `t1.lmm`, `t2.lmm` and `t3.lmm`, set up the equivalent "in"- and "ex"-models and tested with `anova()`, that there is no significant difference in either of the models no matter the covariance structure. This time the correlation parameters seemed more fitting with the numerically largest correlation for all the "un"-models being 0.253, and the numerically largest for all the "ex"-models being 0.0014. Thus with clustered LMMs, we have results from `anova()` and the estimated correlation parameters, that are much more in agreement with each other than they were for the longitudinal models. And we may assume no correlation between different types of movement. Table 10 shows the p-values from using both `KRmodcomp()` and `anova()` on the "ex"-models made with `lmer()`. These results confirm that H_0^{main} cannot be rejected.

Table 10: *The p-values from `KRmodcomp()` and `anova()` with the full "ex"-models of `t1`, `t2` and `t3`, respectively, and their corresponding null-models as input.*

	t1	t2	t3
KRmodcomp()	0.16719	0.20460	0.92591
anova()	0.15634	0.19278	0.92324

8.2 Clustered GEE models

In Figure 18, we saw that the assumption of `t3` being Gaussian may not be correct. Therefore, in this section, I show the results from modelling `t1`, `t2` and `t3` with GEE models releaving me from making any assumptions about the distribution of the response variables.

As in Section 8.1, I will only set up models with unstructured covariances in the main analysis, and I will use models with independence structure to see if we may assume no correlation between types of movement. The models are:

```
1 t1.gee = geeglm(t1 ~ tom + group, id = subject, data = Wrist2, corstr = "unstructured", family =
      gaussian(link = "identity"))
2 t2.gee = geeglm(t2 ~ tom + group, id = subject, data = Wrist2, corstr = "unstructured", family =
      gaussian(link = "identity"))
3 t3.gee = geeglm(t3 ~ tom + group, id = subject, data = Wrist2, corstr = "unstructured", family =
      gaussian(link = "identity"))
```

After replacing `Wrist$group` with `Wrist2$group` in `superGEE()`, giving me the almost identical function `superGEE2()`, I can now analyse the models.

Table 11 shows that the residuals are homoscedastic and almost all are Gaussian. Just like for the clustered LMM for `t3`, the residuals of `t3.gee` are not Gaussian according to the Shapiro-Wilk test, and once

again, the Q-Q plot and histogram seem to disagree with each other. These plots are not shown here, but they are very similar to Figure 19.

Table 11: *The rows for `test="homo"` are the p-values of the Bartlett and Levene's test, respectively, and the row for `test="norm"` are the p-values of the Shapiro-Wilk test.*

	superGEE2(test = c("homo", "norm"))		
	t1.gee	t2.gee	t3.gee
test = "homo"	0.35055	0.60650	0.00000
	0.99783	0.99993	0.93671
test = "norm"	0.10500	0.24098	0.00000

Next, I will test if there is any difference between `t1.gee`, `t2.gee` and `t3.gee` and their corresponding "in"-models. Like in Section 7.4, I let $\boldsymbol{\delta}$ denote the differences in the model based and empirical standard errors. I have gathered the $\boldsymbol{\delta}$'s in matrices for `t1`, `t2` and `t3` shown below:

```
> modcomp.delta.t1
   (Intercept)     tomef     tomur group2    group3
un     0.05496  0.15883  0.14182 0.12770 -0.36816
in     0.42541 -0.57795 -0.43274 1.06641  0.56826

> modcomp.delta.t2
   (Intercept)     tomef     tomur  group2    group3
un    -0.33969 -0.34815 -0.09747 0.05030 -0.51969
in     0.15816 -0.93454 -0.68378 1.32349  0.74645

> modcomp.delta.t3
   (Intercept)     tomef     tomur   group2    group3
un    -0.42868 -0.39995 -0.03886 -0.14085 -0.03231
in    -0.23999 -0.58775 -0.13083  0.30161  0.49650
```

We see that the $\boldsymbol{\delta}$'s for the "un"-models tend to be smaller than those of the "in"-models, indicating that there is correlation between the different types of movement. The sizes of $\hat{\mathbf{a}}$ (see below) seems much more fitting for these model, than they were for the longitudinal GEE-models in Chapter 7.

```
> superGEE2(t1.gee)$ass.param
alpha.1:2 alpha.1:3 alpha.2:3
    0.515     0.418     0.442

> superGEE2(t2.gee)$ass.param
alpha.1:2 alpha.1:3 alpha.2:3
    0.429     0.429     0.850

> superGEE2(t3.gee)$ass.param
alpha.1:2 alpha.1:3 alpha.2:3
    0.168     0.084     0.371
```

Wtih clustered GEE models, we now have that the results from the $\boldsymbol{\delta}$'s and the $\hat{\mathbf{a}}$'s that are in much more agreement with each other, than they were for the longitudinal models.

Finally, Table 12 shows `group` is irreleant to the models.

Table 12: *The row for `$anova1` are the p-values for `group`. The row for `$anova2` are the p-values from the comparison of the model and the null-model. The rows for `$esticon` are the 95%-confidence intervals for $\beta_{gr.2}$ and $\beta_{gr.3}$, and the p-values for these coefficients.*

	`superGEE2()$group`		
	`t1.gee`	`t2.gee`	`t3.gee`
`$anova1`	0.12889	0.10514	0.88910
`$anova2`	0.12889	0.10514	0.88910
`$esticon`	$\beta_{\text{gr.2}} \in [-0.70, 12.88]$, 0.07862 $\beta_{\text{gr.3}} \in [-0.98, 10.77]$, 0.10261	$\beta_{\text{gr.2}} \in [-7.25, 6.44]$, 0.90737 $\beta_{\text{gr.3}} \in [-0.47, 11.10]$, 0.07198	$\beta_{\text{gr.2}} \in [-4.96, 4.36]$, 0.90040 $\beta_{\text{gr.3}} \in [-4.06, 5.78]$, 0.73099

8.3 Conclusion

No matter if we use longitudinal or clustered models, whether we use ANOVA, MANOVA, LMMs or GEE models, there are no tests that indicate, Ibuprofen should have any significant beneficial effect on the recovery of wrist functions, and thus I conclude, that H_0^{main} cannot be rejected.

9 | Discussion

In this chapter, I present points for discussion. The points are presented in chronological order according to the sections in the report, where they appear, and are therefore not given in order of importance.

Subsection 4.2.2: Algorithm 4.2

Despite vigorous searching, I have not been able to find any sources either confirming or disproving that Algorithm 4.2 always converges, and that different choices for initial $\hat{\boldsymbol{\beta}}$ could result in different final $\hat{\boldsymbol{\beta}}$'s. The source, [8], from which I have based this algorithm, mentions nothing about the subject. From having a look at the source code for `lme()`, I cannot tell for sure if the function is based on Algorithm 4.2, but I must assume it is based on a similar algorithm. Hence, my questions about Algorithm 4.2 may be answered through looking at the source code for `lme()`. In the source code, I looked to see if a break is implemented in case convergence has not been reached within, say, 5000 iterations, meaning that convergence is probably not going to happen. I found nothing of the sort. However, the argument `MaxIter` in `lme()` is used for allowing for more iterations in order to reach convergence. There is no upper limit to `MaxIter` (here, I am assuming an upper limit would mean, that if convergence has not been reach with this number of iterations, then convergence is not going to happen at all), which can mean one of two things: either no upper limit is needed as convergence will always happen for big enough `MaxIter` or it is assumed that the user knows that if the function has not converged even with, say, `MaxIter = 5000`, then the user should take that as a sign, that the function will not converge. Wrt. the choice for initial $\hat{\boldsymbol{\beta}}$, I believe it is chosen as $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ because it lowers the number of iterations needed to reach convergence, just like how the initial $\hat{\boldsymbol{\beta}}$ is chosen in Algorithm 6.2.

Subsection 4.3.2: For $\varphi = \sigma^2$, we can set $\mathbf{V}_0 = \sigma^2\mathbf{A}$

There is no (logical) explanation in [7] as to why $\mathbf{V}_0$ can be written as $\sigma^2\mathbf{A}$ when there is no correlation between subjects and autocorrelation within subjects. As

$$\mathbf{V}_0 = \sigma_u^2\mathbf{J}_{n\times n}\otimes\mathbf{R}_0 \;=\; \sigma_u^2\mathbf{J}_{n\times n}\otimes\sigma^2\mathbf{A}$$

only if $\sigma_u^2 = 0$ would $\mathbf{V}_0$ reduce to $\sigma^2\mathbf{A}$, but σ_u^2 is not zero. In A.14, we see that σ_u^2 does not vanish when deriving $\mathbf{V}_0^{-1}$. I have chosen to derive the estimating equations for σ_u^2, σ^2 and ρ in the case with autoregressive structure. I see that in [7], had I chosen a different structure, then σ_u^2 would not have been overlooked when deriving the estimating equation for σ^2. So either σ_u^2 was somehow forgotten in deriving the estimating equation for σ^2 or it was left out on purpose, and the explanation for this was either neglected or thought of as being so obvious, that the authors decided not to include it.

Section 5.4: Model comparison

When comparing the "ar"-, "un"- and "in"-models, I found that only for `ps` was the "un"-model prefered over the other two. But if I looked at the correlation parameters for all the "un"-models, it would seem the "un"-models for `ur` and `ef` probably also should have been prefered over the "in"-models because of the high correlations. But perhaps these correlation were not high at all. When testing whether $\hat{\beta}_i$ could be set equal to zero, I used the function `intervals()`, which gave an interval for $\hat{\beta}_i$. If zero was contained in the interval, we may set $\hat{\beta}_i = 0$. It would have been nice with such a test for the correlation parameters, so that if, say, $\hat{\rho}_{1,2} < c$, for some limit c, we may set $\hat{\rho}_{1,2} = 0$.

An upside to these slightly confusing results is, that it allows me to use Kenward-Rogers approximation. For the Kenward-Rogers approximation, I must use a function in `R` which only has the exchangeable structure implemented. This is no problem, as there is no significant difference in any of my models no matter the covariance structure.

Section 7.3: Identical residuals

With the GEE models, I get identical residuals for both the "ar"-, "un"- and "in"-models for `ps`, and likewise for `ur` and `ef`. This is despite having different $\mathbf{R}_0$, $\hat{\mathbf{a}}$, $\hat{\boldsymbol{\beta}}$ (not shown in the report), and so on in these models, and it does not stem from a coding error in `superGEE()`. Investigating this further, I looked at the residuals for the GEE models in Example 5 (the `respitory`-data) where I had an "in"- and an "un"-model. Here, I also got identical residuals. Therefore, I believe these odd identical residuals are not odd at all.

Section 7.4: Large association parameters

The estimated association parameters were a lot bigger than I would have expected. When looking at $\boldsymbol{\delta}$ for all three type of models for one type of movement (`modcomp.delta.ps`, for instance), $\boldsymbol{\delta}$ tended to either be a little smaller for the "ar"- and "un"-models or about the same as the "in"-model. Hence, I would expect $\hat{\mathbf{a}}$ to be small (although not zero). I cannot say how small, I would have expected it to be, but certainly not larger than 0.5. Or perhaps, these are very reasonable results. I am basing my concern on the large $\hat{\mathbf{a}}$'s on the results from Example 5. In this example, there was a clear difference in the "in"- and "un"-models when looking at their corresponding $\boldsymbol{\delta}$. All entries in $\boldsymbol{\delta}$ for the "un"-model was smaller by about a factor 10 than the corresponding entries in $\boldsymbol{\delta}$ for the "in"-model. This resulted in $\hat{\mathbf{a}} \in [0.209, 0.439]$. In the models for my data, instead of needing *all* entries in $\boldsymbol{\delta}$ for either the "ar"- or "un"-model to be smaller than those in $\boldsymbol{\delta}$ for the "in"-model, perhaps the large $\hat{\mathbf{a}}$ is a result of some entries in $\boldsymbol{\delta}$ for either the "ar"- or "un"-model being *more* than a factor 10 smaller than the corresponding entry in $\boldsymbol{\delta}$ for the "in"-model. So, despite some of the entries in the different $\boldsymbol{\delta}$'s being almost equal, the fact that other entries differ quite some, may be enough to say that the "in"-models are actually less fitting, and thus justifying the large $\hat{\mathbf{a}}$'s.

Chapter 8: Longitudinal vs. clustered models

When using clustered models instead of longitudinal models, I got results from comparing the "in"- and "un"-models that were much more in agreement with the estimated correlations between types of movement. This leads me to think, that when testing for the relevance of the treatment groups, the better set of models for this task, may be the clustered models. One concern I have, though, is the problems with linearity. In the longitudinal models, I just had to add a squared term of `time` to have linearity. Although I did have linearity for some patients in the clustered setup, for some patients this was certainly not so.

Other issues

I believe it is possible that how well and fast a patient recover wrist functions can be related to whether the injured hand is their dominant hand or not. After surgery, if the hand is stiff and sore, you might just choose not to use it if it is your non-dominant hand. Hence, the hand will stay stiff for a longer time, because it is not being used. If instead the injured hand is the dominant hand, you might just use it anyway despite it being stiff and sore simply because it is your instinct. Hence, the wrist will get more "exercise" and will loosen up faster. It is interesting what the results would be, had I been able to include an indicator in my models of whether the injured hand was the dominant one or not.

Appendix A

This appendix contains calculations and results used in the report.

