Using the theory of Longitudinal Data to model the use of antipsychotics in elderly patients with dementia

Project Report

Sally Kit Ipsen

Aalborg University
Mathematics

This document is created using LaTeX for the typesetting.

Mathematics
Aalborg University
http://www.aau.dk

AALBORG UNIVERSITY
STUDENT REPORT

Title:
Using the theory of Longitudinal Data to model the use of antipsychotics in elderly patients with dementia

Theme:
Analysis of Longitudinal Data

Project Period:
September 2014 - April 2015

Project Group:
G5-105b

Participant(s):
Sally Kit Ipsen

Supervisor(s):
Poul Svante Eriksen

Copies: 4

Page Numbers: ??

Date of Completion:
April 10, 2015

Abstract:

In this project the theory of longitudinal data is investigated in order to be able to perform a thorough analysis of the use of antipsychotics in elderly patients with dementia.
We start with a presentation of why such an analysis seems necessary, followed by a description of the data and registers available to fulfill the purpose.
The chapters regarding the theory of longitudinal theory begins with an introduction to the basic theory, continues with description of several estimations methods, and ends with random effects models specified for both binary and count data.
The report is completed with a discussion leading up to conduct an analysis of the use of antipsychotic in elderly patients with dementia using the theory of longitudinal data.

Contents

Preface

This project is written by Sally Kit Ipsen in the fall of 2014 and the first quarter of 2015 during the 9th semester of Mathematics at Institute for Mathematical Sciences at Aalborg University in collaboration with the National Center for Registerbased Research at Aarhus University and supervised by Poul Svante Eriksen.

It is assumed that the reader possesses the mathematical qualifications corresponding to the completion of the bachelor education of Mathematical Sciences at Aalborg University as a minimum.

I would like to thank the National Center for Registerbased Research for providing data, an office and computer equipment, and in particular thank Preben Bo Mortensen, Janne Larsen, Aske Astrup, Christiane Gasse and Ane Møller for help when needed. Also thanks to my supervisor Svante for his always competent guidance.

Reading instructions
References throughout the report will be presented according to the Harvard system.
We always refers to the undersigned.
Figures, tables, and equations are enumerated in reference to the chapter.

Aalborg University, April 10, 2015

Sally Kit Ipsen
<smorte06@student.aau.dk>

Chapter 1

Introduction

Dementia is a chronic or progressive syndrome that characterizes certain symptoms of failing brain function. The clinical term is used to describe a condition with loss of memory, cognitive problems, and deterioration in the ability to perform everyday activities. Internationally, dementia is defined as a syndrome that includes *several* weakened cognitive functions [Socialministeriet, 2010, p. 14]. When elderly people suffer from other diseases such as pneumonia, different kinds of cancer, or inadequate fluid intake, a state of confusion can occur. The symptoms are similar to the symptoms of dementia, but they will disappear when the illness is treated. Many different illnesses can cause dementia, but the most common one is Alzheimer disease which contribute to $60 - 70\%$ of the cases. In a patient with Alzheimer the nerve cells in several areas of the brain slowly perish. Other causes of dementia are blood clots, Parkinson, and long-term alcoholism. Almost all causes have one thing in common; they are not treatable. WHO [2015].

The first signs of dementia are loss of memory, confusion, and a decreasing ability to function in every day life. Especially the relatives notice the first signs but many of the patients are also aware of the loss of memory. As the illness progresses the patient looses the feeling, that something is wrong. The period up until this point can be difficult to go through, and some patients show signs of depression. As the discomfort for the patient decreases, the surroundings and the relatives suffer from an increasing level of problems. Hasselbach [2013]. The degree of severity dementia is divided into 3 categories; mild, moderate, and severe dementia. Particularly in the last category both physical and verbal aggression, and agitation appears frequently [Socialministeriet, 2010, p. 15].

The validity of dementia diagnoses in the Danish nationwide hospital registers has not been evaluated since 2007, but at that time a registered diagnosis of dementia was found to be correct in 85.8% of the cases. This high validity allows for epidemiological studies based on the Danish hospital registers Phung et al. [2007a]. A nationwide study of incidence and prevalence of registered dementia diagnoses in the

Danish national hospital registers was performed in 2009. The results of the study showed that the incidence and prevalence were increasing over time. The published paper concluded that the diagnostic rate of dementia in secondary care over time has improved [Phung et al., 2007b, p. 146]. Approximately 80,000 people in Denmark live with dementia today. Since we live longer, it is anticipated that the 80,000 will be doubled by the year 2035 [Socialministeriet, 2010, p. 9].

The formere mentioned agitation and aggressive behavior, which often occurs in patients with severe dementia, is distressing for the patient and can represent risk for both the patient and for the relatives and nursing staff. Before psycho pharmacological therapy is taking into consideration, a lot of literature suggests non pharmacological approaches and treatment of pain. When pharmacological treatment is needed, it is most common to prescribe antipsychotic medication to address agitation and aggression. However, evidence indicates, first, that this medication only has a limited effect on agitation and, second, that it has a modest but significant improvement in aggression in the short term. There is less positive evidence that the medication helps the patient in the long term. In addition, prescriptions of more than 12 weeks are associated with cumulative risk of severe adverse events including death Ballard and Corbett [2013].

The Danish Health and Medicines Authority has issued warnings and guidelines for the use of antipsychotic medication in elderly patients in 1991, 2000, 2004, 2005, and 2007. Among other reasons is the almost three times higher risk of death found in elderly patients with dementia using antipsychotic medication compared with non-users. for Rationel Farmakoterapi [2006]. The guide from 2007 from the Danish Health and Medicine Authority states that elderly patients with dementia should not be treated with antipsychotics for more than one week. Long-term treatment should only be performed when the symptoms are severe and prolonged despite the relevant non pharmacological therapy, and only in cooperation with a specialist doctor in psychiatry [Sundhedsstyrelsen, 2013, p. 73]. In Denmark, the over-use of antipsychotic medication is only estimated and its actual extent is currently unknown [Sundhedsstyrelsen, 2005, p. 1].

To sum up the problems figure 1.1 shows a sketch with some plotted lines. The darkest line illustrates the increasing incidence and prevalence of dementia due to both changes in diagnostics and attitudes, and to the aging population. The three horizontal lines indicates the warnings and guidelines issues from the Danish Health and Medicines Authority. The two graphs with question marks are examples of the mortality rate and the use of antipsychotics. One could imagine an decreasing mortality rate among dementia patients due to earlier detection and treatment options with antidementia drugs, but no one examined it yet. Likewise with the use of antipsychotic medication; it could be decreasing after a warning or guideline has been issued, but a conclusive study has not been conducted.

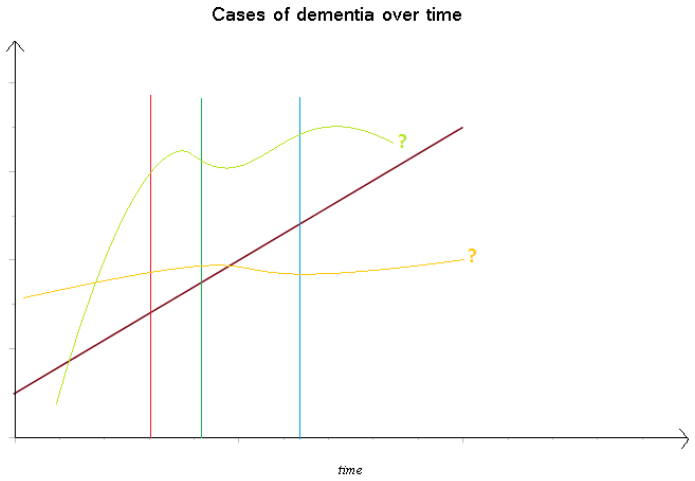

Figure 1.1: Sketch to illustrate the problem.

Based on the guidelines issued from the Danish Health and Medicines Authority, the use and potential over-use of antipsychotic medication in elderly patients is expected to be decreasing. However, no one has performed a thorough analysis [Sundhedsstyrelsen, 2013, p. 73], so a lot of elderly patients suffering from dementia might find their health compromised despite the guidelines. It is important that such an analysis is carried out, because if the guidelines issued by the Danish Health and Medicine Authority are not followed, the authorities need to take action immediately. Very often the elderly patients with dementia are not capable of taking care of themselves, and for that reason we as a society have a responsibility; their lives should not be at risk because of medication which has not even proven effective.

Such a thorough analysis can be difficult to perform only using trendlines and time series analysis, since neither of them take dependency between samples in consideration. The correlation within subjects is an important part, and should not be ignored. Therefore the theory of longitudinal data is useful; this theory allows repeated observations on each experimental unit to depend on each other, and these units can be assumed independent of one another. This means that we can make more robust inferences using the theory of longitudinal data rather than fx a time series analysis.

The aim of this project is to examine the theory of longitudinal data needed in order to conduct an analysis of the use of antipsychotic medication in elderly patients with dementia. Furthermore, the relevant data from the Danish registers are presented.

Chapter 2

Data

This chapter presents the registers from where the data are extracted and an overview of the data included in the study.