A. 1 The *within-group* variance, s_w^2, is an unbiased estimator of the common variance, σ^2, if we have that $\mathrm{E}\left[s_w^2\right] = \sigma^2$:

$$\begin{aligned}
\mathrm{E}\left[s_w^2\right] = \frac{1}{g}\sum_{i=1}^{g}\mathrm{E}\left[s_i^2\right] &\overset{(2.1)}{=} \frac{1}{g}\sum_{i=1}^{g}\frac{1}{n-1}\mathrm{E}\left[\sum_{j=1}^{n}\left(Y_{ij}-\bar{Y}_{i\cdot}\right)^2\right] \\
&= \frac{1}{g(n-1)}\sum_{i=1}^{g}\mathrm{E}\left[\sum_{j=1}^{n}\left(Y_{ij}^2+\bar{Y}_{i\cdot}^2-2Y_{ij}\bar{Y}_{i\cdot}\right)\right] \\
&= \frac{1}{g(n-1)}\sum_{i=1}^{g}\mathrm{E}\left[\sum_{j=1}^{n}Y_{ij}^2+n\bar{Y}_{i\cdot}^2-2\bar{Y}_{i\cdot}\sum_{j=1}^{n}Y_{ij}\right] \\
&= \frac{1}{g(n-1)}\sum_{i=1}^{g}\mathrm{E}\left[\sum_{j=1}^{n}Y_{ij}^2+n\bar{Y}_{i\cdot}^2-2\bar{Y}_{i\cdot}^2\right] \;=\; \frac{1}{g(n-1)}\sum_{i=1}^{g}\mathrm{E}\left[\sum_{j=1}^{n}Y_{ij}^2-n\bar{Y}_{i\cdot}^2\right] \\
&= \frac{1}{g(n-1)}\sum_{i=1}^{g}\left(\sum_{j=1}^{n}\mathrm{E}\left[Y_{ij}^2\right]-n\mathrm{E}\left[\bar{Y}_{i\cdot}^2\right]\right) \\
&= \frac{1}{g(n-1)}\sum_{i=1}^{g}\left(\sum_{j=1}^{n}\left(\mathrm{Var}\left[Y_{ij}\right]+\mathrm{E}\left[Y_{ij}\right]^2\right)-n\left(\mathrm{Var}\left[\bar{Y}_{i\cdot}\right]+\mathrm{E}\left[\bar{Y}_{i\cdot}\right]^2\right)\right) \\
&= \frac{1}{g(n-1)}\sum_{i=1}^{g}\left(\sum_{j=1}^{n}\left(\sigma^2+\mu_i^2\right)-n\left(\frac{\sigma^2}{n}+\mu_i^2\right)\right) \;=\; \frac{1}{g(n-1)}\sum_{i=1}^{g}\left(n\sigma^2+n\mu_i^2-\sigma^2-n\mu_i^2\right) \\
&= \frac{1}{g(n-1)}g\sigma^2(n-1) \;=\; \frac{1}{g}\sum_{i=1}^{g}\sigma^2=\sigma^2.
\end{aligned}$$

The *between-group* variance, s_b^2, is only an unbiased estimator of σ^2 if $\mu_i = \mu$, $\forall i$:

$$\begin{aligned}
\mathrm{E}\left[s_b^2\right] = \frac{n}{g-1}\mathrm{E}\left[\sum_{i=1}^{g}\left(\bar{Y}_{i\cdot}-\bar{Y}_{\cdot\cdot}\right)^2\right] &= \frac{n}{g-1}\sum_{i=1}^{g}\left(\mathrm{Var}\left[\bar{Y}_{i\cdot}\right]+\mathrm{E}\left[\bar{Y}_{i\cdot}\right]-\mathrm{Var}\left[\bar{Y}_{\cdot\cdot}\right]-\mathrm{E}\left[\bar{Y}_{\cdot\cdot}\right]\right) \\
&= \frac{n}{g-1}\sum_{i=1}^{g}\left(\frac{\sigma^2}{n}+\mu-\frac{1}{g^2}g\frac{\sigma^2}{n}-\frac{1}{g}g\mu\right) \;=\; \frac{n}{g-1}\sum_{i=1}^{g}\sigma^2\left(\frac{1}{n}-\frac{1}{nk}\right) \\
&= \frac{n}{g-1}g\sigma^2\left(\frac{g-1}{nk}\right) \;=\; \sigma^2.
\end{aligned}$$

To get from the 1st to the 2nd equality, I used the approach from $\mathrm{E}\left[s_w^2\right]$ when going from the 2nd to 8th equality. Clearly, if $\mu_i \neq \mu$ for some i, then $\mathrm{E}\left[s_b^2\right] \neq \sigma^2$.

A. 2 The *total sum of squares and cross product* matrix, **T**, can be written as a sum of

$$\mathbf{E} = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i\cdot})(\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i\cdot})^T$$

and

$$\mathbf{H} = \sum_{i=1}^{g} n_i(\bar{\mathbf{y}}_{i\cdot} - \bar{\mathbf{y}}_{\cdot\cdot})(\bar{\mathbf{y}}_{i\cdot} - \bar{\mathbf{y}}_{\cdot\cdot})^T:$$

$$\begin{aligned}
\mathbf{T} = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(\mathbf{y}_{ij} - \bar{\mathbf{y}}_{\cdot\cdot})(\mathbf{y}_{ij} - \bar{\mathbf{y}}_{\cdot\cdot})^T &= \sum_{i=1}^{g}\sum_{j=1}^{n_i}\Big((\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i\cdot}) + (\bar{\mathbf{y}}_{i\cdot} - \bar{\mathbf{y}}_{\cdot\cdot})\Big)\Big((\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i\cdot}) + (\bar{\mathbf{y}}_{i\cdot} - \bar{\mathbf{y}}_{\cdot\cdot})\Big)^T \\
&= \sum_{i=1}^{g}\sum_{j=1}^{n_i}\Big((\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i\cdot})(\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i\cdot})^T + 2(\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i\cdot})(\bar{\mathbf{y}}_{i\cdot} - \bar{\mathbf{y}}_{\cdot\cdot})^T + (\bar{\mathbf{y}}_{i\cdot} - \bar{\mathbf{y}}_{\cdot\cdot})(\bar{\mathbf{y}}_{i\cdot} - \bar{\mathbf{y}}_{\cdot\cdot})^T\Big) \\
&= \sum_{i=1}^{g}\sum_{j=1}^{n_i}(\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i\cdot})(\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i\cdot})^T + \sum_{i=1}^{g} n_i(\bar{\mathbf{y}}_{i\cdot} - \bar{\mathbf{y}}_{\cdot\cdot})(\bar{\mathbf{y}}_{i\cdot} - \bar{\mathbf{y}}_{\cdot\cdot})^T + 2\sum_{i=1}^{g}\sum_{j=1}^{n_i}(\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i\cdot})(\bar{\mathbf{y}}_{i\cdot} - \bar{\mathbf{y}}_{\cdot\cdot})^T \\
&= \sum_{i=1}^{g}\sum_{j=1}^{n_i}(\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i\cdot})(\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i\cdot})^T + \sum_{i=1}^{g} n_i(\bar{\mathbf{y}}_{i\cdot} - \bar{\mathbf{y}}_{\cdot\cdot})(\bar{\mathbf{y}}_{i\cdot} - \bar{\mathbf{y}}_{\cdot\cdot})^T + 2\sum_{i=1}^{g}\sum_{j=1}^{n_i}\Big(\mathbf{y}_{ij}\bar{\mathbf{y}}_{i\cdot}^T - \mathbf{y}_{ij}\bar{\mathbf{y}}_{\cdot\cdot}^T - \bar{\mathbf{y}}_{i\cdot}\bar{\mathbf{y}}_{i\cdot}^T + \bar{\mathbf{y}}_{i\cdot}\bar{\mathbf{y}}_{\cdot\cdot}^T\Big) \\
&= \sum_{i=1}^{g}\sum_{j=1}^{n_i}(\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i\cdot})(\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i\cdot})^T + \sum_{i=1}^{g} n_i(\bar{\mathbf{y}}_{i\cdot} - \bar{\mathbf{y}}_{\cdot\cdot})(\bar{\mathbf{y}}_{i\cdot} - \bar{\mathbf{y}}_{\cdot\cdot})^T + 2\sum_{i=1}^{g}\sum_{j=1}^{n_i}\Big(\mathbf{y}_{ij}\bar{\mathbf{y}}_{i\cdot}^T - \mathbf{y}_{ij}\bar{\mathbf{y}}_{\cdot\cdot}^T\Big) \\
&\quad - 2\sum_{i=1}^{g}\big(n_i\bar{\mathbf{y}}_{i\cdot}\bar{\mathbf{y}}_{i\cdot}^T + n_i\bar{\mathbf{y}}_{i\cdot}\bar{\mathbf{y}}_{\cdot\cdot}^T\big) \\
&= \sum_{i=1}^{g}\sum_{j=1}^{n_i}(\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i\cdot})(\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i\cdot})^T + \sum_{i=1}^{g} n_i(\bar{\mathbf{y}}_{i\cdot} - \bar{\mathbf{y}}_{\cdot\cdot})(\bar{\mathbf{y}}_{i\cdot} - \bar{\mathbf{y}}_{\cdot\cdot})^T + 2\sum_{i=1}^{g}\sum_{j=1}^{n_i}\Big(\mathbf{y}_{ij}\bar{\mathbf{y}}_{i\cdot}^T - \mathbf{y}_{ij}\bar{\mathbf{y}}_{\cdot\cdot}^T\Big) \\
&\quad - 2\sum_{i=1}^{g}\left(\sum_{j=1}^{n_i}\mathbf{y}_{ij}\bar{\mathbf{y}}_{i\cdot}^T + \sum_{j=1}^{n_i}\mathbf{y}_{ij}\bar{\mathbf{y}}_{\cdot\cdot}^T\right) \\
&= \sum_{i=1}^{g}\sum_{j=1}^{n_i}(\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i\cdot})(\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i\cdot})^T + \sum_{i=1}^{g} n_i(\bar{\mathbf{y}}_{i\cdot} - \bar{\mathbf{y}}_{\cdot\cdot})(\bar{\mathbf{y}}_{i\cdot} - \bar{\mathbf{y}}_{\cdot\cdot})^T = \mathbf{E} + \mathbf{H}.
\end{aligned}$$

A. 3 Let λ be an eigenvalue of $\mathbf{HE}^{-1} = \mathbf{A}$ with $\mathbf{v}$ as the corresponding eigenvector. Then $\mathbf{Av} = \lambda\mathbf{v}$. Let $\mathbf{B} = \mathbf{A} + \mathbf{I}_{p\times p}$. Then

$$\mathbf{Bv} = \big(\mathbf{A} + \mathbf{I}_{p\times p}\big)\mathbf{v} = \mathbf{Av} + \mathbf{I}_{p\times p}\mathbf{v} = \lambda\mathbf{v} + \mathbf{v} = (\lambda + 1)\mathbf{v}.$$

This means that $\lambda + 1$ is an eigenvalue of **B**. In Appendix B (see B.4), it is explained that the determinant of a matrix equals the product of the eigenvalues. Using this and the fact that $|\mathbf{I}_{p\times p}| = 1$, we

get

$$\Lambda_W = \frac{|\mathbf{E}|}{|\mathbf{H}+\mathbf{E}|} = \frac{|\mathbf{E}||\mathbf{E}^{-1}|}{|\mathbf{H}+\mathbf{E}||\mathbf{E}^{-1}|} = \frac{|\mathbf{I}_{p\times p}|}{|\mathbf{A}+\mathbf{I}_{p\times p}|}$$
$$= |\mathbf{A}+\mathbf{I}_{p\times p}|^{-1} = |\mathbf{B}|^{-1} = \prod_{i=1}^{p}\frac{1}{1+\lambda_i}.$$

Let $\mathbf{C} = \mathbf{A}(\mathbf{A}+\mathbf{I}_{p\times p})^{-1} = \mathbf{A}\mathbf{B}^{-1}$. Then

$$\mathbf{C}\mathbf{v} = \mathbf{A}\mathbf{B}^{-1}\mathbf{v} = \mathbf{A}(\lambda+1)^{-1}\mathbf{v} = \frac{1}{\lambda+1}\mathbf{A}\mathbf{v} = \frac{\lambda}{\lambda+1}\mathbf{v}.$$

This means that $\frac{\lambda}{\lambda+1}$ is an eigenvalue of $\mathbf{C}$. In Appendix B (B.4), it is explained that the trace of a matrix equals the sum of the eigenvalues. Using this and the trace is invariant under cyclic permutations, we get

$$\Lambda_P = \operatorname{tr}\left\{\mathbf{H}(\mathbf{H}+\mathbf{E})^{-1}\right\} = \operatorname{tr}\left\{\mathbf{H}(\mathbf{H}+\mathbf{E})^{-1}\mathbf{E}\mathbf{E}^{-1}\right\}$$
$$= \operatorname{tr}\left\{\mathbf{E}^{-1}\mathbf{H}(\mathbf{H}+\mathbf{E})^{-1}\left(\mathbf{E}^{-1}\right)^{-1}\right\} = \operatorname{tr}\left\{\mathbf{A}(\mathbf{H}\mathbf{E}^{-1}+\mathbf{E}\mathbf{E}^{-1})^{-1}\right\}$$
$$= \operatorname{tr}\left\{\mathbf{A}(\mathbf{A}+\mathbf{I}_{p\times p})^{-1}\right\} = \operatorname{tr}\{\mathbf{C}\} = \sum_{i=1}^{p}\frac{\lambda_i}{1+\lambda_i}.$$