2.1 Data sources

The Danish National Patient Register, from now on called NPR, was established in 1977. At that point the register only covered somatic inpatients, but over the years it has been expanded, so it today covers both somatic and psychiatric in- and outpatients in all hospitals. The data in NPR can be divided into two categories; the clinical kind and the administrative kind. The data are organized so that each variable has a limited number of codes. In 1994 a major change took place because Denmark adapted the International Classification of Diseases version 10, (ICD-10). Before that, diagnoses were coded according to ICD-8 Lynge et al. [2011]. Reporting all activity from private hospitals is mandatory, but it is known to be incomplete. The National Board of Health estimated in 2008, that 5% of all operations were missing Lynge et al. [2011].

In the Danish **Lægemiddelstatistikregisteret** (LMSREG) we have access to information about prescripted medicine. It contains information about the overall sale of medicine in Denmark. This registration began in January 1994 with data from the pharmacies. In 1997 the information was supplemented with reporting from Statens Serum Institut, the hospital pharmacies, and veterinary clinics. Hence the register is said to be complete from the year 1997. The data are sent twice a year from LMSREG to Statistics Denmark, where they can be accessed by scientists via Laegemiddeldatabasen (LMDB). LMSREG does not contain information about herbal medicine, and it only includes prescription drugs, ie. medicine sold at a pharmacy to a citizen with a CPR-number. Consequently, all the data are person referable and can be cross referenced with other databases [Statistik, 2013, p. 4].

Since 1970, **The Danish Psychiatric Central Research Register**, (PCRR), has electronic record of patients treated at psychiatric departments in Denmark. In 1938, there were eight mental hospital that treated all psychiatric inpatients in Denmark. A systematic collection of nationwide clinical data on patients admitted to Danish mental hospitals began on that year. The first psychiatric departments in general hospitals were not established until 1956, but hereafter they also contributed to this non-electronic register. Then in April 1969 a nationwide electronic database was established and named Psychiatric Central Register. Consequently, the first year with complete information on every psychiatric admission is 1970. Data on outpatient treatment and emergency room contacts was included from 1995 and onwards. The authority responsible for the data in PCRR is The National Board of Health. No systematic validation studies of the clinical diagnoses has been conducted, but validation of some diagnoses exist in several studies, and these show good results. It is important to remember that in Denmark most cases of mild to moderate mental disorders are diagnosed and treated by the general practitioners, and therefore not registered in PCRR Mors et al. [2011].

2.2 Data description

The patients suffering from dementia are found in NPR, LMSREG and PCRR through the following definition: if a subject is given a ICD-10 code corresponding to any kind of dementia *or* if the subject at any time has a prescription for antidementia medicine, the subject is defined as a patient with dementia. Subjects diagnosed before their 65th birthday are excluded from the study, because the validity of dementia diagnoses is lower for patients under 65 years of age.

In the following each variable from LMSREG, available in the study, is described.
EKSD is the date of the customer transaction [Statistik, 2013, p. 11].
APK is the number of packs [Statistik, 2013, p. 14].
DOSO is the dosage of the prescription [Statistik, 2013, p. 42].
ATC is the ATC-code of the medicine [Statistik, 2013, p. 78].
DOSFORM is the form of the medicine, for instance tablets or injections [Statistik, 2013, p. 79].
PACKSIZE is the size of the pack. If the doseform is tablet, this numeral is the number of pills, but if the doseform is injection, it can be different things e.g. grams or milliliters [Statistik, 2013, p. 81].
STRNUM is the strength of the medicament, ie. how many mg per pill [Statistik, 2013, p. 87].
STRUNIT is the unit of the strength of the medicine [Statistik, 2013, p. 88].
Volume is one pack's numeric volume. It is typically measured as DDD (Defined Daily Dosage), which is a value set by WHO, but there are also other types of volume measurements [Statistik, 2013, p. 89].

VolTypeCode is a code that indicates which measure is used in the Volume variable [Statistik, 2013, p. 90].

The most important ones here are EKSD and Volume, since they provide information about time and quantity.

Chapter 3

Theory of Longitudinal Data

This chapter introduces longitudinal data and describes its basic theory.

3.1 Introduction

A study where individuals are measured repeatedly through time is called a longitudinal study. A major advantage of this kind of study is its ability to separate cohort and age effects. An example of this is shown in figure 3.1, which illustrates hypothetical data on the relationship between reading ability and age. In part *a*

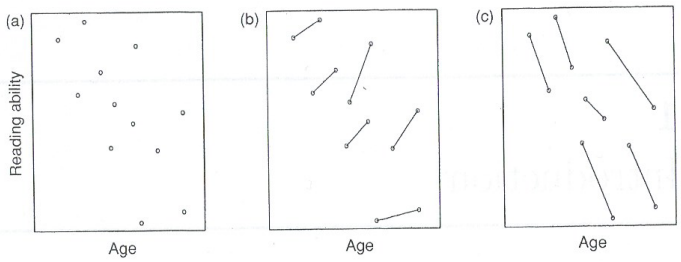

Figure 3.1: Hypothetical data describing the relation between reading ability and age.

reading ability is plotted against age for a hypothetical cross-sectional study of children. We see that reading ability appears to be poorer among older children. In part *b* we suppose the same data were obtained in a longitudinal study in which each individual was measured twice. It is clear that while younger children began at a higher reading level, everyone improved with time. If the data set were as in part *c*, the cross-sectional and longitudinal study show the same unlikely story; reading ability decreases with age. The example illustrates, that longitudinal studies can distinguish changes over time within individuals from differences among people in their same baseline level, which cross-sectional studies cannot. [Diggle et al., 2002, p. 1].

Longitudinal data can be collected either prospectively, following subjects forward in time, or retrospectively, by extracting multiple measurements on each person from historical records. The statistical methods described in this project apply to both situations. It is most common to collect longitudinal data prospectively, but we also use data collected retrospectively.

Because the set of observations on one subject tends to be intercorrelated, longitudinal data require special statistical methods that take correlation into account to draw valid scientific conclusions. This correlation is also an issue when analyzing a single long time series of measurements, but the analysis of longitudinal data tends to be simpler because subjects are usually assumed independent. The assumption about independence allows us to borrow strength across people, which means that the consistency of a pattern across subjects can provide valid inferences. Because of that we can obtain more robust results from longitudinal studies than from time series. [Diggle et al., 2002, p. 2].

Different types of longitudinal studies have some things in common; there are repeated observations on each experimental unit, these units can be assumed independent of one another, and the multiple responses within each unit are likely to be correlated. But there are also important differences among different examples. Some responses are continuous variables, while others may be binary or a count. Linear models can be used in the first case, but will not suffice in the two latter. Hence the choice of statistical model must depend on the type of outcome variable. [Diggle et al., 2002, p. 14].

Notation

In general, capital letters represent random variables or matrices, while small letters are for specific observations. Scalars and matrices will be in normal type, and vectors will be in bold type.

Let Y_{ij} represent a response variable and x_{ij} a vector of length p of explanatory variables observed at time t_{ij}, for observation $j = 1, \ldots, n_i$ on subject $i = 1, \ldots, m$.

The mean and variance of Y_{ij} are represented by $E(Y_{ij}) = \mu_{ij}$ and $Var(Y_{ij}) = \upsilon_{ij}$. The set of repeated outcomes for subject i are collected into an n_i-vector, $\mathbf{Y_i} = (Y_{i1}, \ldots, Y_{in_i})$, with mean $E(\mathbf{Y_i}) = \mu_\mathbf{i}$ and $n_i \times n_i$ covariance matrix $Var(\mathbf{Y_i}) = V_i$, where the jk element of V_i is the covariance between Y_{ij} and Y_{ik} denoted by $Cov(Y_{ij}, Y_{ik}) = \upsilon_{ijk}$. R_i denotes the $n_i \times n_i$ correlation matrix of $\mathbf{Y_i}$. The responses for all units are referred to as $\mathbf{Y} = (\mathbf{Y_1}, \ldots, \mathbf{Y_m})$, which is an N-vector with $N = \sum_{i=1}^{m} n_i$.

3.2 Graphical presentation of longitudinal data

This section is best illustrated with an example. This example consist of bodyweights of 48 pigs in 9 successive weeks of follow-up. A scatterplot of the response variable against time is always a good first graph for longitudinal data. A number of impor-

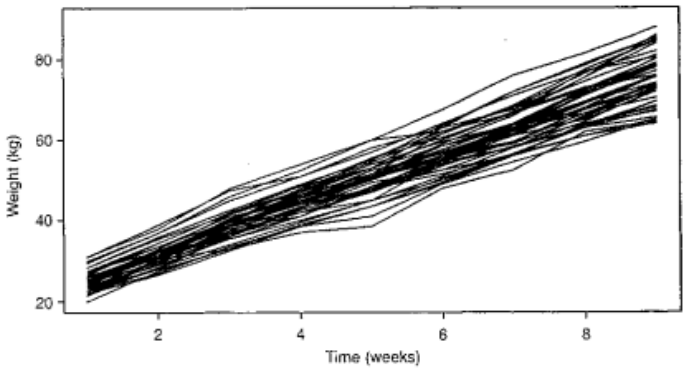

Figure 3.2: Scatterplot of pigs example

tant patterns are apparent in the simple graph in figure 3.2. First, all animals are gaining weight. Second, the pigs which are largest at the beginning of the observation period tend to be largest throughout. This phenomenon is called *tracking*. Third, at the beginning of the study the spread among the 48 pigs is substantially smaller than at the end. [Diggle et al., 2002, p. 34-35].

Although it is hard to pick out individual response profiles, the scatterplot is an adequete display for exploring these growth data. That part can be done slightly better by making a display obtained from the one in figure 3.2 by standardizing each observation. The data are standardized by subtracting the mean and then dividing by the standard deviation of the 48 observations at each time. Mathematically it is expressed by the following equation:

$$y_{ij}^{*} = \frac{(y_{ij} - y_j)}{s_j}. \tag{3.1}$$

The effect of this standardization is shown in figure 3.3. This plot is able to highlight the degree of tracking whereby animals tend to maintain their relative size over time. [Diggle et al., 2002, p. 35].

With large data sets, connected line graphs become unduly cluttered, and an alternative strategy for the basic time plot is needed. An example of $CD4+$ cell number is now introduced. The HIV (the human immune deficiency virus) attacks an immune cell called $CD4+$ cell. An uninfected individual has around 1100 cells per millilitre of blood and this number decreases in number with time from infection.

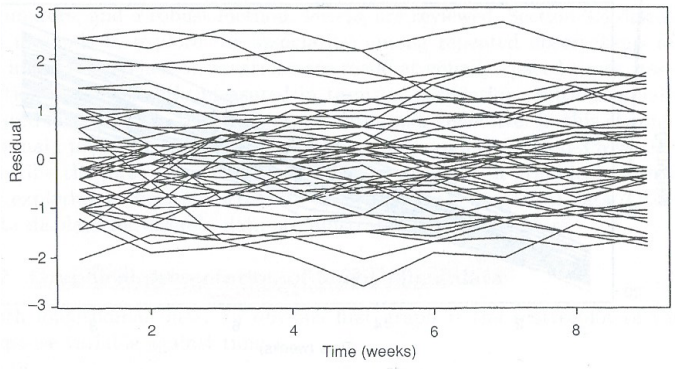

Figure 3.3: Residuals of pigs example.

Hence, an infected person's $CD4+$ cell number can be used to monitor disease progression. This example consists of 2376 values of $CD4+$ cell number for 369 infected men. The values are plotted against time (years) since seroconversion (the time when HIV becomes detectable). The scatterplot in figure 3.4 displays the data for all men, and with each persons repeated observations connected. This plot is unfortunately

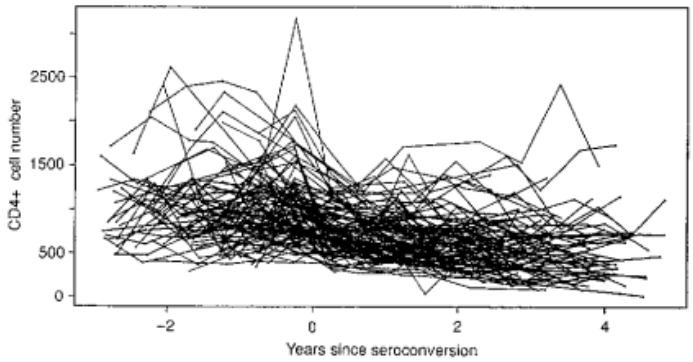

Figure 3.4: Scatterplot of $CD4+$ example.

extremely busy, which reduces its usefulness. The issue is, that on the one hand repeated measurements are connected to display changes through time for individuals, while on the other hand presenting every person's curve creates little more than confusion in large data sets. An ideal solution could be to display each individual's data with a very thin gray line and to use darker lines for the typical pattern, for example the average, as well as for a subset of persons. [Diggle et al., 2002, p. 35-37].

There is an alternative solution, which involves connecting the repeated measurements for only a judicious selection of individuals. The simplest way is to just choose a subset at random, but two problems might arise with that method. First, it is possible to obtain a non-representative group by chance, and second, the display is unlikely to uncover outlying individuals. Therefore a second approach is preferable:

the individual curves are ordered with respect to some characteristic that is relevant to the model of interest, and then the data for individuals with selected quantiles for this ordering statistic are connected. Resistant statistics are preferred for this ordering, because one or a few outlying observations should not determine an individual's summary score. Examples of resistant statistics are the median, median absolute deviation, and the biweight trend. Of course, when the data naturally divide into for example treatment groups, a separate plot can be made for each group, or one plot can be produced with a separate summary curve for each group and distinct plotting symbols for the data from each group. [Diggle et al., 2002, p. 37-38].

Figure 3.5 shows the residuals from the average curve with the repeated values for nine individuals connected. These persons had median residual value at the extrema or the $5th$, $10th$, $25th$, $50th$, $75th$, $90th$, or $95th$ percentile. Residuals are used instead of raw data, as they sometimes help to uncover more subtle patterns than individual curves. These data have considerably more variation across time than the

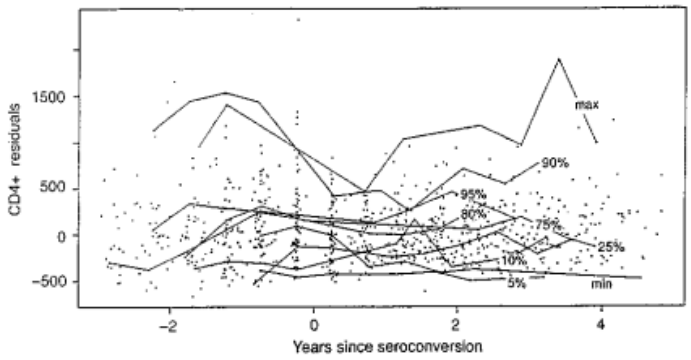

Figure 3.5: Residuals of CD4+ example

data shown in figure 3.3. [Diggle et al., 2002, p. 38-39].

3.3 Fitting smooth curves to longitudinal data

Three different non-parametric regression methods on how to fit the mean response profile as a function of time will be described in this section. The methods are called *Kernel*, *Spline* and *Lowess* estimation respectively.

With the purpose of simplifying the following descriptions, it is assumed that there is only one observation on each individual at each time. The observation observed at time t_i is denoted y_i.

Kernel estimation is illustrated in figure 3.6. In the figure we see three windows, each centered at different time point. All points visible in the windows are highlighted with circles. The average Y value of the points in one window, is the

estimated mean response at time t_i. This means that the estimated mean response at time t_1, $\hat{\mu}(t_1)$, is calculated as the average value of the points visible in the window centered at time t_1. To estimate the smooth curve, let a window slide from left to

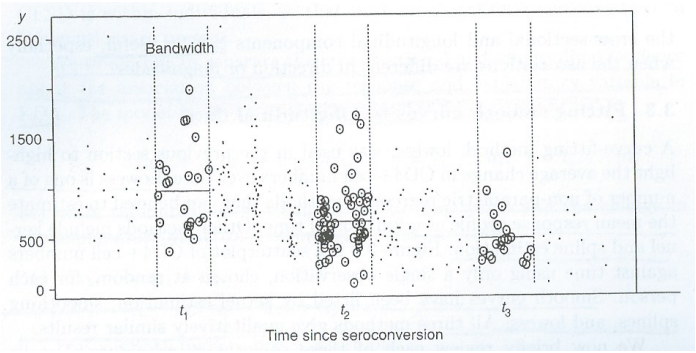

Figure 3.6: Illustration of the Kernel smoother.

right across the data. At every time calculate the average of the points within the window. Different window sizes called bandwidths can be chosen. The extremes are the narrowest window that will interpolate the data, and the broadest window, that includes all observations, which will result in a horizontal line, a constant value equal to the average value of all observed values. Generally, a wider window results in a smoother curve.

Instead of just taking the straight average of the points in each window, we can use a weighting function that puts more weight on points close to t and less weight to points further away from t. The straight average method is called "using a box-car window". The alternative with the weighting function is a better strategy. The definition of the kernel estimation is expressed in this equation:

$$\hat{\mu}(t) = \frac{\sum_{i=1}^{m} w(t, t_i, h) y_i}{\sum_{i=1}^{m} w(t, t_i, h)},$$

where the function w is

$$w(t, t_i, h) = K\left(\frac{t - t_i}{h}\right),$$

and h is the bandwidth of the kernel. The Gaussian kernel, which is $K(u) = exp(-0,5u^2)$, is a commonly used weighting function. As wider windows create smoother curves, so will larger values of h. A comparison between two kernel estimators with different bandwidths is shown in figure 3.7. [Diggle et al., 2002, pp. 41-43]. **Spline** is the second smoothing method described in this section. The function $s(t)$, which minimizes the criterion

$$J(\lambda) = \sum_{i=1}^{m} (y_i - s(t_i))^2 + \lambda \int (s''(t))^2 dt, \tag{3.2}$$

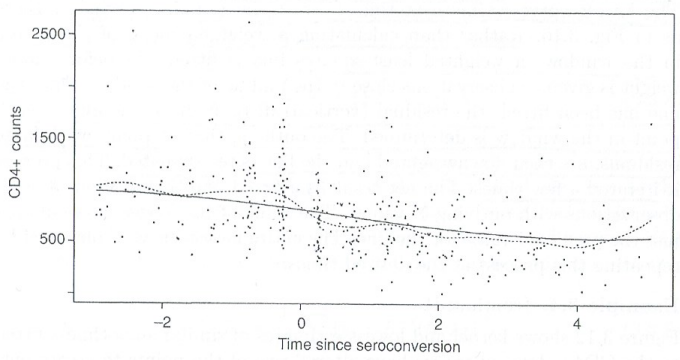

Figure 3.7: Large (solid) vs. small (dotted) bandwidth in the Kernel smoother.

is called a cubic smoothing spline, where $s''(t)$ is the second derivative of $s(t)$. The first part is a measure of the accuracy of the function $s(t)$ compared to the observations y_i. The part with the integral quantifies how rough and bumpy the curvature is. The quantity λ determines how smooth the spline is. Smaller λ's give rougher curves whereas larger λ's give smoother curves.