A. 4 Let $\operatorname{Var}[U_i] = \sigma_u^2$, $\forall i$, and $\operatorname{Var}\left[\epsilon_{ij}\right] = \sigma^2$, $\forall j$, and remember that u_i and ϵ_{ij} are idependent, and the ϵ_{ij}s are mutually independent. Observations within groups are correlated:

$$\operatorname{Corr}\left[Y_{ij}, Y_{il}\right] = \frac{\operatorname{Cov}\left[Y_{ij}, Y_{il}\right]}{\sqrt{\operatorname{Var}\left[Y_{ij}\right]\operatorname{Var}\left[Y_{il}\right]}} = \frac{\frac{1}{2}\left(\operatorname{Var}\left[Y_{ij}+Y_{il}\right] - \operatorname{Var}\left[Y_{ij}\right] - \operatorname{Var}\left[Y_{il}\right]\right)}{\sqrt{\operatorname{Var}\left[Y_{ij}\right]\operatorname{Var}\left[Y_{il}\right]}}$$
$$= \frac{\frac{1}{2}\left(\operatorname{Var}\left[Y_{ij}+Y_{il}\right] - \left(\sigma_u^2+\sigma^2\right) - \left(\sigma_u^2+\sigma^2\right)\right)}{\sqrt{\left(\sigma_u^2+\sigma^2\right)\left(\sigma_u^2+\sigma^2\right)}} = \frac{\frac{1}{2}\operatorname{Var}\left[Y_{ij}+Y_{il}\right] - \left(\sigma_u^2+\sigma^2\right)}{\sigma_u^2+\sigma^2}$$
$$= \frac{\frac{1}{2}\operatorname{Var}\left[2\mu + (x_{ij}+x_{il})\beta + 2U_i + \epsilon_{ij} + \epsilon_{il}\right] - \sigma_u^2 - \sigma^2}{\sigma_u^2+\sigma^2}$$
$$= \frac{2\operatorname{Var}[U_i] + \frac{1}{2}\operatorname{Var}\left[\epsilon_{ij}+\epsilon_{il}\right] - \sigma_u^2 - \sigma^2}{\sigma_u^2+\sigma^2} = \frac{\frac{1}{2}\sigma^2 + \frac{1}{2}\sigma^2 - \sigma^2 + \sigma_u^2}{\sigma_u^2+\sigma^2}$$
$$= \frac{\sigma_u^2}{\sigma_u^2+\sigma^2}.$$

In the 2nd equality, $\operatorname{Var}[X+Y] = \operatorname{Var}[X] + \operatorname{Var}[Y] + 2\operatorname{Cov}[X,Y]$ is used to change the numerator.

As $\operatorname{Cov}[X,Y] = 0$ when X and Y are independent, obviously $\operatorname{Corr}\left[Y_{ij}, Y_{kl}\right] = 0$ when $i \neq k$, i.e. when Y_{ij} and Y_{kl} are observations from different groups.

A. 5 The score function wrt. $\boldsymbol{\beta}$ is

$$S_{\boldsymbol{\beta}}(\boldsymbol{\beta};\mathbf{y}) = \frac{\partial}{\partial\boldsymbol{\beta}}\ell(\boldsymbol{\beta};\mathbf{y}) \overset{(4.10)}{=} -\frac{1}{2}\frac{\partial}{\partial\boldsymbol{\beta}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) = -\frac{1}{2}\frac{\partial}{\partial\boldsymbol{\beta}}\left(\mathbf{y}\mathbf{V}^{-1}\mathbf{y}-2\mathbf{y}^T\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta}\right)$$

$$= \frac{\partial}{\partial\boldsymbol{\beta}}\left(\mathbf{y}^T\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta}-\frac{1}{2}\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta}\right) = \left(\mathbf{y}^T\mathbf{V}^{-1}\mathbf{X}\right)^T-\left(\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\right)^T = \mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}-\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta}.$$

MLE of $\boldsymbol{\beta}$ is then

$$S_{\boldsymbol{\beta}}(\boldsymbol{\beta};\mathbf{y}) = \mathbf{0} \Leftrightarrow \mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}-\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta} = \mathbf{0} \Leftrightarrow \hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}.$$

A. 6 If we want to predict the random variable X as much as possible by a constant c, the best predictor for X will be $c = \mathrm{E}[X]$:

$$\mathrm{E}\left[(X-c)^2\right] = \mathrm{E}\left[(X-\mathrm{E}[X]+\mathrm{E}[X]-c)^2\right] = \mathrm{E}\left[(X-\mathrm{E}[X])^2\right]+2(\mathrm{E}[X]-c)\mathrm{E}[X-\mathrm{E}[X]]+(\mathrm{E}[X]-c)^2$$

$$= \mathrm{E}\left[(X-\mathrm{E}[X])^2\right]+(\mathrm{E}[X]-c)^2 = \mathrm{Var}[X]+(\mathrm{E}[X]-c)^2.$$

A. 7 In order to find $\mathbf{u}$, we must solve $S_u(\boldsymbol{\beta},\boldsymbol{\varphi};\mathbf{u},\mathbf{y}) = \mathbf{0}$, where the score function is given as

$$\frac{\partial}{\partial\mathbf{u}}\ell(\boldsymbol{\beta},\boldsymbol{\varphi},\mathbf{u};\mathbf{y}) = -\frac{1}{2}\frac{\partial}{\partial\mathbf{u}}\left(\log(|\mathbf{R}(\boldsymbol{\varphi})|)+(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u})^T\mathbf{R}(\boldsymbol{\varphi})^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u})+\log(|\mathbf{G}(\boldsymbol{\varphi})|)+\mathbf{u}^T\mathbf{G}(\boldsymbol{\varphi})^{-1}\mathbf{u}\right)$$

$$= -\frac{1}{2}\frac{\partial}{\partial\mathbf{u}}\left((\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u})^T\mathbf{R}(\boldsymbol{\varphi})^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u})+\mathbf{u}^T\mathbf{G}(\boldsymbol{\varphi})^{-1}\mathbf{u}\right)$$

$$= -\frac{1}{2}\frac{\partial}{\partial\mathbf{u}}\left(\mathbf{y}^T\mathbf{R}(\boldsymbol{\varphi})^{-1}\mathbf{y}+\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{R}(\boldsymbol{\varphi})^{-1}\mathbf{X}\boldsymbol{\beta}+\mathbf{u}^T\mathbf{Z}^T\mathbf{R}(\boldsymbol{\varphi})^{-1}\mathbf{Z}\mathbf{u}-2\mathbf{y}^T\mathbf{R}(\boldsymbol{\varphi})^{-1}\mathbf{X}\boldsymbol{\beta}\right.$$
$$\left.-2\mathbf{y}^T\mathbf{R}(\boldsymbol{\varphi})^{-1}\mathbf{Z}\mathbf{u}-2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{R}(\boldsymbol{\varphi})^{-1}\mathbf{Z}\mathbf{u}+\mathbf{u}^T\mathbf{G}(\boldsymbol{\varphi})^{-1}\mathbf{u}\right)$$

$$= -\frac{1}{2}\frac{\partial}{\partial\mathbf{u}}\left(\mathbf{u}^T\mathbf{Z}^T\mathbf{R}(\boldsymbol{\varphi})^{-1}\mathbf{Z}\mathbf{u}-2\mathbf{y}^T\mathbf{R}(\boldsymbol{\varphi})^{-1}\mathbf{Z}\mathbf{u}+2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{R}(\boldsymbol{\varphi})^{-1}\mathbf{Z}\mathbf{u}+\mathbf{u}^T\mathbf{G}(\boldsymbol{\varphi})^{-1}\mathbf{u}\right)$$

$$= -\left(\mathbf{u}^T\mathbf{Z}^T\mathbf{R}(\boldsymbol{\varphi})^{-1}\mathbf{Z}\right)^T+\left(\mathbf{y}^T\mathbf{R}(\boldsymbol{\varphi})^{-1}\mathbf{Z}\right)^T-\left(\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{R}(\boldsymbol{\varphi})^{-1}\mathbf{Z}\right)^T-\mathbf{G}(\boldsymbol{\varphi})^{-1}\mathbf{u}$$

$$= \mathbf{Z}^T\mathbf{R}(\boldsymbol{\varphi})^{-1}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u}\right)-\mathbf{G}(\boldsymbol{\varphi})^{-1}\mathbf{u}.$$

A. 8 We have that $\mathbf{U} \sim N_m(\mathbf{0},\mathbf{G})$ and $\boldsymbol{\epsilon} \sim N_N(\mathbf{0},\mathbf{R})$, and $\mathbf{u}$ and $\boldsymbol{\epsilon}$ are independent. Thus

$$\mathrm{Var}[\mathbf{Y}] = \mathrm{Var}\left[\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{U}+\boldsymbol{\epsilon}\right] = \mathbf{Z}\mathrm{Var}[\mathbf{U}]\mathbf{Z}^T+\mathrm{Var}[\boldsymbol{\epsilon}] = \mathbf{Z}\mathbf{G}\mathbf{Z}^T+\mathbf{R}.$$

A. 9 In the following are the calculations showing that the restricted log-likelihood can be written as $\ell_{\mathrm{p}}(\boldsymbol{\varphi}) - \frac{1}{2}\log\left(|\mathbf{X}^T\mathbf{V}(\boldsymbol{\varphi})^{-1}\mathbf{X}|\right)$. Let

$$\mathbf{A}(\boldsymbol{\varphi}) = \mathbf{X}^T\mathbf{V}(\boldsymbol{\varphi})^{-1}\mathbf{X} \tag{1}$$
$$\mathbf{B}(\boldsymbol{\varphi}) = \mathbf{A}(\boldsymbol{\varphi})^{-1}\mathbf{X}^T\mathbf{V}(\boldsymbol{\varphi})^{-1}. \tag{2}$$

The term inside $\exp(\cdot)$ in $L(\boldsymbol{\beta},\boldsymbol{\varphi};\mathbf{y})$ (Equation (4.9)) can be rewritten as

$$\begin{aligned}
(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{V}(\boldsymbol{\varphi})^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) &= \mathbf{y}^T\mathbf{V}(\boldsymbol{\varphi})^{-1}\mathbf{y} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{V}(\boldsymbol{\varphi})^{-1}\mathbf{X}\boldsymbol{\beta} - 2\mathbf{y}^T\mathbf{V}(\boldsymbol{\varphi})^{-1}\mathbf{X}\boldsymbol{\beta} \\
&= \mathbf{y}^T\mathbf{V}(\boldsymbol{\varphi})^{-1}\mathbf{y} + \boldsymbol{\beta}^T\mathbf{A}(\boldsymbol{\varphi})\boldsymbol{\beta} - 2\mathbf{y}^T\mathbf{V}(\boldsymbol{\varphi})^{-1}\mathbf{X}\boldsymbol{\beta} \\
&= \mathbf{y}^T\mathbf{V}(\boldsymbol{\varphi})^{-1}\mathbf{y} + \boldsymbol{\beta}^T\mathbf{A}(\boldsymbol{\varphi})\boldsymbol{\beta} - 2\mathbf{y}^T\mathbf{V}(\boldsymbol{\varphi})^{-1}\mathbf{X}\boldsymbol{\beta} + \mathbf{y}^T\mathbf{V}(\boldsymbol{\varphi})^{-1}\mathbf{X}\mathbf{A}(\boldsymbol{\varphi})^{-1}\mathbf{X}^T\mathbf{V}(\boldsymbol{\varphi})^{-1}\mathbf{y} \\
&\quad - \mathbf{y}^T\mathbf{V}(\boldsymbol{\varphi})^{-1}\mathbf{X}\mathbf{A}(\boldsymbol{\varphi})^{-1}\mathbf{X}^T\mathbf{V}(\boldsymbol{\varphi})^{-1}\mathbf{y} \\
&= \left(\boldsymbol{\beta}^T - \mathbf{y}^T\mathbf{V}(\boldsymbol{\varphi})^{-1}\mathbf{X}\mathbf{A}(\boldsymbol{\varphi})^{-1}\right)\left(\mathbf{A}(\boldsymbol{\varphi})\boldsymbol{\beta} - \mathbf{X}^T\mathbf{V}(\boldsymbol{\varphi})^{-1}\mathbf{y}\right) + \mathbf{y}^T\mathbf{V}(\boldsymbol{\varphi})^{-1}\mathbf{y} \\
&\quad - \mathbf{y}^T\mathbf{V}(\boldsymbol{\varphi})^{-1}\mathbf{X}\mathbf{A}(\boldsymbol{\varphi})^{-1}\mathbf{X}^T\mathbf{V}(\boldsymbol{\varphi})^{-1}\mathbf{y} \\
&= \left(\boldsymbol{\beta}^T - \mathbf{y}^T\mathbf{V}(\boldsymbol{\varphi})^{-1}\mathbf{X}\mathbf{A}(\boldsymbol{\varphi})^{-1}\right)\mathbf{A}(\boldsymbol{\varphi})\left(\boldsymbol{\beta} - \mathbf{A}(\boldsymbol{\varphi})^{-1}\mathbf{X}^T\mathbf{V}(\boldsymbol{\varphi})^{-1}\mathbf{y}\right) + \mathbf{y}^T\mathbf{V}(\boldsymbol{\varphi})^{-1}\mathbf{y} \\
&\quad - \mathbf{y}^T\mathbf{V}(\boldsymbol{\varphi})^{-1}\mathbf{X}\mathbf{A}(\boldsymbol{\varphi})^{-1}\mathbf{X}^T\mathbf{V}(\boldsymbol{\varphi})^{-1}\mathbf{y} \\
&= \left(\boldsymbol{\beta} - \mathbf{A}(\boldsymbol{\varphi})\mathbf{X}^T\mathbf{V}(\boldsymbol{\varphi})^{-1}\mathbf{y}\right)^T\mathbf{A}(\boldsymbol{\varphi})\left(\boldsymbol{\beta} - \mathbf{A}(\boldsymbol{\varphi})^{-1}\mathbf{X}^T\mathbf{V}(\boldsymbol{\varphi})^{-1}\mathbf{y}\right) + \mathbf{y}^T\mathbf{V}(\boldsymbol{\varphi})^{-1}\mathbf{y} \\
&\quad - \mathbf{y}^T\left(\mathbf{A}(\boldsymbol{\varphi})^{-1}\mathbf{X}^T\mathbf{V}(\boldsymbol{\varphi})^{-1}\right)^T\mathbf{A}(\boldsymbol{\varphi})\left(\mathbf{A}(\boldsymbol{\varphi})^{-1}\mathbf{X}^T\mathbf{V}(\boldsymbol{\varphi})^{-1}\right)\mathbf{y} \\
&= \left(\boldsymbol{\beta} - \mathbf{B}(\boldsymbol{\varphi})\mathbf{y}\right)^T\mathbf{A}(\boldsymbol{\varphi})\left(\boldsymbol{\beta} - \mathbf{B}(\boldsymbol{\varphi})\mathbf{y}\right) + \mathbf{y}^T\mathbf{V}(\boldsymbol{\varphi})^{-1}\mathbf{y} - \mathbf{y}^T\mathbf{B}(\boldsymbol{\varphi})^T\mathbf{A}(\boldsymbol{\varphi})\mathbf{B}(\boldsymbol{\varphi})\mathbf{y}.
\end{aligned} \tag{3}$$