It can be shown that the spline minimizing the criterion in equation 3.2, is a twice-differentiable piecewise cubic polynomial. The spline can be calculated using the observations and by solving linear equations. [Diggle et al., 2002, p. 44]

Lowess is an extension of the kernel method that is less sensitive to outliers. Therefore, this smoothing is said to be more robust. The point of origin is the same as in the Kernel smoother; a window is centered as in figure 3.6. But, instead of calculating a weighted mean of the points, we fit a weighted least-squares line. As in the Kernel estimation, the points closer to the center of the window are given more weight than the observations further away. When this line has been fitted, we determine the residual to each point in the window. Large residuals belong to outliers and are then downweighted before the line is fitted again. The process is repeated a few times, consequently it is an iterative method. The resulting line is insensitive to outliers. The predicted value for the line at t_i is the value of the Lowess estimate at that point. To obtain the entire Lowess curve repeat this method for the requested times.

A comparison between Kernel (solid) and Lowess (dotted) estimation is shown in figure 3.8, where two observations have been altered to create outliers. It is obvious that the Kernel smoother is more affected by those outliers than the Lowess smoother. [Diggle et al., 2002, p. 44].

When working with longitudinal data, it is better to use robust smoothing methods. In all three methods described here, there is a bandwidth parameter that determines

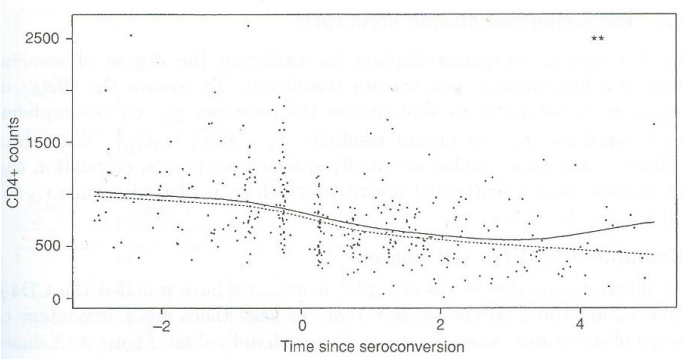

Figure 3.8: Comparison between Kernel (solid) and Lowess (dotted) smoother.

the smoothness of the estimated curve. There is a trade-off between variance and bias when we choose this smoothing parameter; a larger bandwidth equals smaller variance of the fitted curve, but are also more bias. The key is to choose a bandwidth that balances bias and variance. In figure 3.9 we see a comparison of all three smoothing methods, and overall the results are similar, at least in the example with the $CD4+$ data. [Diggle et al., 2002, p. 41,45].

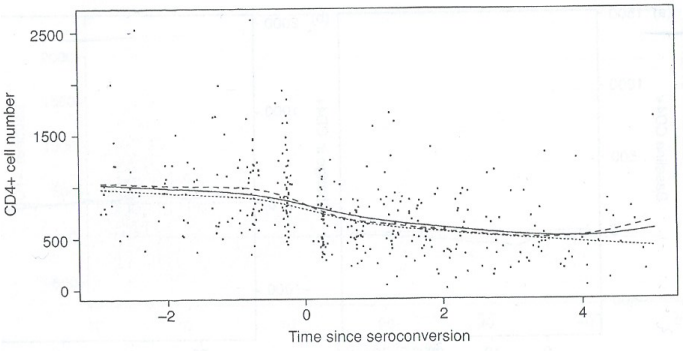

Figure 3.9: Comparison between the three smoothing methods; Kernel (solid), Spline (dashed), and Lowess (dotted).

Chapter 4

Models for longitudinal data

The aim of this chapter is to describe a general linear modeling framework for longitudinal data. Here the inferences made about the estimated parameters recognize the likely correlation structure in the data. This can be achieved by building explicit parametric models of the covariance structure and checking their validity. In some cases it is possible to use another method of inference, robust to misspecification of the covariance structure. That means the inference made is valid, even though the true covariance structure is not as assumed. We start the chapter describing a general linear model for correlated data. Afterward, we display some approaches to parameter estimation, which are not restricted to specific models for the covariance structure. For parameters describing the mean response, least-squares estimation is used, while for the covariance parameters either maximum likelihood or restricted maximum likelihood are used. [Diggle et al., 2002, p. 54].

4.1 The general linear model

As usual we have m subjects each with n observations y_{ij}, where $j = 1, \ldots, n$ denotes the sequence of observed measurements on the ith subject. With p explanatory variables we have, that each y_{ij} is associated with values x_{ijk}, where $k = 1, \ldots, p$. The y_{ij} are assumed to be realizations of random variables:

$$Y_{ij} = \beta_1 x_{ij1} + \cdots + \beta_p x_{ijp} + \epsilon_{ij}.$$

Here the ϵ_{ij} are random sequences of length n respectively matched to the m subjects. It is a classical assumption that the ϵ_{ij} are mutually independent random variables, but in this model according to the longitudinal structure of the data, it is expected that ϵ_{ij} are correlated within subjects. [Diggle et al., 2002, p. 55].

Expressed as matrix formulation we have $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{in})$ denoting the observed sequence of measurements on the ith subject. Furthermore the complete set of $N = nm$ measurements for m units are expressed as $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_m)^T$. We have an $N \times p$ matrix of explanatory variables labeled X, where the $(n(i-1)+j)$th

row is denoted $(x_{ij1}, x_{ij2}, \ldots, x_{ijp})$. The variance matrix for the vector of measurements on a single subject is represented by a non-zero $n \times n$ block $\sigma^2 V_0$. When combined we end up with a block-diagonal matrix called $\sigma^2 V$. With all of this assumed, the general linear model for longitudinal data processes $\mathbf{y}$ as drawn from a multivariate Gaussian random vector labeled Y. We write the model like this:

$$\mathbf{Y} \sim MVN(X\beta, \sigma^2 V). \tag{4.1}$$

The block-diagonal structure of $\sigma^2 V$ is crucial if we wish to use robust methods analyzing the data. This is because we are then able to use the replication across units to estimate $\sigma^2 V$ without having to make parametric assumptions about its form. [Diggle et al., 2002, p. 55].

4.2 Weighted least-squares estimation

The aim of this subsection is to look further into estimating the regression parameter β in the model described in equation 4.1. The value, $\tilde{\beta}_W$, which minimizes the quadratic form

$$(\mathbf{y} - X\beta)^T W (\mathbf{y} - X\beta), \tag{4.2}$$

is the so called *weighted least-squares* estimator of β, where W is a symmetric $N \times N$ weight matrix. It is later shown what impact this matrix has on the properties of the estimate of β. The dimensions of the other parameters are as follows: $\mathbf{y}$: $N \times 1$, X: $N \times p$ and β: $p \times 1$. Here p denotes the number of explanatory variables, and $N = nm$. Using matrix manipulations on expression 4.2 we can calculate the estimator for β. We start by multiplying W into the parenthesis, and since $(AB)^T = (B^T A^T)$ we can rewrite expression 4.2 as:

$$\left(\mathbf{y}^T - \beta^T X^T\right)(W\mathbf{y} - WX\beta).$$

Then by multiplying the two parenthesis we obtain:

$$\mathbf{y}^T W\mathbf{y} - \mathbf{y}^T WX\beta - \beta^T X^T W\mathbf{y} + \beta^T X^T WX\beta.$$

The two terms in the middle are both scalars, because their dimensions are $1 \times N \times N \times N \times N \times p \times p \times 1 = 1$ and $1 \times p \times p \times N \times N \times N \times N \times 1 = 1$ respectively. Therefore they can be added:

$$\mathbf{y}^T W\mathbf{y} - 2\mathbf{y}^T WX\beta + \beta^T X^T WX\beta.$$

To determine the β, that minimizes this expression, it is differentiated with respect to β and equated to zero:

$$\begin{aligned} \frac{\partial}{\partial\beta}\left(\mathbf{y}^T W\mathbf{y} - 2\mathbf{y}^T WX\beta + \beta^T X^T WX\beta\right) &= \\ -2\left(\mathbf{y}^T WX\right)^T + 2X^T WX\beta &= 0. \end{aligned}$$

In order to differentiate matrices two rules were used [Azzalini, 1996, p. 170]:

$$\frac{d}{dx}Ax = A^T \qquad \text{and} \qquad \frac{d}{dx}x^TBx = 2Bx.$$

Standard matrix manipulation and algebra are used to obtain the result:

$$\begin{aligned} -2\left(\mathbf{y}^TWX\right)^T + 2X^TWX\beta &= 0 \\ 2X^TWX\beta &= 2X^TW^T\mathbf{y} \\ X^TWX\beta &= X^TW^T\mathbf{y} \\ \beta &= \left(X^TWX\right)^{-1}X^TW\mathbf{y}. \end{aligned}$$

It is now clear, that the value, $\tilde{\beta}_W$, which minimizes expression 4.2, is:

$$\tilde{\beta}_W = \left(X^TWX\right)^{-1}X^TW\mathbf{y}. \tag{4.3}$$

No matter what value is chosen for W, the weighted least-squares estimator $\tilde{\beta}_W$ is unbiased. This is because the measurements, $\mathbf{y}$, are realizations of a random vector $\mathbf{Y}$, where the mean is $E(\mathbf{Y}) = \mathbf{X}\beta$, as described in the previous section.