Notice, that $\hat{\boldsymbol{\beta}}(\boldsymbol{\varphi})$ (Equation (4.13)) can be expressed as

$$\hat{\boldsymbol{\beta}}(\boldsymbol{\varphi}) \overset{(1)}{=} \mathbf{A}(\boldsymbol{\varphi})^{-1}\mathbf{X}^T\mathbf{V}(\boldsymbol{\varphi})^{-1}\mathbf{y} \overset{(2)}{=} \mathbf{B}(\boldsymbol{\varphi})\mathbf{y} \tag{4}$$

and so, the last two terms in Equation (3) can be written as

$$\begin{aligned}
\mathbf{y}^T\mathbf{V}(\boldsymbol{\varphi})^{-1}\mathbf{y} - \mathbf{y}^T\mathbf{B}(\boldsymbol{\varphi})^T\mathbf{A}(\boldsymbol{\varphi})\mathbf{B}(\boldsymbol{\varphi})\mathbf{y} &= \mathbf{y}^T\mathbf{V}(\boldsymbol{\varphi})^{-1}\mathbf{y} + \mathbf{y}^T\mathbf{B}(\boldsymbol{\varphi})^T\mathbf{A}(\boldsymbol{\varphi})\mathbf{B}(\boldsymbol{\varphi})\mathbf{y} - 2\mathbf{y}^T\mathbf{B}(\boldsymbol{\varphi})^T\mathbf{A}(\boldsymbol{\varphi})\mathbf{B}(\boldsymbol{\varphi})\mathbf{y} \\
&\overset{(4)}{=} \mathbf{y}^T\mathbf{V}(\boldsymbol{\varphi})^{-1}\mathbf{y} + \hat{\boldsymbol{\beta}}(\boldsymbol{\varphi})^T\mathbf{A}(\boldsymbol{\varphi})\hat{\boldsymbol{\beta}}(\boldsymbol{\varphi}) - 2\mathbf{y}^T\mathbf{B}(\boldsymbol{\varphi})^T\mathbf{A}(\boldsymbol{\varphi})\hat{\boldsymbol{\beta}}(\boldsymbol{\varphi}) \\
&\overset{(2)}{=} \mathbf{y}^T\mathbf{V}(\boldsymbol{\varphi})^{-1}\mathbf{y} + \hat{\boldsymbol{\beta}}(\boldsymbol{\varphi})^T\mathbf{A}(\boldsymbol{\varphi})\hat{\boldsymbol{\beta}}(\boldsymbol{\varphi}) - 2\mathbf{y}^T\mathbf{V}(\boldsymbol{\varphi})^{-1}\mathbf{X}\mathbf{A}(\boldsymbol{\varphi})^{-1}\mathbf{A}(\boldsymbol{\varphi})\hat{\boldsymbol{\beta}}(\boldsymbol{\varphi}) \\
&\overset{(1)}{=} \mathbf{y}^T\mathbf{V}(\boldsymbol{\varphi})^{-1}\mathbf{y} + \hat{\boldsymbol{\beta}}(\boldsymbol{\varphi})^T\mathbf{X}^T\mathbf{V}(\boldsymbol{\varphi})^{-1}\mathbf{X}\hat{\boldsymbol{\beta}}(\boldsymbol{\varphi}) - 2\mathbf{y}^T\mathbf{V}(\boldsymbol{\varphi})^{-1}\mathbf{X}\hat{\boldsymbol{\beta}}(\boldsymbol{\varphi}) \\
&= (\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}(\boldsymbol{\varphi}))^T\mathbf{V}(\boldsymbol{\varphi})^{-1}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}(\boldsymbol{\varphi})).
\end{aligned} \tag{5}$$

Using Equations (3) and (5), the term inside $\exp(\cdot)$ is then

$$(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{V}(\boldsymbol{\varphi})^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) = \left(\boldsymbol{\beta} - \mathbf{B}(\boldsymbol{\varphi})\mathbf{y}\right)^T\mathbf{A}(\boldsymbol{\varphi})\left(\boldsymbol{\beta} - \mathbf{B}(\boldsymbol{\varphi})\mathbf{y}\right) + (\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}(\boldsymbol{\varphi}))^T\mathbf{V}(\boldsymbol{\varphi})^{-1}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}(\boldsymbol{\varphi})). \tag{6}$$

Thus we get

$$\begin{aligned}
\int L(\boldsymbol{\beta},\boldsymbol{\varphi};\mathbf{y})\,d\boldsymbol{\beta} &\overset{(4.9)}{=} \int \left(\frac{1}{\sqrt{2\pi}^N}|\mathbf{V}(\boldsymbol{\varphi})|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{V}(\boldsymbol{\varphi})^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right)\right) d\boldsymbol{\beta} \\
&\overset{(6)}{=} \int \frac{1}{\sqrt{2\pi}^N}|\mathbf{V}(\boldsymbol{\varphi})|^{-\frac{1}{2}} \\
&\quad \cdot\exp\left(-\frac{1}{2}\left(\left(\boldsymbol{\beta}-\mathbf{B}(\boldsymbol{\varphi})\mathbf{y}\right)^T\mathbf{A}(\boldsymbol{\varphi})\left(\boldsymbol{\beta}-\mathbf{B}(\boldsymbol{\varphi})\mathbf{y}\right)+(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}(\boldsymbol{\varphi}))^T\mathbf{V}(\boldsymbol{\varphi})^{-1}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}(\boldsymbol{\varphi}))\right)\right) d\boldsymbol{\beta} \\
&= \frac{1}{\sqrt{2\pi}^N}|\mathbf{V}(\boldsymbol{\varphi})|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}(\boldsymbol{\varphi}))^T\mathbf{V}(\boldsymbol{\varphi})^{-1}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}(\boldsymbol{\varphi}))\right) \\
&\quad \cdot\int \exp\left(-\frac{1}{2}\left(\boldsymbol{\beta}-\mathbf{B}(\boldsymbol{\varphi})\mathbf{y}\right)^T\mathbf{A}(\boldsymbol{\varphi})\left(\boldsymbol{\beta}-\mathbf{B}(\boldsymbol{\varphi})\mathbf{y}\right)\right) d\boldsymbol{\beta} \\
&= \frac{1}{\sqrt{2\pi}^N}|\mathbf{V}(\boldsymbol{\varphi})|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}(\boldsymbol{\varphi}))^T\mathbf{V}(\boldsymbol{\varphi})^{-1}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}(\boldsymbol{\varphi}))\right) \\
&\quad \cdot\sqrt{2\pi}^N|\mathbf{A}(\boldsymbol{\varphi})|^{-\frac{1}{2}}\int \frac{1}{\sqrt{2\pi}^N}|\mathbf{A}(\boldsymbol{\varphi})|^{\frac{1}{2}}\exp\left(-\frac{1}{2}\left(\boldsymbol{\beta}-\mathbf{B}(\boldsymbol{\varphi})\mathbf{y}\right)^T\mathbf{A}(\boldsymbol{\varphi})\left(\boldsymbol{\beta}-\mathbf{B}(\boldsymbol{\varphi})\mathbf{y}\right)\right) d\boldsymbol{\beta} \\
&= \frac{1}{\sqrt{2\pi}^N}|\mathbf{V}(\boldsymbol{\varphi})|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}(\boldsymbol{\varphi}))^T\mathbf{V}(\boldsymbol{\varphi})^{-1}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}(\boldsymbol{\varphi}))\right)\cdot\sqrt{2\pi}^N|\mathbf{A}(\boldsymbol{\varphi})|^{-\frac{1}{2}}. \qquad (7)
\end{aligned}$$

The last equation deserves a short explanation: In the second to last equation, I multiplied by 1, so that the term in the integral became a likelihood function. In fact, the term in the integral has the form of a multivariate normal density with $\boldsymbol{\beta}$ as the variable. Integrating the density wrt. $\boldsymbol{\beta}$ is of course just 1, giving the last equation.

Letting the squared roots cancel each other out, the restricted log-likelihood thus becomes

$$\begin{aligned}
\ell_{\mathrm{R}}(\boldsymbol{\varphi}) &\overset{(4.15),(7)}{=} -\frac{1}{2}\log\left(|\mathbf{V}(\boldsymbol{\varphi})|\right)-\frac{1}{2}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}(\boldsymbol{\varphi}))^T\mathbf{V}(\boldsymbol{\varphi})^{-1}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}(\boldsymbol{\varphi}))-\frac{1}{2}\log\left(|\mathbf{A}(\boldsymbol{\varphi})|\right) \\
&\overset{(4.14)}{=} \ell_{\mathrm{p}}(\boldsymbol{\varphi})-\frac{1}{2}\log\left(|\mathbf{A}(\boldsymbol{\varphi})|\right) \overset{(1)}{=} \ell_{\mathrm{p}}(\boldsymbol{\varphi})-\frac{1}{2}\log\left(|\mathbf{X}^T\mathbf{V}(\boldsymbol{\varphi})^{-1}\mathbf{X}|\right).
\end{aligned}$$

A. 10 The BLUP of $\mathbf{u}$, when $\mathbf{G}=\sigma_u^2\mathbf{I}_{m\times m}$, $\mathbf{Z}=\mathbf{I}_{m\times m}\otimes\mathbf{1}_n^T$ and $\mathbf{V}=\mathbf{I}_{m\times m}\otimes\left(\sigma_u^2\mathbf{J}_{n\times n}+\mathbf{R}_0\right)$:

$$\begin{aligned}
\hat{\mathbf{u}} &\overset{(4.12)}{=} \mathbf{G}\mathbf{Z}^T\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}) \overset{(4.20),(4.21)}{=} \sigma_u^2\mathbf{I}_{m\times m}(\mathbf{I}_{m\times m}\otimes\mathbf{1}_n^T)\left(\mathbf{I}_{m\times m}\otimes\left(\sigma_u^2\mathbf{J}_{n\times n}+\mathbf{R}_0\right)\right)^{-1}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}) \\
&= \sigma_u^2(\mathbf{I}_{m\times m}\otimes 1)(\mathbf{I}_{m\times m}\otimes\mathbf{1}_n^T)\left(\mathbf{I}_{m\times m}\otimes\left(\sigma_u^2\mathbf{J}_{n\times n}+\mathbf{R}_0\right)^{-1}\right)(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}) \\
&= \sigma_u^2\left(\mathbf{I}_{m\times m}\otimes\mathbf{1}_n^T\left(\sigma_u^2\mathbf{J}_{n\times n}+\mathbf{R}_0\right)^{-1}\right)(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}) = \sigma_u^2\left(\mathbf{I}_{m\times m}\otimes\mathbf{1}_n^T\frac{1}{\sigma_u^2}\left(\mathbf{J}_{n\times n}+\frac{1}{\sigma_u^2}\mathbf{R}_0\right)^{-1}\right)(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}) \\
&= \left(\mathbf{I}_{m\times m}\otimes\mathbf{1}_n^T\left(\mathbf{1}_n\mathbf{1}_n^T+\frac{1}{\sigma_u^2}\mathbf{R}_0\right)^{-1}\right)(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}) = \left(\mathbf{I}_{m\times m}\otimes\mathbf{1}_n^T\mathbf{R}_0^{-1}\left(\mathbf{1}_n^T\mathbf{R}_0^{-1}\mathbf{1}_n+\frac{1}{\sigma_u^2}\right)^{-1}\right)(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}) \\
&= \frac{1}{r_0+\frac{1}{\sigma_u^2}}\left(\mathbf{I}_{m\times m}\otimes\mathbf{1}_n^T\mathbf{R}_0^{-1}\right)(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}),
\end{aligned}$$

remembering that $r_0=\mathbf{1}_n^T\mathbf{R}_0^{-1}\mathbf{1}_n$.

A. 11 The estimate of $\boldsymbol{\beta}$ is independent of $\boldsymbol{\varphi}$:

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y} \\
&= \left((\mathbf{1}_m \otimes \mathbf{I}_{n\times n})^T(\mathbf{I}_{m\times m}\otimes\mathbf{V}_0)^{-1}(\mathbf{1}_m\otimes\mathbf{I}_{n\times n})\right)^{-1}(\mathbf{1}_m\otimes\mathbf{I}_{n\times n})^T(\mathbf{I}_{m\times m}\otimes\mathbf{V}_0)^{-1}\mathbf{y} \\
&= \left((\mathbf{1}_m^T\otimes\mathbf{I}_{n\times n})(\mathbf{I}_{m\times m}\otimes\mathbf{V}_0^{-1})(\mathbf{1}_m\otimes\mathbf{I}_{n\times n})\right)^{-1}(\mathbf{1}_m^T\otimes\mathbf{I}_{n\times n})(\mathbf{I}_{m\times m}\otimes\mathbf{V}_0^{-1})\mathbf{y} \\
&= \left(\mathbf{1}_m^T\mathbf{I}_{m\times m}\mathbf{1}_m\otimes\mathbf{I}_{n\times n}\mathbf{V}_0^{-1}\mathbf{I}_{n\times n}\right)^{-1}(\mathbf{1}_m^T\mathbf{I}_{m\times m}\otimes\mathbf{I}_{n\times n}\mathbf{V}_0^{-1})\mathbf{y} \\
&= \left(m\mathbf{V}_0^{-1}\right)^{-1}(\mathbf{1}_m^T\otimes\mathbf{V}_0^{-1})\mathbf{y} \;=\; \frac{1}{m}(1\otimes\mathbf{V}_0)(\mathbf{1}_m^T\otimes\mathbf{V}_0^{-1})\mathbf{y} \;=\; \frac{1}{m}(\mathbf{1}_m^T\otimes\mathbf{I}_{n\times n})\mathbf{y} \\
&= \frac{1}{m}\begin{bmatrix}\mathbf{I}_{n\times n} & \mathbf{I}_{n\times n} & \cdots & \mathbf{I}_{n\times n}\end{bmatrix}\begin{bmatrix}\mathbf{y}_1 & \mathbf{y}_2 & \cdots & \mathbf{y}_m\end{bmatrix}^T \\
&= \frac{1}{m}\begin{bmatrix}1&0&\cdots&0&\cdots&1&0&\cdots&0\\0&1&\cdots&0&\cdots&0&1&\cdots&0\\\vdots&\vdots&\ddots&\vdots&\cdots&\vdots&\vdots&\ddots&\vdots\\0&0&\cdots&1&\cdots&0&0&\cdots&1\end{bmatrix}\begin{bmatrix}y_{11} & \cdots & y_{1n} & \cdots & \cdots & y_{m1} & \cdots & y_{mn}\end{bmatrix}^T \\
&= \frac{1}{m}\begin{bmatrix}\sum_{i=1}^m y_{i1}\\ \vdots \\ \sum_{i=1}^m y_{in}\end{bmatrix} \;=\; \begin{bmatrix}\bar{y}_{\cdot 1}\\ \vdots \\ \bar{y}_{\cdot n}\end{bmatrix}.
\end{aligned}$$

A. 12 The expression of the term $\mathbf{1}_n^T\mathbf{R}_0^{-1}(\mathbf{y}_i - \hat{\boldsymbol{\beta}})$ for the autoregressive covariance structure:

$$\begin{aligned}
\mathbf{1}_n^T\mathbf{R}_0^{-1}(\mathbf{y}_i - \hat{\boldsymbol{\beta}}) &= \frac{1}{\sigma^2(1-\rho^2)}\begin{bmatrix}1-\rho\\(1-\rho)^2\\\vdots\\(1-\rho)^2\\1-\rho\end{bmatrix}^T\begin{bmatrix}y_{i1}-\bar{y}_{\cdot 1}\\y_{i2}-\bar{y}_{\cdot 2}\\\vdots\\y_{i,n-1}-\bar{y}_{\cdot n-1}\\y_{in}-\bar{y}_{\cdot n}\end{bmatrix} \\
&= \frac{1}{\sigma^2(1-\rho^2)}\begin{bmatrix}(1-\rho)^2+\rho(1-\rho)\\(1-\rho)^2\\\vdots\\(1-\rho)^2\\(1-\rho)^2+\rho(1-\rho)\end{bmatrix}^T\begin{bmatrix}y_{i1}-\bar{y}_{\cdot 1}\\y_{i2}-\bar{y}_{\cdot 2}\\\vdots\\y_{i,n-1}-\bar{y}_{\cdot n-1}\\y_{in}-\bar{y}_{\cdot n}\end{bmatrix} \\
&= \frac{1}{\sigma^2(1-\rho^2)}\left((1-\rho)^2\sum_{j=1}^n\left(y_{ij}-\bar{y}_{i\cdot}\right)+\rho(1-\rho)(y_{i1}-\bar{y}_{\cdot 1}+y_{in}-\bar{y}_{\cdot n})\right) \\
&= \frac{1-\rho}{\sigma^2(1-\rho^2)}\Big((1-\rho)n(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})+\rho(y_{i1}-\bar{y}_{\cdot 1}+y_{in}-\bar{y}_{\cdot n})\Big).
\end{aligned}$$

A. 13 The expression of $\hat{u}_i$ for the autoregressive covariance structure:

$$\begin{aligned}
\hat{u}_i &\overset{(4.23)}{=} \frac{1}{\frac{(1-\rho)(2+(n-2)(1-\rho))}{\sigma^2(1-\rho^2)}+\frac{1}{\sigma_u^2}}\cdot\frac{1-\rho}{\sigma^2(1-\rho^2)}\Big((1-\rho)n(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})+\rho(y_{i1}-\bar{y}_{\cdot 1}+y_{in}-\bar{y}_{\cdot n})\Big)\\
&= \frac{1-\rho}{(1-\rho)(2+(n-2)(1-\rho))+\frac{\sigma^2(1-\rho^2)}{\sigma_u^2}}\Big((1-\rho)n(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})+\rho(y_{i1}-\bar{y}_{\cdot 1}+y_{in}-\bar{y}_{\cdot n})\Big)\\
&= \frac{(1-\rho)\frac{\sigma_u^2}{1-\rho}}{\Big((1-\rho)(2+(n-2)(1-\rho))+\frac{\sigma^2(1-\rho^2)}{\sigma_u^2}\Big)\frac{\sigma_u^2}{1-\rho}}\Big((1-\rho)n(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})+\rho(y_{i1}-\bar{y}_{\cdot 1}+y_{in}-\bar{y}_{\cdot n})\Big)\\
&= \frac{\sigma_u^2}{\sigma_u^2(2+(n-2)(1-\rho))+\frac{\sigma^2(1-\rho^2)}{1-\rho}}\Big((1-\rho)n(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})+\rho(y_{i1}-\bar{y}_{\cdot 1}+y_{in}-\bar{y}_{\cdot n})\Big)\\
&= \frac{\sigma_u^2\Big((1-\rho)n(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})+\rho(y_{i1}-\bar{y}_{\cdot 1}+y_{in}-\bar{y}_{\cdot n})\Big)}{\sigma^2(1+\rho)+\sigma_u^2(n-\rho(n-2))}.
\end{aligned}$$

A. 14 We have, that $\mathbf{V}_0=\sigma_u^2\mathbf{J}_{n\times n}+\mathbf{R}_0$, and we must find $\mathbf{V}_0^{-1}$:

$$\begin{aligned}
&\mathbf{V}_0=\sigma_u^2\mathbf{J}_{n\times n}+\mathbf{R}_0 \;\Leftrightarrow\; \mathbf{I}_{n\times n}=\sigma_u^2\mathbf{J}_{n\times n}\mathbf{V}_0^{-1}+\mathbf{R}_0\mathbf{V}_0^{-1} \;\Leftrightarrow\; \mathbf{R}_0\mathbf{V}_0^{-1}=\mathbf{I}_{n\times n}-\sigma_u^2\mathbf{J}_{n\times n}\mathbf{V}_0^{-1}\\
&\Leftrightarrow \mathbf{V}_0^{-1}=\mathbf{R}_0^{-1}-\sigma_u^2\mathbf{R}_0^{-1}\mathbf{J}_{n\times n}\mathbf{V}_0^{-1} \;=\; \mathbf{R}_0^{-1}-\sigma_u^2\mathbf{R}_0^{-1}\mathbf{J}_{n\times n}\left(\sigma_u^2\mathbf{J}_{n\times n}+\mathbf{R}_0\right)^{-1}\\
&\qquad=\mathbf{R}_0^{-1}-\mathbf{R}_0^{-1}\mathbf{J}_{n\times n}\left(\mathbf{J}_{n\times n}+\frac{1}{\sigma_u^2}\mathbf{R}_0\right)^{-1} \;=\; \mathbf{R}_0^{-1}-\mathbf{R}_0^{-1}\mathbf{J}_{n\times n}\mathbf{R}_0^{-1}\left(\mathbf{1}_n^T\mathbf{R}_0^{-1}\mathbf{1}_n+\frac{1}{\sigma_u^2}\right)^{-1}\\
&\qquad=\mathbf{R}_0^{-1}-\frac{1}{r_0+\frac{1}{\sigma_u^2}}\mathbf{R}_0^{-1}\mathbf{J}_{n\times n}\mathbf{R}_0^{-1} \;=\; \mathbf{R}_0^{-1}-\frac{1}{r_0+\frac{1}{\sigma_u^2}}\mathbf{R}_0^{-1}\mathbf{1}_n\mathbf{1}_n^T\mathbf{R}_0^{-1}, \qquad (8)
\end{aligned}$$

remembering that $r_0=\mathbf{1}_n^T\mathbf{R}_0^{-1}\mathbf{1}_n$. Thus for $\varphi=\sigma_u^2$, $LHS(\varphi)$ and $RHS(\varphi)$ becomes

$$\begin{aligned}
LHS(\sigma_u^2) &\overset{(4.25)}{=} m\mathrm{tr}\left\{\mathbf{V}_0^{-1}\frac{\partial\mathbf{V}_0}{\partial\sigma_u^2}\right\} \;=\; m\mathrm{tr}\left\{\mathbf{V}_0^{-1}\frac{\partial}{\partial\sigma_u^2}(\sigma_u^2\mathbf{J}_{n\times n}+\mathbf{R}_0)\right\} \;=\; m\mathrm{tr}\left\{\mathbf{V}_0^{-1}\mathbf{J}_{n\times n}\right\} \;=\; m\mathbf{1}_n^T\mathbf{V}_0^{-1}\mathbf{1}_n\\
&\overset{(8)}{=} m\mathbf{1}_n^T\left(\mathbf{R}_0^{-1}-\frac{1}{r_0+\frac{1}{\sigma_u^2}}\mathbf{R}_0^{-1}\mathbf{1}_n\mathbf{1}_n^T\mathbf{R}_0^{-1}\right)\mathbf{1}_n \;=\; m\left(\mathbf{1}_n^T\mathbf{R}_0^{-1}\mathbf{1}_n-\frac{1}{r_0+\frac{1}{\sigma_u^2}}\mathbf{1}_n^T\mathbf{R}_0^{-1}\mathbf{1}_n\mathbf{1}_n^T\mathbf{R}_0^{-1}\mathbf{1}_n\right)\\
&= m\left(r_0-\frac{1}{r_0+\frac{1}{\sigma_u^2}}r_0^2\right) \;=\; mr_0\left(1-\frac{r_0}{r_0+\frac{1}{\sigma_u^2}}\right) \;=\; mr_0\left(\frac{r_0+\frac{1}{\sigma_u^2}-r_0}{r_0+\frac{1}{\sigma_u^2}}\right) \;=\; m\frac{r_0}{r_0\sigma_u^2+1}
\end{aligned}$$

and

$$\begin{aligned}
RHS(\sigma_u^2) &\overset{(4.25)}{=} -(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})^T\frac{\partial\mathbf{V}^{-1}}{\partial\sigma_u^2}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}) = -(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})^T\frac{\partial}{\partial\sigma_u^2}\left(\mathbf{I}_{m\times m}\otimes\left(\mathbf{R}_0^{-1}-\frac{1}{r_0+\frac{1}{\sigma_u^2}}\mathbf{R}_0^{-1}\mathbf{1}_n\mathbf{1}_n^T\mathbf{R}_0^{-1}\right)\right)(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})\\
&= (\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})^T\frac{\partial}{\partial\sigma_u^2}\frac{1}{r_0+\frac{1}{\sigma_u^2}}\left(\mathbf{I}_{m\times m}\otimes\mathbf{R}_0^{-1}\mathbf{1}_n\mathbf{1}_n^T\mathbf{R}_0^{-1}\right)(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})\\
&= (\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})^T\frac{\partial}{\partial\sigma_u^2}\frac{\sigma_u^2}{r_0\sigma_u^2+1}\left(\mathbf{I}_{m\times m}\otimes\mathbf{R}_0^{-1}\mathbf{1}_n\mathbf{1}_n^T\mathbf{R}_0^{-1}\right)(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})\\
&= (\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})^T\frac{1}{(r_0\sigma_u^2+1)^2}\left(\mathbf{I}_{m\times m}\otimes\mathbf{R}_0^{-1}\mathbf{1}_n\mathbf{1}_n^T\mathbf{R}_0^{-1}\right)(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}).
\end{aligned}$$