In the model expressed in equation 4.1 we also see, that the variance is

$$Var(\mathbf{Y}) = \sigma^2V.$$

Therefore we can conclude, that the variance of the estimator is:

$$Var(\tilde{\beta}_W) = \sigma^2\left(\left(X^TWX\right)^{-1}X^TW\right)V\left(WX\left(X^TWX\right)^{-1}\right). \tag{4.4}$$

[Diggle et al., 2002, p. 59-60].

Now we take a look at the estimator in two the cases, where $W = I$ and $W = V^{-1}$ respectively. The first case in which W equals the identity matrix, I, the weighted least-squares estimator in equation 4.3 is reduced to the ordinary least-squares estimator. Hence

$$\tilde{\beta}_I = \left(X^TX\right)^{-1}X^T\mathbf{y}$$

with

$$Var(\tilde{\beta}_I) = \sigma^2\left(X^TX\right)^{-1}X^TVX\left(X^TX\right)^{-1}. \tag{4.5}$$

In the second case, where $W = V^{-1}$, the weighted least squares becomes

$$\hat{\beta} = \left(X^TV^{-1}X\right)^{-1}X^TV^{-1}\mathbf{y} \tag{4.6}$$

with

$$Var(\hat{\beta}) = \sigma^2\left(X^TV^{-1}X\right)^{-1}. \tag{4.7}$$

Under the multivariate Gaussian assumption in equation 4.1 $\hat{\beta}$ denotes the maximum likelihood estimator for β. MLE is the most efficient weighted least-squares estimator and determined by using $W = V^{-1}$. But this most optimal weighting matrix can only be calculated if the complete correlation structure of the data is known. In practice, the correlation structure is difficult to identify. The relative efficiency of the two different estimates, $\tilde{\beta}_W$ and $\hat{\beta}$, can be calculated from their respective variance matrices in equations 4.4 and 4.7. By comparing the calculated efficiencies it is possible to evaluate the loss of efficiency connected to using a different weighting matrix W. [Diggle et al., 2002, p. 60].

Through comparisons it is found that in *some* cases with a balanced design, the ordinary least-squares estimator, $\tilde{\beta}_I$, is completely sufficient for point estimation. But this is far from being always true. In an example with two-treatment crossover design the comparison shows that a careful balancing of between-subjects and within-subjects comparisons of the two treatments is necessary in order to obtain efficient estimation. And this balancing depends critically on the covariance structure of the data. [Diggle et al., 2002, p. 62-63].

From the form of the variance of $\tilde{\beta}$ in equation 4.5 it is given, that interval estimation for β requires information about the variance matrix, $\sigma^2 V$. This is required even when the OLS estimator is reasonably efficient. For example take a look at the variance of the least-square estimator:

$$Var(\tilde{\beta}_I) = \sigma^2 (X^T X)^{-1}. \tag{4.8}$$

This formula assumes that $V = I$, and when it is not so, it can be seriously misleading. If the correlation structure in the data is ignored, and the interval estimation for β is based on the formula in equation 4.8, where σ^2 is replaced by the residual mean square, its usual estimator, we have what is called a naive approach. When $V \neq I$, two sources of error occur; first formula 4.8 is wrong, and second $\tilde{\sigma}^2$ is no longer an unbiased estimator. In order to evaluate the combined effect from the two sources of error, a comparison between the diagonal elements of $Var(\tilde{\beta})$ as defined in equation 4.5 and the corresponding elements of the matrix $E(\tilde{\sigma}^2)(X^T X)^{-1}$ is necessary. We will not dig deeper into such comparisons here, but the conclusion based on others' analysis is that there can be a big risk of seriously over- or underestimating the variance of β, when a positive autocorrelation is present and the approach is a naive use of OLS. Such over- or underestimation also depends on the design matrix. [Diggle et al., 2002, p. 63-64].

4.3 Maximum likelihood estimation

In the general linear model we wish to estimate the parameters of interest; β, σ^2 and V_0. One method of doing that is to estimate them simultaneously using the

likelihood function under the Gaussian assumption in expression 4.1. In this case the related log-likelihood function for observed data $\mathbf{y}$ is this:

$$L(\beta, \sigma^2, V_0) = -0,5 \left(nmlog(\sigma^2) + mlog(|V_0|) + \sigma^{-2}(\mathbf{y} - \mathbf{X}\beta)^T V^{-1}(\mathbf{y} - \mathbf{X}\beta) \right). \tag{4.9}$$

For a given V_0 the weighted least-squares estimator in equation 4.6 is the maximum likelihood estimator for β:

$$\hat{\beta}(V_0) = \left(X^T V^{-1} X \right)^{-1} X^T V^{-1} \mathbf{y}. \tag{4.10}$$

This estimator substituted into equation 4.9 produces this:

$$L(\hat{\beta}(V_0), \sigma^2, V_0) = -0,5 \left(nmlog(\sigma^2) + mlog(|V_0|) + \sigma^{-2} RSS(V_0) \right), \tag{4.11}$$

where RSS denotes the residual sum of squares:

$$RSS(V_0) = (\mathbf{y} - \mathbf{X}\hat{\beta}(V_0))' V^{-1} (\mathbf{y} - \mathbf{X}\hat{\beta}(V_0)).$$

If equation 4.11 is differentiated with respect to σ^2 and equated to zero, the maximum likelihood estimator for σ^2 is obtained. Hence for a fixed V_0 we have,

$$\hat{\sigma^2}(V_0) = \frac{RSS(V_0)}{nm}. \tag{4.12}$$

Now both equations 4.10 and 4.12 are substituted into the log-likelihood function in equation 4.9, and a so-called *reduced* or *partial maximized* log-likelihood for V_0, apart from a constant term, is obtained:

$$\begin{aligned} L_r(V_0) &= L\left(\hat{\beta}(V_0), \hat{\sigma^2}(V_0), V_0 \right) \\ &= -\frac{1}{2} m \left(nlog(RSS(V_0)) + log(|V_0|) \right). \end{aligned}$$

Last $\hat{V_0}$ is yielded from maximizing $L_r(V_0)$, and when substituted into equations 4.10 and 4.12, the following maximum likelihood estimators are achieved [Diggle et al., 2002, p. 64-65]:

$$\hat{\beta} \equiv \hat{\beta}(\hat{V_0}) \qquad \text{and} \qquad \hat{\sigma}^2 \equiv \hat{\sigma^2}(\hat{V_0}).$$

When using maximum likelihood for the simultaneous estimation of the three parameters, the form of the design matrix X is directly involved. Therefore if a wrong form for X is assumed, the estimators for σ^2 and V_0 may not be consistent. This problem can be dealt with by using an over-elaborate model for the mean response profiles in estimating the covariance structure. Unfortunately it is not always possible to use this approach, which might be problematic in another way. The approach requires using a design matrix with a large number of columns, and hereby biased estimates occur. Hence maximum likelihood estimation presents a conflict; do we want consistent estimates or unbiased estimates? When we are confident that an adequate model can be defined using a design matrix with a small number of columns, it is not a serious conflict. Otherwise other methods of estimation have to be considered. One of these methods is the method of restricted maximum likelihood. [Diggle et al., 2002, p. 65].

4.4 Restricted maximum likelihood estimation

Restricted maximum likelihood, REML, estimation is a method of estimating variance components in a general linear model. As described in the previous section, the problem with the standard maximum likelihood estimation is that biased estimators of the covariance parameters are produced. Recall the case of the general linear model with independent errors:

$$\mathbf{Y} \sim MVN\left(X\beta, \sigma^2 I\right), \tag{4.13}$$

where the maximum likelihood estimator for σ^2 is:

$$\hat{\sigma}^2 = \frac{RSS}{(nm)}.$$

The usual unbiased estimator, when p is the number of elements in β, is:

$$\tilde{\sigma}^2 = \frac{RSS}{(nm-p)}. \tag{4.14}$$

Actually, $\tilde{\sigma}^2$ is exactly the REML estimator for σ^2 in the model in expression 4.13.