A. 15 For $\varphi = \sigma^2$, assuming that $\mathbf{V}_0 = \sigma^2\mathbf{A}$ is correct, $LHS(\varphi)$ and $RHS(\varphi)$ becomes

$$\begin{aligned} LHS(\sigma^2) = m\mathrm{tr}\left\{\mathbf{V}_0^{-1}\frac{\partial \mathbf{V}_0}{\partial \sigma^2}\right\} &= m\mathrm{tr}\left\{\frac{1}{\sigma^2}\mathbf{A}^{-1}\frac{\partial}{\partial \sigma^2}\sigma^2\mathbf{A}\right\} \\ = m\mathrm{tr}\left\{\frac{1}{\sigma^2}\mathbf{A}^{-1}\mathbf{A}\right\} &= m\frac{1}{\sigma^2}\mathrm{tr}\{\mathbf{I}_{n\times n}\} = \frac{1}{\sigma^2}mn \end{aligned}$$

and with $\mathbf{V}^{-1} = \mathbf{I}_{m\times m}\otimes\frac{1}{\sigma^2}\mathbf{A}^{-1}$, we get

$$\begin{aligned} RHS(\sigma^2) &= -(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})^T\frac{\partial \mathbf{V}^{-1}}{\partial \sigma^2}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}) \\ &= \frac{1}{\sigma^4}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})^T\left(\mathbf{I}_{m\times m}\otimes\mathbf{A}^{-1}\right)(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}) \\ &= \frac{1}{\sigma^4}\sum_{i=1}^{m}(\mathbf{y}_i-\hat{\boldsymbol{\beta}})^T\mathbf{A}^{-1}(\mathbf{y}_i-\hat{\boldsymbol{\beta}}) \\ &= \frac{1}{\sigma^4}\sum_{i=1}^{m}\begin{bmatrix} y_{i1}-\bar{y}_{\cdot 1} \\ y_{i2}-\bar{y}_{\cdot 2} \\ \vdots \\ y_{in}-\bar{y}_{\cdot n}\end{bmatrix}^T\mathbf{A}^{-1}\begin{bmatrix} y_{i1}-\bar{y}_{\cdot 1} \\ y_{i2}-\bar{y}_{\cdot 2} \\ \vdots \\ y_{in}-\bar{y}_{\cdot n}\end{bmatrix} \\ &= \frac{1}{\sigma^4(1-\rho^2)}\sum_{i=1}^{m}\begin{bmatrix}\delta_{i1}\\ \delta_{i2}\\ \delta_{i3}\\ \vdots\\ \delta_{i,n-1}\\ \delta_{in}\end{bmatrix}^T\begin{bmatrix} 1 & -\rho & 0 & \dots & 0 & 0 \\ -\rho & 1+\rho^2 & -\rho & \dots & 0 & 0 \\ 0 & -\rho & 1+\rho^2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1+\rho^2 & -\rho \\ 0 & 0 & 0 & \dots & -\rho & 1\end{bmatrix}\begin{bmatrix}\delta_{i1}\\ \delta_{i2}\\ \delta_{i3}\\ \vdots\\ \delta_{i,n-1}\\ \delta_{in}\end{bmatrix} \\ &= \frac{1}{\sigma^4(1-\rho^2)}\sum_{i=1}^{m}\begin{bmatrix}\delta_{i1}\\ \delta_{i2}\\ \delta_{i3}\\ \vdots\\ \delta_{i,n-1}\\ \delta_{in}\end{bmatrix}^T\begin{bmatrix} \delta_{i1}-\rho\delta_{i2} \\ -\rho\delta_{i1}+(1+\rho^2)\delta_{i2}-\rho\delta_{i3} \\ -\rho\delta_{i2}+(1+\rho^2)\delta_{i3}-\rho\delta_{i4} \\ \vdots \\ -\rho\delta_{i,n-2}+(1+\rho^2)\delta_{i,n-1}-\rho\delta_{in} \\ -\rho\delta_{i,n-1}+\delta_{in}\end{bmatrix} \\ &= \frac{1}{\sigma^4(1-\rho^2)}\sum_{i=1}^{m}\Big(\delta_{i1}(\delta_{i1}-\rho\delta_{i2})+\delta_{i2}(-\rho\delta_{i1}+(1+\rho^2)\delta_{i2}-\rho\delta_{i3}) \\ &\quad +\delta_{i3}(-\rho\delta_{i2}+(1+\rho^2)\delta_{i3}-\rho\delta_{i4})+\ldots+\delta_{i,n-1}(-\rho\delta_{i,n-2}+(1+\rho^2)\delta_{i,n-1}-\rho\delta_{in}) \\ &\quad +\delta_{in}(-\rho\delta_{i,n-1}+\delta_{in})\Big) \\ &= \frac{1}{\sigma^4(1-\rho^2)}\sum_{i=1}^{m}\left(\delta_{i1}^2+\delta_{in}^2-2\rho\sum_{j=2}^{n}\delta_{i,j-1}\delta_{ij}+(1+\rho^2)\sum_{j=2}^{n-1}\delta_{ij}^2\right) \\ &= \frac{1}{\sigma^4(1-\rho^2)}\sum_{i=1}^{m}\left((1+\rho^2)\sum_{j=1}^{n}\delta_{ij}^2-\rho^2(\delta_{i1}^2+\delta_{in}^2)-2\rho\sum_{j=2}^{n}\delta_{i,j-1}\delta_{ij}\right). \end{aligned}$$

A. 16 For $\varphi = \rho$, $LHS(\varphi)$ and $RHS(\varphi)$ becomes

$$\begin{aligned} LHS(\rho) = m\text{tr}\left\{\mathbf{V}_0^{-1}\frac{\partial \mathbf{V}_0}{\partial \rho}\right\} &= m\text{tr}\left\{\frac{1}{\sigma^2}\mathbf{A}^{-1}\frac{\partial}{\partial \rho}\sigma^2\mathbf{A}\right\} \\ = m\text{tr}\left\{\mathbf{A}^{-1}\frac{\partial}{\partial \rho}\mathbf{A}\right\} &= -\frac{\rho}{1-\rho^2}2m(n-1). \end{aligned}$$

Here, I skip the calculation of $\mathbf{A}^{-1}\frac{\partial}{\partial \rho}\mathbf{A}$, as it is quite comprehensive, and relying on the result in [7]. And

$$\begin{aligned} RHS(\rho) &= -(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})^T\frac{\partial \mathbf{V}^{-1}}{\partial \rho}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}) \\ &= -(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})^T\left(\frac{1}{\sigma^2}\mathbf{I}_{m\times m}\otimes\frac{\partial}{\partial \rho}\mathbf{A}^{-1}\right)(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}) \\ &= \frac{1}{\sigma^2}\sum_{i=1}^{m}\left(\begin{bmatrix}\delta_{i1}\\ \delta_{i2}\\ \delta_{i3}\\ \vdots\\ \delta_{i,n-1}\\ \delta_{in}\end{bmatrix}^T\frac{\partial}{\partial \rho}\mathbf{A}^{-1}\begin{bmatrix}\delta_{i1}\\ \delta_{i2}\\ \delta_{i3}\\ \vdots\\ \delta_{i,n-1}\\ \delta_{in}\end{bmatrix}\right) \\ &= \frac{1}{\sigma^2(1-\rho^2)^2} \\ &\quad \sum_{i=1}^{m}\left(\begin{bmatrix}\delta_{i1}\\ \delta_{i2}\\ \delta_{i3}\\ \vdots\\ \delta_{i,n-1}\\ \delta_{in}\end{bmatrix}^T\begin{bmatrix}2\rho & -(1+\rho^2) & 0 & \dots & 0 & 0\\ -(1+\rho^2) & 4\rho & -(1+\rho^2) & \dots & 0 & 0\\ 0 & -(1+\rho^2) & 4\rho & \dots & 0 & 0\\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots\\ 0 & 0 & 0 & \dots & 4\rho & -(1+\rho^2)\\ 0 & 0 & 0 & \dots & -(1+\rho^2) & 2\rho\end{bmatrix}\begin{bmatrix}\delta_{i1}\\ \delta_{i2}\\ \delta_{i3}\\ \vdots\\ \delta_{i,n-1}\\ \delta_{in}\end{bmatrix}\right) \\ &= -\frac{1}{\sigma^2(1-\rho^2)^2}\sum_{i=1}^{m}\left(4\rho\sum_{j=1}^{n}\delta_{ij}^2 - 2\rho(\delta_{i1}^2+\delta_{in}^2) - 2(1+\rho^2)\sum_{j=2}^{n}\delta_{i,j-1}\delta_{ij}\right). \end{aligned}$$

The last equation comes from the expression of $RHS(\sigma^2)$. The calculations in $RHS(\rho)$ are the same as in $RHS(\sigma^2)$, only $-\frac{1}{\sigma^4}\mathbf{A}^{-1}$ is replaced by $\frac{1}{\sigma^2}\frac{\partial}{\partial \rho}\mathbf{A}^{-1}$.

A. 17 The estimate of μ_i is a weighted average between the overall mean, μ, and the group average observation, $\bar{y}_{i\cdot}$. We have, that

$$\begin{aligned} \text{Var}\left[\bar{Y}_{i\cdot}\right] = \text{Var}\left[\frac{1}{n}\sum_{j=1}^{n}Y_{ij}\right] &= \text{Var}\left[\frac{1}{n}\sum_{j=1}^{n}\mu_i\right] + \text{Var}\left[\frac{1}{n}\sum_{j=1}^{n}\epsilon_{ij}\right] = \text{Var}\left[\frac{1}{n}n\mu_i\right] + \frac{1}{n^2}\sum_{j=1}^{n}\text{Var}\left[\epsilon_{ij}\right] \\ &= \text{Var}\left[\mu_i\right] + \frac{1}{n^2}n\text{Var}\left[\epsilon_{ij}\right] = \sigma_u^2 + \frac{\sigma^2}{n}, \end{aligned}$$

and

$$\begin{aligned}\mathrm{E}\left[\bar{Y}_{i\cdot}\right] &= \mathrm{E}\left[\frac{1}{n}\sum_{j=1}^{n} Y_{ij}\right] \;=\; \mathrm{E}\left[\frac{1}{n}\sum_{j=1}^{n}\mu_i\right] + \mathrm{E}\left[\frac{1}{n}\sum_{j=1}^{n}\epsilon_{ij}\right] \\ &= \mathrm{E}\left[\frac{1}{n}n\mu_i\right] + \frac{1}{n}\sum_{j=1}^{n}\mathrm{E}\left[\epsilon_{ij}\right] \;=\; \mathrm{E}\left[\mu_i\right] \;=\; \mu.\end{aligned}$$

This means that the best prediction of the group average mean is found by conditioning on what we already know, i.e. $\bar{y}_{i\cdot}$:

$$\begin{aligned}\hat{\mu}_i = \mathrm{E}\left[\mu_i \mid \bar{Y}_{i\cdot}\right] &= \mathrm{E}\left[\mu_i\right] + \mathrm{Cov}\left[\mu_i, \bar{Y}_{i\cdot}\right]\frac{1}{\mathrm{Var}\left[\bar{Y}_{i\cdot}\right]}\left(\bar{y}_{i\cdot} - \mathrm{E}\left[\bar{Y}_{i\cdot}\right]\right) \\ &= \mu + \left(\mathrm{E}\left[\mu_i\bar{Y}_{i\cdot}\right] - \mathrm{E}\left[\mu_i\right]\mathrm{E}\left[\bar{Y}_{i\cdot}\right]\right)\frac{1}{\sigma_u^2 + \frac{\sigma^2}{n}}\left(\bar{y}_{i\cdot} - \mu\right) \\ &= \mu + \left(\mathrm{E}\left[\mu_i\frac{1}{n}\sum_{j=1}^{n}(\mu_i + \epsilon_{ij})\right] - \mu^2\right)\frac{1}{\sigma_u^2 + \frac{\sigma^2}{n}}\left(\bar{y}_{i\cdot} - \mu\right) \\ &= \mu + \left(\mathrm{E}\left[\mu_i^2\right] - \mathrm{E}\left[\frac{1}{n}\sum_{j=1}^{n}\mu_i\epsilon_{ij}\right] - \mu^2\right)\frac{1}{\sigma_u^2 + \frac{\sigma^2}{n}}\left(\bar{y}_{i\cdot} - \mu\right) \\ &= \mu + \left(\sigma_u^2 + \mu^2 - \frac{1}{n}\sum_{j=1}^{n}\mathrm{E}\left[\mu_i\right]\mathrm{E}\left[\epsilon_{ij}\right] - \mu^2\right)\frac{1}{\sigma_u^2 + \frac{\sigma^2}{n}}\left(\bar{y}_{i\cdot} - \mu\right) \\ &= \mu + \frac{\sigma_u^2}{\sigma_u^2 + \frac{\sigma^2}{n}}\left(\bar{y}_{i\cdot} - \mu\right) \;=\; \left(1 - \frac{\sigma_u^2}{\sigma_u^2 + \frac{\sigma^2}{n}}\right)\mu + \frac{\sigma_u^2}{\sigma_u^2 + \frac{\sigma^2}{n}}\bar{y}_{i\cdot}.\end{aligned}$$

A. 18 We have, that $U_i \sim N\left(0, \sigma_u^2\right)$. The covariance between the total residuals at two timepoint is just the variance of the random intercepts:

$$\begin{aligned}\mathrm{Cov}\left[Y_{ij}, Y_{ik}\right] &= \mathrm{Cov}\left[U_i + \epsilon_{ij}, U_i + \epsilon_{ik}\right] \\ &= \mathrm{E}\left[(U_i + \epsilon_{ij})(U_i + \epsilon_{ik})\right] - \mathrm{E}\left[U_i + \epsilon_{ij}\right]\mathrm{E}\left[U_i + \epsilon_{ik}\right] \\ &= \mathrm{E}\left[U_i^2\right] + \mathrm{E}\left[\epsilon_{ij}\epsilon_{ik}\right] + \mathrm{E}\left[U_i\epsilon_{ik}\right] + \mathrm{E}\left[U_i\epsilon_{ik}\right] - 0 \\ &= \sigma_u^2.\end{aligned}$$

A. 19 The covariance of the total residual is

$$\begin{aligned}
\operatorname{Cov}\left[Y_{ij}, Y_{ik}\right] &= \operatorname{Cov}\left[U_i + x_{ij}b_i + \epsilon_{ij}, U_i + x_{ik}b_i + \epsilon_{ik}\right] \\
&= \mathrm{E}\left[(U_i + x_{ij}b_i + \epsilon_{ij})(U_i + x_{ik}b_i + \epsilon_{ik})\right] - \mathrm{E}\left[U_i + x_{ij}b_i + \epsilon_{ij}\right]\mathrm{E}\left[U_i + x_{ik}b_i + \epsilon_{ik}\right] \\
&= \mathrm{E}\left[U_i^2\right] + x_{ij}x_{ik}\mathrm{E}\left[b_i^2\right] + \mathrm{E}\left[\epsilon_{ij}\epsilon_{ik}\right] + (x_{ij} + x_{ik})\mathrm{E}\left[U_i b_i\right] + \mathrm{E}\left[U_i(\epsilon_{ij} + \epsilon_{ik})\right] \\
&\quad + x_{ij}\mathrm{E}\left[b_i\epsilon_{ik}\right] + x_{ik}\mathrm{E}\left[b_i\epsilon_{ij}\right] - 0 \\
&= \sigma_u^2 + x_{ij}x_{ik}\sigma_b^2 + (x_{ij} + x_{ik})\mathrm{E}\left[U_i b_i\right] \\
&= \sigma_u^2 + x_{ij}x_{ik}\sigma_b^2 + (x_{ij} + x_{ik})\Big(\operatorname{Cov}\left[U_i, b_i\right] + \mathrm{E}\left[U_i\right]\mathrm{E}\left[b_i\right]\Big) \\
&= \sigma_u^2 + x_{ij}x_{ik}\sigma_b^2 + (x_{ij} + x_{ik})\sigma_{ub}.
\end{aligned}$$

A. 20 When a random variable, Y_i, has a distribution from the exponential family, the mean value of Y_i can be expressed as $b'(\theta_i)$. To show this, we use the fact that the mean of a random variable, X, is defined as $\mathrm{E}[X] = \int x f(x)\, dx$:

$$\begin{aligned}
\frac{d}{d\theta_i}\int f(y_i;\theta_i,\phi)\, dy_i &= \int \frac{d}{d\theta_i} f(y_i;\theta_i,\phi,)\, dy_i \\
&\overset{(6.1)}{=} \int h\left(y_i,\phi\right)\frac{d}{d\theta_i}\exp\left(\frac{y_i\theta_i - b(\theta_i)}{\phi}\right) dy_i \\
&= \int h\left(y_i,\phi\right)\exp\left(\frac{y_i\theta_i - b(\theta_i)}{\phi}\right)\left(\frac{y_i - b'(\theta_i)}{\phi}\right) dy_i \\
&= \frac{1}{\phi}\int \left(y_i f(y_i;\theta_i,\phi) - b'(\theta_i) f(y_i;\theta_i,\phi)\right)\, dy_i \qquad (9) \\
&= \frac{1}{\phi}\left(\int y_i f(y_i;\theta_i,\phi)\, dy_i - b'(\theta_i)\int f(y_i;\theta_i,\phi)\, dy_i\right) \qquad (10) \\
&= \frac{1}{\phi}\left(\mathrm{E}\left[Y_i\right] - b'(\theta_i)\right) \\
&= 0 \\
&\Updownarrow \\
\mathrm{E}\left[Y_i\right] &= b'(\theta_i).
\end{aligned}$$

The variance of Y_i can be expressed as $\phi b''(\theta_i)$. To show this, we use the fact that the variance of a

random variable, X, is defined as $\mathrm{Var}[X] = \mathrm{E}\left[(X - \mathrm{E}[X])^2\right] = \mathrm{E}\left[X^2\right] - \mathrm{E}[X]^2$:

$$\begin{aligned}
\frac{d^2}{d\theta_i^2}\int f(y_i;\theta_i,\phi)\,dy_i &= \int \frac{d^2}{d\theta_i^2} f(y_i;\theta_i,\phi)\,dy_i \\
&\overset{(10)}{=} \frac{1}{\phi}\left(\int y_i \frac{d}{d\theta_i} f(y_i;\theta_i,\phi)\,dy_i - \int \frac{d}{d\theta_i} b'(\theta_i) f(y_i;\theta_i,\phi)\,dy_i\right) \\
&\overset{(9)}{=} \frac{1}{\phi}\left[\frac{1}{\phi_i}\int y_i\left(y_i f(y_i;\theta_i,\phi) - b'(\theta_i) f(y_i;\theta_i,\phi)\right)\,dy_i\right. \\
&\quad \left. - \int \left(b''(\theta_i) f(y_i;\theta_i,\phi) + b'(\theta_i)\frac{d}{d\theta_i} f(y_i;\theta_i,\phi)\right)\,dy_i\right] \\
&= \frac{1}{\phi^2}\left(\mathrm{E}\left[Y_i^2\right] - b'(\theta_i)\right) - \frac{1}{\phi} b''(\theta_i)\int f(y_i;\theta_i,\phi)\,dy_i \\
&\quad - \frac{1}{\phi} b'(\theta)\int \frac{d}{d\theta_i} f(y_i;\theta_i,\phi)\,dy_i \\
&= \frac{1}{\phi^2}\left(\mathrm{E}\left[Y_i^2\right] - b'(\theta_i)\right) - \frac{1}{\phi} b''(\theta_i) - \frac{1}{\phi} b'(\theta_i)\left(\frac{1}{\phi}\left(\mathrm{E}[Y_i] - b'(\theta_i)\right)\right) \\
&= \frac{1}{\phi^2}\left(\mathrm{E}\left[Y_i^2\right] - \mathrm{E}[Y_i]\right) - \frac{1}{\phi} b''(\theta_i) - \frac{1}{\phi} b'(\theta_i)\left(\frac{1}{\phi}\left(\mathrm{E}[Y_i] - \mathrm{E}[Y_i]\right)\right) \\
&= \frac{1}{\phi^2}\left(\mathrm{Var}[Y_i]\right) - \frac{1}{\phi} b''(\theta_i) \\
&= 0 \\
&\Updownarrow \\
\mathrm{Var}[Y_i] &= \phi b''(\theta_i).
\end{aligned}$$

A. 21 Remember, that $\mu_i = b'(\theta_i)$, $\theta_i = b'(\mu_i)^{-1}$ and $V(\mu_i) = b''\left(b'(\mu_i)^{-1}\right)$. Thus the score function is

$$\begin{aligned}
\frac{\partial \ell_i}{\partial \mu_i} = \frac{\partial \ell_i}{\partial \theta_i} \times \frac{\partial \theta_i}{\partial \mu_i} &\overset{(6.1)}{=} \frac{y_i - b'(\theta_i)}{\phi} \times \left(\frac{\partial \mu_i}{\partial \theta_i}\right)^{-1} \\
&= \frac{y_i - b'(\theta_i)}{\phi} \times \left(\frac{\partial b'(\theta_i)}{\partial \theta_i}\right)^{-1} \\
&= \frac{y_i - b'(\theta_i)}{\phi} \times \left(b''(\theta_i)\right)^{-1} \\
&= \frac{y_i - b'(\theta_i)}{\phi} \times \frac{1}{b''(b'(\mu_i)^{-1})} \\
&= \frac{y_i - b'(\theta_i)}{\phi} \times \frac{1}{V(\mu_i)}.
\end{aligned}$$

A. 22 The function q_i satisfies the properties $\mathrm{E}\left[q_i\right] = 0$ and $\mathrm{Var}\left[q_i\right] = -\mathrm{E}\left[\frac{\partial q_i}{\partial \mu_i}\right]$, meaning it mimics a proper score function:

$$\begin{aligned}
\mathrm{E}\left[q_i\right] &= \frac{\mathrm{E}\left[Y_i\right] - \mu_i}{\phi V(\mu_i)} = 0 \\
\mathrm{Var}\left[q_i\right] &= \frac{\mathrm{Var}\left[Y_i - \mu_i\right]}{\phi^2 V(\mu_i)^2} = \frac{\phi V(\mu_i)}{\phi^2 V(\mu_i)^2} = \frac{1}{\phi V(\mu_i)} \\
-\mathrm{E}\left[\frac{\partial q_i}{\partial \mu_i}\right] &= -\mathrm{E}\left[\frac{Y_i}{\phi}\frac{\partial}{\partial \mu_i}\frac{1}{V(\mu_i)} - \frac{1}{\phi}\frac{\partial}{\partial \mu_i}\frac{\mu_i}{V(\mu_i)}\right] = -\mathrm{E}\left[-\frac{Y_i}{\phi}\frac{V'(\mu_i)}{V(\mu_i)^2} - \frac{1}{\phi}\frac{V(\mu_i) - \mu_i V'(\mu_i)}{V(\mu_i)^2}\right] \\
&= -\mathrm{E}\left[\frac{-Y_i V'(\mu_i) - V(\mu_i) + \mu_i V'(\mu_i)}{\phi V(\mu_i)^2}\right] = -\left(\frac{-\mathrm{E}\left[Y_i\right] V'(\mu_i) - V(\mu_i) + \mu_i V'(\mu_i)}{\phi V(\mu_i)^2}\right) \\
&= \frac{V(\mu_i)}{\phi V(\mu_i)^2} = \frac{1}{\phi V(\mu_i)}.
\end{aligned}$$

A. 23 The mean value of r_{jk} is just the correlation ρ_{jk}:

$$\begin{aligned}
\mathrm{E}\left[r_{jk}\right] &= \frac{\mathrm{E}\left[(Y_{ij} - \mu_{ij})(Y_{ik} - \mu_{ik})\right]}{\phi\sqrt{V(\mu_{ij})V(\mu_{ik})}} = \frac{\mathrm{Cov}\left[Y_{ij}, Y_{ik}\right]}{\sqrt{\phi^2 V(\mu_{ij})V(\mu_{ik})}} = \frac{\mathrm{Cov}\left[Y_{ij}, Y_{ik}\right]}{\sqrt{\mathrm{Var}\left[Y_{ij}\right]\mathrm{Var}\left[Y_{ik}\right]}} \\
&= \mathrm{Corr}\left[Y_{ij}, Y_{ik}\right] = \rho_{jk}.
\end{aligned}$$

Appendix B

This appendix contains additional theory that may be helpful for the reader.

B. 1 (The `anova`-function in `R`) When giving `anova()` two inputs, the function performs a comparison of the models, which must be nested. The comparison is based on AIC (*Akaike's Information Criterion*). AIC is given as

$$AIC = 2p - 2\log(\hat{L}),$$

where p is the number of parameters in the model and $\hat{L}$ is the maximum value of the likelihood function for the model. When given an increasing input, $\log(\cdot)$ is an increasing function. For a fixed p, AIC will decrease as $\log(L)$ increases. And AIC will be at its minimum when $\log(L)$ is at its maximum, i.e. when $L = \hat{L}$. As the objective is always, that we want L maximized, the objective of the AIC is to have it minimized. When comparing two models with, say, AIC_{model1} and AIC_{model2}, the better model will be the one with the smallest AIC. But how big does the difference in AIC have to be for us to say that $AIC_{\text{model1}} \neq AIC_{\text{model2}}$? For this, `anova()` also provides a p-values. The hypothesis tested in `anova()` is

$$H_0^{\text{anova}_1} : \textit{The AIC of the models are the same.}$$

For a p-value below the level α, $H_0^{\text{anova}_1}$ is rejected.

When giving `anova()` just one input, the function tests whether the model terms are significant, i.e. whether the estimated coefficients are significant. The hypothesis is

$$H_0^{\text{anova}_2} : \beta_1, \ldots, \beta_p \textit{ are significant.}$$

The test statistic is

$$F = \frac{N-k}{k-1} \cdot \frac{\sum_{i=1}^{k} n_i (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i\cdot})^2},$$

where k is the number of samples, n_i is the size of sample i, $N = \sum_{i=1}^{k} n_i$, y_{ij} is the observation of the jth subject in the ith sample, $\bar{y}_{i\cdot}$ is the sample mean for the ith sample, and $\bar{y}_{\cdot\cdot}$ is the overall mean. The test statistic is approximately F-distributed with $k-1$ and $N-k$ degrees of freedom, and $H_0^{\text{anova}_2}$ is rejected if $F > F(\alpha, k-1, N-k)$ for a chosen level α.

Based on `https://en.wikipedia.org/wiki/F-test`.

B. 2 (*F*-test) An F-test is not one particular test, but rather any statistical test in which the test statistic follows an F-distribution under H_0. Examples of F-tests are given in B.4, B.6 and B.7. The F-test is sensitive to non-normality. When doing analysis of variance, the assumption of homoscedasticity is more easily violated, when the normality assumption is violated.

Based on `https://en.wikipedia.org/wiki/Mauchly%27s_sphericity_test`.

B. 3 (Eigenvalues) An eigenvalue is a scalar, that satisfies $\mathbf{Av} = \lambda\mathbf{v}$ for a matrix $\mathbf{A}$, where $\mathbf{v}$ is called the corresponding eigenvector.

Let λ_i for $i = 1, \ldots, n$ be the eigenvalues (including multiplicity) of $\mathbf{A} \in \mathbb{R}^{n\times n}$. Then

$$\operatorname{tr}\{\mathbf{A}\} = \sum_{i=1}^{n} \lambda_i \quad \text{and} \quad |\mathbf{A}| = \prod_{i=1}^{n} \lambda_i.$$

Eigenvalues can also be thought of as roots as they are the solutions to the characteristic polynomial $|\mathbf{A} - \lambda\mathbf{I}_{n\times n}|$.

Based on https://en.wikipedia.org/wiki/Determinant#Relation_to_eigenvalues_and_trace.