Now look at the general linear model with dependent errors:

$$\mathbf{Y} \sim MVN\left(X\beta, \sigma^2 V\right).$$

Here the REML estimator is based on a linearly transformed set of data $\mathbf{Y}^* = A\mathbf{Y}$, such that the distribution of $\mathbf{Y}^*$ does not depend on β. Now the REML estimator is defined as a maximum likelihood estimator for $\mathbf{Y}^*$. Taking A to be the matrix, which converts $\mathbf{Y}$ into OLS residuals,

$$A = I - X\left(X^T X\right)^{-1} X^T,$$

is one method of obtaining the requested distribution of $\mathbf{Y}^*$. In this case, no matter what value of β, $\mathbf{Y}^*$ has a singular multivariate Gaussian distribution with zero mean. By using a matrix A with only $nm - p$ rows, we could achieve a non-singular distribution. It turns out that any full-rank matrix with the property that $E(\mathbf{Y}^*) = 0$ for all β, gives the same result. Therefore, the estimators for σ^2 and V do not depend on which rows are used or on the particular choice of A. Hence the transformation from $\mathbf{Y}$ to $\mathbf{Y}^*$ does not need to be explicit. [Diggle et al., 2002, p. 66].

The REML estimator for σ^2 is:

$$\tilde{\sigma}^2(V_0) = \frac{RSS(V_0)}{nm-p}. \tag{4.15}$$

Again p denotes the number of elements of β. We have this reduced log-likelihood function:

$$L^*(V_0) = -\frac{1}{2}m\left(n\log(RSS(V_0)) + \log(|V_0|)\right) - \frac{1}{2}\log\left(\left|X^T V^{-1} X\right|\right), \tag{4.16}$$

and the REML estimator for V_0 maximizes it. The result, $\tilde{V}_0$, substituted into equations 4.10 and 4.15 provides the following REML estimators:

$$\tilde{\beta} \equiv \hat{\beta}(\tilde{V}_0) \qquad \text{and} \qquad \tilde{\sigma}^2 \equiv \tilde{\sigma^2}(\tilde{V}_0).$$

[Diggle et al., 2002, p. 68-69].

4.5 Robust estimation

The basis for this method of robust estimation of β is to use the generalized least-square estimator $\tilde{\beta}_W$, which is defined in equation 4.3, together with an estimated variance matrix defined as

$$\hat{R}_W = \left((X^T W X)^{-1} X^T W\right) \hat{V} \left(W X (X^T W X)^{-1}\right). \tag{4.17}$$

Here the scale-parameter σ^2 is re-absorbed into V, and whatever the true covariance structure is, the estimate $\hat{V}$ is consistent for V. From now on inference is made as if

$$\tilde{\beta}_W \sim MVN(\beta, \hat{R}_W). \tag{4.18}$$

The matrix W^{-1} is called the working variance matrix. It is different from the true variance matrix V. A relatively simple form for W^{-1} is applied, but of course we still hope that it captures the qualitative structure of V. The crucial distinction between this and other approaches is that a poorly chosen W^{-1} only affects the efficiency of the inferences made for β, not their validity. Consequently, confidence intervals and tests of hypotheses made from the model in expression 4.18 will be asymptotically correct no matter what the true form of V is.

Using the OLS estimator $\tilde{\beta}$ is the simplest possible way. It would be equivalent to assuming that the measurements are uncorrelated within a subject. A more efficient way for data with a smoothly decaying autocorrelation structure, is to use a block-diagonal W^{-1}. This working variance matrix would have non-zero elements of the form $\exp(-c|t_j - t_k|)$, where c is a chosen positive constant that roughly matches the rate of decay anticipated. A more complicated W^{-1} is usually unnecessary. [Diggle et al., 2002, p. 70].

Now look at an experiment chosen to fit the saturated model for the mean response. The robust estimator $\hat{V}$ required in equation 4.17, is achieved by executing the following steps according to the REML principle. The measurements are made at each

of n time-points t_j on m_h experimental units in the hth of g experimental treatment groups. The complete set of measurements are written as:

$$y_{hij}, \qquad h = 1, \ldots, g, \qquad i = 1, \ldots, m_h, \qquad j = 1, \ldots, n.$$

The mean response according to the saturated model is

$$E(Y_{hij}) = \mu_{hj}, \qquad h = 1, \ldots, g, \qquad j = 1, \ldots, n.$$

The covariance structure according to a saturated model is $V = Var(\mathbf{Y})$; a block-diagonal, positive definite but otherwise arbitrary $n \times n$ matrix, where all non-zero blocks are equal to V_0.

The design matrix X has a rather special form. With $g = 2$ treatments and replications $m_1 = 2$ and $m_2 = 3$, it is this:

$$X = \begin{bmatrix} I & O \\ I & O \\ O & I \\ O & I \\ O & I \end{bmatrix}.$$

Here I denotes the $n \times n$ identity matrix, while O is the $n \times n$ matrix of zeros. The estimators for the mean, μ_{hj}, are the corresponding sample mean, meaning

$$\hat{\mu}_{hj} = m_h^{-1} \sum_{i=1}^{m_h} y_{hij}.$$

Furthermore the REML estimator for the variance matrix V_0 is

$$\hat{V}_0 = \left(\sum_{i=1}^{g} m_i - g \right)^{-1} \sum_{h=1}^{g} \sum_{i=1}^{m_h} (\mathbf{y}_{hi} - \hat{\mu}_h)(\mathbf{y}_{hi} - \hat{\mu}_h)^T, \tag{4.19}$$

where $\mathbf{y}_{hi} = (y_{hi1}, \ldots, y_{hin})^T$ and $\hat{\mu}_h = (\hat{\mu}_{h1}, \ldots, \hat{\mu}_{hn})^T$. Now $\hat{V}_0$ is a block-diagonal matrix with non-zero blocks and exactly the required estimate $\hat{V}$.

When the data comes from observational studies with continuously varying covariates, it is typically no longer possible to achieve an explicit expression for the REML estimate of V_0. In this case and others, where the saturated model strategy is not feasible, the same basic idea can still be applied. No assumptions about the form of the variance matrix are made, and then an X matrix corresponding to the most elaborate model for the mean response is used and the REML estimate $\hat{V}_0$ is obtained by numerical maximization of equation 4.16. [Diggle et al., 2002, p. 70-71].

Either case, the computed $\hat{V}$ is substituted into equation 4.17, and the model in

equation 4.18 is used to make robust inference for β. These inferences about β are typically made using an X matrix with a lot fewer columns than the ng of the corresponding matrix in the saturated model. The standard approach for general linear models can be used if we wish to test linear hypotheses about β within this model. When the hypothesis $Q\beta = \mathbf{0}$, where Q is a full rank matrix of dimensions $q \times p$ for some $q < p$, is getting tested, we start by looking at the model expressed in 4.18 and thereby obtaining this model:

$$Q\hat{\beta}_W \sim MVN(Q\beta, Q\hat{R}_W Q^T).$$

A suitable test statistic for the hypothesis is

$$T = \hat{\beta}_W^T Q^T (Q\hat{\beta}_W^T Q^T)^{-1} Q\hat{\beta}_W.$$

Here the approximate null sampling distribution of T is chi-squared with q degrees of freedom.

The following robust method is a modified, but useful approach to use when measurement times are not common to all units. The non-zero blocks in the required variance matrix, V, corresponding to the sets of measurements within units, are no longer constant between units. The set of measurements on the ith unit are put in an $n_i \times n_i$ matrix named V_{0i}, and the mean vector of these measurements is denoted μ_i. The mean vector is estimated by using, $\hat{\mu}_i$, the OLS estimates from the most complicated model that we are prepared to consider for the mean response. Then the following formula defines an estimate of V_{0i}:

$$\hat{V}_{0i} = (\mathbf{y}_{hi} - \hat{\mu}_h)(\mathbf{y}_{hi} - \hat{\mu}_h)^T. \tag{4.20}$$

Now $\hat{V}$ is the block-diagonal matrix with non-zero blocks $\hat{V}_{0i}$ defined by equation 4.20, and $\hat{V}$ is exactly the robust estimate of V. [Diggle et al., 2002, p. 72]

The approaches described in this section are relatively simple to use. Provided that the experimental design allows the fitting of a saturated model for the mean response, the REML estimates can be computed without many problems. Since the rest of calculations only involve standard matrix manipulation it should be fairly obtainable. The methods are designed so that consistent inferences for the mean response parameter are a result from a correct specification of the mean structure, no matter what the true covariance structure is. [Diggle et al., 2002, p. 79].