B. 4 (Shapiro-Wilk test) The Shapiro-Wilk test tests the hypothesis

$$H_0^{\text{shapiro}} : \textit{sample is from a normally distributed population.}$$

The test statistic is

$$W = \frac{\left(\sum_{i=1}^{n} a_i y_{(i)}\right)^2}{\sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2},$$

where data is arranged in ascending order with $y_{(i)}$ being the ith smallest observation in the sample, $\bar{y}$ is the sample mean and

$$\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} = \frac{m^T V^{-1}}{\left(m^T V^{-1} V^{-1} m\right)^{\frac{1}{2}}}$$

with $m = \mathrm{E}[Y_i] \overset{\text{iid.}}{\sim} N(0,1)$ and V being the covariance matrix of the $y_{(i)}$s. A p-value lower than the chosen level α means H_0 is rejected. The p-value is found in a table using the sample size, n, and the calculated test statistic, W.

Based on https://en.m.wikipedia.org/wiki/Shapiro–Wilk_test.

B. 5 (Kruskal-Wallis) When the assumption of Gaussian samples fail, one can use the Kruskal-Wallis test as an alternative to the one-way ANOVA. This test makes no assumptions about the distribution of the samples. It tests the hypothesis

$$H_0^{\text{kruskal}} : \textit{the medians of all the groups are equal.}$$

The test works by first arranging all the values in all the samples by rank. Let r_{ij} denote the rank of the observation y_{ij}. The test statistic is

$$K = (N-1)\frac{\sum_{i=1}^{g} n_i \left(\bar{r}_{i\cdot} - \bar{r}\right)^2}{\sum_{i=1}^{k}\sum_{j=1}^{n_i} \left(\bar{r}_{ij} - \bar{r}\right)^2},$$

where $\bar{r}_{i\cdot} = \sum_{j=1}^{n_i} r_{ij}$ is the average rank for observations in group i, $\bar{r} = \frac{N+1}{2}$ is the average of all the ranks, N is the total number of observation and n_i is the number of observations in group i. The

test statistic is approximately χ^2-distributed with $g-1$ degrees of freedom, and H_0^{kruskal} is rejected if $K > \chi^2_{g-1,\alpha}$ for a chosen level α.

Based on https://en.wikipedia.org/wiki/Kruskal–Wallis_one-way_analysis_of_variance.

B. 6 (Bartlett test) The Bartlett test tests the hypothesis

$$H_0^{\text{bartlett}} : \textit{variances across samples are equal.}$$

The test statistic is

$$\chi^2 = \frac{(N-k)\log\left(s_p^2\right) - \sum_{i=1}^{k}(n_i-1)\log\left(s_p^2\right)}{1 + \frac{1}{3(k-1)}\left(\sum_{i=1}^{k}\frac{1}{n_i-1} - \frac{1}{N-k}\right)},$$

where k is the number of samples, n_i is the size of sample i, $N = \sum_{i=1}^{k} n_i$, and $s_p^2 = \frac{1}{N-k}\sum_{i=1}^{k} s_i^2$ with s_i^2 being the ith sample variance. The test statistic is approximately χ^2-distributed with $k-1$ degrees of freedom, and H_0 is rejected if $\chi^2 > \chi^2_{k-1,\alpha}$ for a chosen level α.

Based on https://en.wikipedia.org/wiki/Bartlett%27s_test.

B. 7 (Levene's test) The Levene's test tests the same hypothesis as the Bartlett test, but does so using the F-distribution. The test statistic is

$$F = \frac{N-k}{k-1} \cdot \frac{\sum_{i=1}^{k} n_i (z_{i\cdot} - z_{\cdot\cdot})^2}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(z_{ij} - z_{i\cdot})^2},$$

where k is the number of samples, n_i is the size of sample i, $N = \sum_{i=1}^{k} n_i$, $z_{ij} = |y_{ij} - \bar{y}_{i\cdot}|$ with y_{ij} being the observation of the jth subject in the ith sample and $\bar{y}_{i\cdot}$ is the sample mean for the ith sample, $z_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i} z_{ij}$ and $z_{\cdot\cdot} = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i} z_{ij}$. The test statistic is approximately F-distributed with $k-1$ and $N-k$ degrees of freedom, and H_0^{levene} is rejected if $F > F(\alpha, k-1, N-k)$ for a chosen level α.

Based on https://en.m.wikipedia.org/wiki/Levene%27s_test.

B. 8 (Conditional multivariate distribution) For

$$\begin{pmatrix}\mathbf{X}_1\\ \mathbf{X}_2\end{pmatrix} \sim N_{n+m}\left(\begin{bmatrix}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_1 & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_2\end{bmatrix}\right)$$

we have that $\mathbf{X}_1 \mid \mathbf{X}_2 = \mathbf{x}_2 \sim N_n\left(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_2^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_{12}\right)$.

B. 9 (Direct product) Below are all the rules for direct product used in this report.

Direct product: For $\mathbf{A} \in \mathbb{R}^{n\times m}$ and $\mathbf{B} \in \mathbb{R}^{s\times t}$, the direct product is

$$\mathbf{A}\otimes\mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \dots & a_{1m}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & \dots & a_{2m}\mathbf{B} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1}\mathbf{B} & a_{n2}\mathbf{B} & \dots & a_{1nm}\mathbf{B} \end{bmatrix} \in \mathbb{R}^{ns\times mt}.$$

Direct sum: For matrices $\mathbf{A}_j \in \mathbb{R}^{n_j\times m_j}$, the direct sum is

$$\bigotimes_{j=1}^{J}\mathbf{A}_j = \mathbf{A}_1\otimes\mathbf{A}_2\otimes\cdots\otimes\mathbf{A}_J.$$

Sum: The sum of two direct sums for matrices $\mathbf{A}_j \in \mathbb{R}^{n_j\times m_j}$ and $\mathbf{B}_j \in \mathbb{R}^{n_j\times m_j}$ is

$$\left(\bigotimes_{j=1}^{J}\mathbf{A}_j\right) + \left(\bigotimes_{j=1}^{J}\mathbf{B}_j\right) = \bigotimes_{j=1}^{J}(\mathbf{A}_j+\mathbf{B}_j).$$

Product: The product of two direct sums for matrices $\mathbf{A}_j \in \mathbb{R}^{n_j\times m_j}$ and $\mathbf{B}_j \in \mathbb{R}^{m_j\times k_j}$ is

$$\left(\bigotimes_{j=1}^{J}\mathbf{A}_j\right)\left(\bigotimes_{j=1}^{J}\mathbf{B}_j\right) = \bigotimes_{j=1}^{J}(\mathbf{A}_j\mathbf{B}_j).$$

Transpose: The transpose of the direct product of matrices $\mathbf{A} \in \mathbb{R}^{n\times m}$, $\mathbf{B} \in \mathbb{R}^{k\times l}$, and $\mathbf{C} \in \mathbb{R}^{s\times t}$ is

$$(\mathbf{A}\otimes\mathbf{B}\otimes\mathbf{C})^T = \mathbf{A}^T\otimes\mathbf{B}^T\otimes\mathbf{C}^T$$

Inverse: The inverse of the direct product of matrices $\mathbf{A} \in \mathbb{R}^{n\times m}$ and $\mathbf{B} \in \mathbb{R}^{s\times t}$ is

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

Commutative law: For a constant c and a matrix $\mathbf{A} \in \mathbb{R}^{n\times m}$, we have

$$c\otimes\mathbf{A} = \mathbf{A}\otimes c = c\mathbf{A}.$$

Distributive law: For matrices $\mathbf{A} \in \mathbb{R}^{n\times m}$, $\mathbf{B} \in \mathbb{R}^{n\times m}$, and $\mathbf{C} \in \mathbb{R}^{s\times t}$, we have

$$(\mathbf{A}+\mathbf{B})\otimes\mathbf{C} = (\mathbf{A}\otimes\mathbf{C}) + (\mathbf{B}\otimes\mathbf{C})$$
$$\mathbf{C}\otimes(\mathbf{A}+\mathbf{B}) = (\mathbf{C}\otimes\mathbf{A}) + (\mathbf{C}\otimes\mathbf{B}).$$

Based on `https://en.wikipedia.org/wiki/Kronecker_product`.

B. 10 (Derivative of determinant) For a square matrix $\mathbf{A}$ dependent on t, we have

$$\frac{d}{dt}|\mathbf{A}(t)| = |\mathbf{A}(t)|\operatorname{tr}\left\{\mathbf{A}(t)^{-1}\frac{d}{dt}\mathbf{A}(t)\right\}.$$

Based on `https://en.wikipedia.org/wiki/Jacobi%27s_formula`.

B. 11 (Likelihood ratio test) Let $\hat{\boldsymbol{\theta}}$ be the maximum likelihood estimate of the parameter $\boldsymbol{\theta} \in \mathbb{R}^{p\times 1}$. We want to test the hypothesis

$$H_0^{\text{LRT}} : \boldsymbol{\theta} = \boldsymbol{\theta}_0,$$

where $\boldsymbol{\theta}_0$ is a proposed value of $\boldsymbol{\theta}$. To test H_0^{LRT}, we look at the ratio between the likelihood with the proposed value, $L(\boldsymbol{\theta}_0)$, and the likelihood with the estimated value, $L(\hat{\boldsymbol{\theta}})$, i.e. we look at $L(\boldsymbol{\theta}_0)/L(\hat{\boldsymbol{\theta}})$. If this ratio is very small, then the data with $\hat{\boldsymbol{\theta}}$ is more plausible than the data with $\boldsymbol{\theta}_0$, which means H_0^{LRT} is rejected. More precisely, the likelihood ratio test statistic is given by

$$-2LR = -2\left(\ell(\boldsymbol{\theta}_0) - \ell(\hat{\boldsymbol{\theta}})\right)$$

and H_0^{LRT} is rejected if $-2LR > \chi^2_{p,\alpha}$ for a chosen level α.

Based on `https://en.m.wikipedia.org/wiki/Likelihood-ratio_test`.

B. 12 (Wald test) Assume that $\hat{\boldsymbol{\theta}}$ is a consistent estimate of the parameter $\boldsymbol{\theta} \in \mathbb{R}^{p\times 1}$. Then

$$\hat{\boldsymbol{\theta}} \sim N_p\left(\boldsymbol{\theta}, I(\boldsymbol{\theta})^{-1}\right),$$

assymptotically, where $I(\boldsymbol{\theta})$ is the information matrix. The standard error of $\hat{\theta}_i$ is $\hat{\sigma}_i = \sqrt{\text{Var}\left[\hat{\boldsymbol{\theta}}\right]_{ii}}$. The Wald statistic is

$$W_i = \frac{\hat{\theta}_i - \theta_{i,0}}{\hat{\sigma}_i},$$

where $\theta_{i,0}$ is the proposed value of θ_i. Under

$$H_0^{\text{wald}} : \theta_i = \theta_{i,0}$$

W_i will be approximately $N(0,1)$-distributed. It is not uncommon to write the statistic as

$$W_i^2 = \frac{\left(\hat{\theta}_i - \theta_{i,0}\right)^2}{\hat{\sigma}_i^2},$$

which under H_0^{wald} is χ^2-distributed with 1 degree of freedom. We reject H_0^{wald} if $W_i > \chi^2_{1,\alpha}$ for a chosen level α.

Based on `https://en.m.wikipedia.org/wiki/Wald_test`.

References

[1] Author(s): P. Olofsson & M. Andersson
Title: *Probability, Statistics, and Stochastic Processes*, 2nd Ed., chapter 7
Publisher: Wiley
Year: 2012

[2] Author(s): D. Hedeker & R. D. Gibbons
Title: *Longitudinal Data Analysis*, chapters 2, 4, 6 and 8
Publisher: Wiley
Year: 2006

[3] Author(s): A. Field, University of Sussex,
Title: *A Bluffer's Guide to ... Sphericity*,
Year: 2017

[4] Author(s): The Pennsylvania State University,
www: `https://onlinecourses.science.psu.edu/stat505/node/159`, chapters 8.2 and 8.3
Year: 2018

[5] Author(s): G. Carey, University of Colorado, Boulder,
www: `http://ibgwww.colorado.edu/~carey/p7291dir/handouts/manova1.pdf`
Year: 1998

[6] Author(s): E. Demidenko
Title: *Mixed Models: Theory and Applications*, chapters 1 and 2
Publisher: Wiley
Year: 2004

[7] Author(s): C. E. McCullogh & S. R. Searle
Title: *Generalized, Linear, and Mixed Models*, chapters 5 and 7
Publisher: Wiley
Year: 2001

[8] Author(s): H. Madsen & P. Thyregod
Title: *Introduction to General and Generalized Linear Models*, chapter 5
Publisher: CRC Press
Year: 2011

[9] Author(s): C. Czado (Technische Universität München)
www: `http://www2.stat.duke.edu/~sayan/Sta613/2017/lec/LMM.pdf`
Year: 2017

[10] Author(s): B. S. Everitt & T. Hothorn
Title: *A Handbook of Statistical Analysis Using R*, chapter 12
Publisher: CRC Press
Year: 2010

[11] Author(s): U. Halekoh & S. Højsgaard
Title: *A Kenward-Roger Approximation and Parametric Bootstrap Methods for Tests in Linear Mixed Models – the R Package pbkrtest,*
Publisher: American Statistical Association
Year: 2014

[12] Author(s): M. A. Islam & R. I. Chowdhury
Title: *Analysis of Repeated Measures Data,* chapters 11 and 12
Publisher: Springer
Year: 2017

[13] Author(s): G. Fitzmaurice, M. Davidian, G. Verbeke & G. Molenberghs
Title: *Longitudinal Data Analysis,* chapter 3
Publisher: CRC Press
Year: 2009