Chapter 5

Parametric model for covariance structure

In this chapter we take another look at the general linear model in equation 4.1. This time the covariance structure of the sequence of measurements on each experimental unit is to be investigated, and therefore a different notation is used. We have a vector of n_i measurements on the ith unit: $\mathbf{y}_i = (y_{i1}, \ldots, y_{in_i})$, and a corresponding vector of times at which these measurements were taken: $\mathbf{t}_i = (t_{i1}, \ldots, t_{in_i})$. The m units, corresponding to all the data $\mathbf{y}$, are written altogether as: $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_m)$, $\mathbf{t} = (\mathbf{t}_1, \ldots, \mathbf{t}_m)$. Furthermore $N = \sum_{i=1}^{m} n_i$. It is assumed that the measurements are drawn from mutually independent Gaussian random vectors, $\mathbf{Y}_i$, where:

$$\mathbf{Y}_i \sim MVN\left(X_i\beta, V_i(\mathbf{t}_i, \alpha)\right). \tag{5.1}$$

Here X_i denotes the $n_i \times p$ matrix of explanatory variables, and the dimensions of β and α are p and q respectively. The model for the entire set of data is written as

$$\mathbf{Y} \sim MVN\left(X\beta, V(\mathbf{t}, \alpha)\right), \tag{5.2}$$

where X is obtained by stacking the matrices for each unit, X_i, and $V(\cdot)$ is block-diagonal with the non-zero blocks $V_i(\cdot)$. Hence, the matrix X has dimension $N \times p$, and $V(\cdot)$ is an $N \times N$ matrix. By this notation it is implied that continuous time is the natural setting for most longitudinal data. Therefore specific models are derived by assuming that the sequences Y_{ij}, $j = 1, \ldots, n_i$, are sampled from independent copies of an underlying stochastic process, $\{Y(t), t \in \Re\}$, which is in continuous time. This comes together as: $Y_{ij} = Y_i(t_{ij})$, with $j = 1, \ldots, n_i$ and $i = 1, \ldots, m$. [Diggle et al., 2002, p. 81-82]

The principal tools used to describe the properties of the model are the covariance function and its relation, the variogram. Before moving on we present the variogram of a stochastic process, $\{Y(t)\}$, and it is defined as the function

$$\gamma(u) = \frac{1}{2}E\left(\left(Y(t) - Y(t-u)\right)^2\right), \qquad u \geq 0.$$

Let $\rho(u)$ denote the correlation between $Y(t)$ and $Y(t-u)$, and $\sigma^2 = Var\left(Y(t)\right)$, then for a stationary process $Y(t)$ the variogram equals:

$$\gamma(u) = \sigma^2 \left(1 - \rho(u)\right).$$

Within the general framework of equation 5.1 different kinds of stationary and non-stationary behavior arise, and in the following section we look at an example of a model that fits the aim of this project. Methods for fitting the model to data are developed and described later on. [Diggle et al., 2002, p. 92].

5.1 Model

There exist at least three different sources of random variation in longitudinal data, and of course we wish to include as many of these as possible in a useful model. So before developing the model, we look at these likely sources in order to understand them qualitatively.

1. Random effects: A simple example of random effects is when the general level of the response profile varies between units, in other words, some individuals are intrinsically high responders, while others are low responders. Random effects are defined as the stochastic variation between units expressed in various aspects of their behavior. The units are sampled at random from a population.

2. Serial correlation: Some part of an individual's observed measurement is a response to time-varying stochastic processes working within that individual. This kind of variation results in a correlation between pairs of measurements on the same individual which depends on the time separation between the pair of measurements. As the time separation increases, the correlation typically decreases.

3. Measurement error: The measurement process itself adds a component of variation to the data, and this variation is defined as measurement error.

We use an additive formulation to incorporate the three features into a specific model. First, the mean and covariance structure must be explicitly separated. This is done by writing the model in expression 5.2 as

$$\mathbf{Y} = X\beta + \epsilon.$$

The last term is modeled according to this:

$$\epsilon \sim MVN\left(0, V(\mathbf{t}, \alpha)\right).$$

The element of ϵ that corresponds to the jth measurement of the ith unit is denoted by ϵ_{ij}. Now the ϵ_{ij} is split up additively into the three different sources of variation using this formulation:

$$\epsilon_{ij} = \mathbf{d}_{ij}^T \mathbf{U}_i + W_i(t_{ij}) + Z_{ij}.$$

Here $\mathbf{U}_i$ corresponds to random effects, and it is a set of m mutually independent r-element Gaussian random vectors, each with mean vector zero and covariance matrix labeled G. The term $\mathbf{d}_{ij}$ denotes the r-element vectors of explanatory variables attached to individual measurements. Also $W_i(t_{ij})$ corresponds to serial correlation, and is sampled from m independent copies of a stationary Gaussian process with zero mean, variance σ^2 and correlation function $\rho(u)$. Lastly Z_{ij} corresponds to the measurement error; it is a set of N mutually independent Gaussian random variables, each with zero mean and variance τ^2. [Diggle et al., 2002, p. 82-83].

When this model is applied, it may be useful to use a transformation on the data. For example, if the data have an underlying multiplicative structure, it would be reasonable to use a logarithmic transformation on them, to obtain an additive structure instead.

The vector of random variables, ϵ_{ij}, associated with the ith unit, is written as $\epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{in_i})$. The $n_i \times r$ matrix with jth row $\mathbf{d}_{ij}$ is named D_i. The correlation between $W_i(t_{ij})$ and $W_i(t_{ik})$ is h_{ijk}, and $h_{ijk} = \rho(|t_{ij} - t_{ik}|)$. The $n_i \times n_i$ matrix consisting of (j,k)th elements h_{ijk} is labeled H_i. Now the covariance matrix of ϵ_i is

$$Var(\epsilon_i) = D_i G D_i^T + \sigma^2 H_i + \tau^2 I_i, \tag{5.3}$$

where I_i is the $n_i \times n_i$ identity matrix. Because measurements from different units are independent the subscript i can be dropped. Hereby equation 5.3 becomes

$$Var(\epsilon) = DGD^T + \sigma^2 H + \tau^2 I, \tag{5.4}$$

where $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$ is a non-specific sequence of measurements from one individual. In the following model $\mathbf{t} = (t_1, \ldots, t_n)$ corresponds to the set of times at which the measurements are made. [Diggle et al., 2002, p. 83-84].

Random intercept

The model is called random intercept, because the value of U represent an amount by which *all* measurements on the individual in question are raised or lowered relative to the population average, i.e. U is a random intercept. This is valid, when U is a univariate Gaussian random variable with mean zero, variance ν^2, and $d_j = 1$. By implementing these in equation 5.4 the following variance matrix is obtained:

$$Var(\epsilon) = \nu^2 J + \sigma^2 H + \tau^2 I.$$

Here J is an $n \times n$ matrix of 1's. Since the variance of each $\epsilon_\mathbf{j}$ is $Var(\epsilon_\mathbf{j}) = \nu^2 + \sigma^2 + \tau^2$ and the elements of H are specified by a correlation function $\rho(u)$, the variogram for the model becomes

$$\gamma(u) = \nu^2 + \tau^2 + \sigma^2 \left(1 - \rho(u)\right).$$

Figure 5.1 shows the behavior of the variogram when $\rho(u) = \exp(-u^2)$. For this

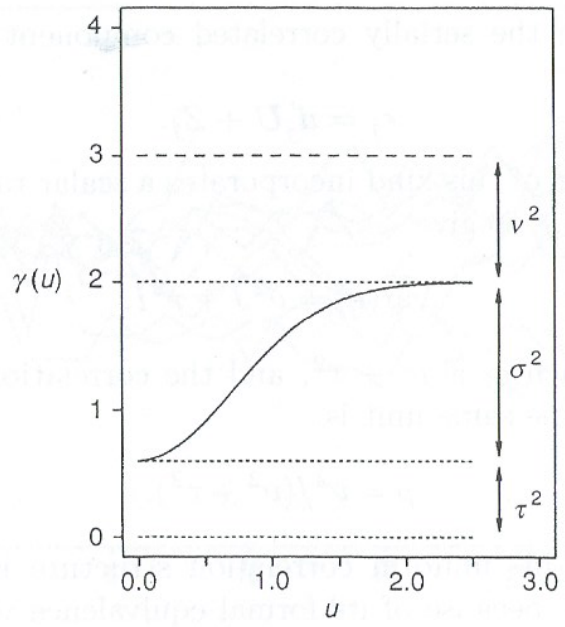

Figure 5.1: Variogram when $\rho(u) = \exp(u^2)$.

specific model the limit of the variogram function as $u \to \infty$ is always less than the variance of ϵ_j:

$$\lim_{u \to \infty} \gamma(u) < Var(\epsilon_j).$$

This indicates that no matter how big a time difference, u, the variogram function does not exceeds the variance of ϵ_j. [Diggle et al., 2002, p. 89-91].

5.2 Model-fitting

The model-fitting process includes the following four steps:

1. Formulation, where the general form of the model is chosen.

2. Estimation, where numerical values are attached to the parameters.

3. Inference, where confidence intervals are calculated and hypothesis about parameters of direct interest are tested.

4. Diagnostics, where we control that the model fits the data.

Formulation

At this stage it is assumed that the identification and treatment of outliers has already been resolved. We start focusing on the mean and covariance structure of the data. Regarding the mean; when the data is well-replicated, it is possible and meaningful to look at time-plots of observed averages. Otherwise it is helpful to use a non-parametric smoothing process on the data. Regarding the covariance structure; by subtracting the OLS estimate of the corresponding mean response from each measurement the residuals are obtained. Here the corresponding mean response is based on the most elaborate model. Looking at time-plots, scatterplot matrices and empirical variogram plots of the residuals will give an idea of the structure of the data. Time-plots with non-stationary variation indicate, that it is necessary to use either a transformation of the data or a non-stationary model like the random effects model just described. If the patterns appears to be stationary, the underlying covariance structure can be estimated using the empirical variogram. All in all, different data causes different parametric models, and it is important to consider that an incorrect or poorly chosen model can result in biased estimates. [Diggle et al., 2002, p. 94-95].

Estimation

The aim of this step is to estimate numerical values for the parameters of interest. The general form of the model is this:

$$\mathbf{Y} \sim MVN\left(X\beta, \sigma^2 V(\alpha)\right). \tag{5.5}$$

Therefore the parameters of interest are β, σ^2 and α. Their estimates are labeled $\hat{\beta}$, $\hat{\sigma}^2$ and $\hat{\alpha}$, respectively. For special cases of the model in expression 5.5 specific estimation methods exist. But in the following a general method is derived using the same strategy as for the maximum likelihood estimates. The only difference is that now we make use of the parametric structure of the variance matrix. Hereby, for a given α equations 4.10 and 4.14 hold in their modified forms:

$$\hat{\beta}(\alpha) = \left(X^T V(\alpha)^{-1} X\right)^{-1} X^T V(\alpha)^{-1} \mathbf{y}, \tag{5.6}$$

and

$$\hat{\sigma}^2(\alpha) = \frac{RSS(\alpha)}{N-p},$$

where

$$RSS(\alpha) = \left(\mathbf{y} - X\hat{\beta}(\alpha)\right)^T V(\alpha)^{-1} \left(\mathbf{y} - X\hat{\beta}(\alpha)\right),$$

and the total number of measurements on all m individuals is $N = \sum_{i=1}^{m} n_i$. The parameter α can be estimated by using REML estimation as defined in section 4.4. In that case the REML estimate for α maximizes the function

$$L^*(\tilde{\alpha}) = -\frac{1}{2}\left(N\log(RSS(\alpha)) + \log\left(|V(\alpha)|\right) + \log\left(\left|X^T V(\alpha)^{-1} X\right|\right)\right),$$

which results in this REML estimate of β: $\hat{\beta} = \hat{\beta}(\alpha)$. The parameter can also be estimated using maximum likelihood estimation, in which case the estimate for α maximizes

$$L(\hat{\alpha}) = -\frac{1}{2}\left(N\log\left(RSS(\alpha)\right) + \sum_{i=1}^{m}\log\left(|V_i(\mathbf{t}_i,\alpha)|\right)\right).$$

[Diggle et al., 2002, p. 95-97]

Inference

Now we wish to make inference about β. The model in expression 5.5 combined with the result from equation 5.6 generates a basic for making such inference:

$$\hat{\beta}(\alpha) \sim MVN\left(\beta, \sigma^2\left(X^T V(\alpha)^{-1} X\right)^{-1}\right).$$

It is assumed that this model continues to hold approximately when the unknown values of σ^2 and α are substituted by the REML estimates $\hat{\sigma}^2$ and $\hat{\alpha}$. Therefore it is written as

$$\hat{\beta} \sim MVN\left(\beta, \hat{V}\right), \tag{5.7}$$

where

$$\hat{V} = \hat{\sigma}^2\left(X^T V(\hat{\alpha})^{-1} X\right)^{-1}.$$

An application of the model in expression 5.7 is to look at the standard errors of individual elements of β and to estimate those. Another application is to look at the general linear transformations of the form

$$\psi = D\beta,$$

and calculate confidence regions for these. Here D is a full-rank matrix with $r \times p$ elements, where $r \leq p$. If $\psi = D\hat{\beta}$, then we have that

$$\hat{\psi} \sim MVN(\psi, D\hat{V}D^T), \tag{5.8}$$

from which it is possible to calculate confidence regions for ψ as described in the following. From the model in expression 5.8 we have that:

$$T(\psi) = \left(\hat{\psi} - \psi\right)^T \left(D\hat{V}D^T\right)^{-1}\left(\hat{\psi} - \psi\right)$$

is approximately distributed as χ^2_r. Now the q-critical value of χ^2_r is denoted by $c_r(q)$, and therefore

$$P\left(\chi^2_r \geq c_r(q)\right) = q.$$

Hence, a $100(1-q)\%$ confidence region for ψ is

$$\left[\psi : T(\psi) \leq c_r(q)\right].$$

The above described method of inference is based on a maximal model for β, from where the parameters σ^2 and α are estimated. [Diggle et al., 2002, p. 97-98].

Diagnostics

The model-fitting process is completed by checking the model against the data. We wish to highlight any systematic discrepancies, which can be done by comparing the data with the fitted model in a certain way. A relatively simple and effective check of the mean is to superimpose the fitted mean response profiles on a time-plot of the average observed one, and likewise for the covariance structure, where the fitted variogram is superimposed on a plot of the empirical variogram. If inconsistencies are revealed, they can be incorporated into a revised model, which then undergoes the fitting process over again. [Diggle et al., 2002, p. 98].

Chapter 6

Generalized linear models

This chapter presents generalized linear models. First, the general class of generalized linear models is introduced followed by a look at Poisson regression. Later, generalized linear models for longitudinal data are described.

The expected value of a given, but unknown quantity can be predicted by means of ordinary linear regression, but this is only appropriate when the response variable has a normal distribution. However, many types of response variables do not follow this distribution. Generalized linear models (GLMs) were formulated as a way of unifying regression models for independent, discrete, and continuous responses. This generalization allows the response variable to have arbitrary distributions, and thereby provides a common statistical methodology for different types of response variables. [Diggle et al., 2002, p. 343].

6.1 The general case

Based on a single response Y_i and a vector of p explanatory variables, $\mathbf{x_i}$, associated with each of m experimental units, the objective is to state the mean response's, $E(Y_i) = \mu_i$, dependency on the explanatory variables. We assume the mean response to be related to covariates named $\mathbf{x}$ through a so called *link function* named h:

$$h(\mu_i) = \mathbf{x}_i'\boldsymbol{\beta}. \tag{6.1}$$

The variance of Y_i is expressed with a $variancefunction$. This is known as v, and it is specified by the mean μ_i;

$$Var(Y_i) = v_i = \phi v(\mu_i),$$

where ϕ is a scaling factor. For some members of the GLM family, ϕ is a known constant, while for others it is a parameter that needs to be estimated. [Diggle et al., 2002, p. 345].

The exponential family of distributions include the Poisson distribution, the Gaussian distribution, the binomial distribution, and the two-parameter gamma distribution. With a likelihood function of this form

$$f(y_i) = exp\left(\frac{y_i\theta_i - \psi(\theta_i)}{\phi} + c(y_i, \phi)\right), \tag{6.2}$$

each member of the exponential family of distributions corresponds to a class of GLMs. Now we take a closer look at the mean, $E[Y] = \mu_i$, in order to determine the relation between this and the *natural parameter*, labeled θ_i. According to the definition of the mean, [Olofsson, 2005, p. 100], we know that

$$E[Y] = \int yf(y)dy$$

in the continuous case. Furthermore we know [Olofsson, 2005, p. 88], that

$$\int f(y)dy = 1. \tag{6.3}$$

By replacing $f(y)$ in equation 6.3 by the likelihood function from equation 6.2 we obtain this expression:

$$1 = \int exp\left(\frac{y_i\theta_i - \psi(\theta_i)}{\phi} + c(y_i, \phi)\right)dy. \tag{6.4}$$

The term that does not involve y is taken out of the integral, and we get

$$exp(\frac{\psi(\theta_i)}{\phi}) = \int exp\left(\frac{y_i\theta_i}{\phi} + c(y_i, \phi)\right)dy.$$

First the logarithm is taken on both sides:

$$\frac{\psi(\theta_i)}{\phi} = log\left(\int exp\left(\frac{y_i\theta_i}{\phi} + c(y_i, \phi)\right)dy\right),$$

and then the equation is differentiated with respect to θ_i:

$$\frac{1}{\phi}\frac{\partial\psi(\theta_i)}{\partial\theta_i} = \frac{\int exp\left(\frac{y_i\theta_i}{\phi} + c(y_i, \phi)\right)\frac{y_i}{\phi}dy}{\int exp\left(\frac{y_i\theta_i}{\phi} + c(y_i, \phi)\right)dy}. \tag{6.5}$$

Now, by equation 6.4, we see that

$$\int exp\left(\frac{y_i\theta_i}{\phi} + c(y_i, \phi)\right)dy = exp\left(\frac{\psi(\theta_i)}{\phi}\right),$$

and hereby equation 6.5 reduces to:

$$\begin{aligned}
\frac{\partial\psi(\theta_i)}{\partial\theta_i} &= \frac{\int exp\left(\frac{y_i\theta_i}{\phi} + c(y_i, \phi)\right)y_i dy}{exp\left(\frac{\psi(\theta_i)}{\phi}\right)} \\
&= \int exp\left(\frac{y_i\theta_i}{\phi} - \frac{\psi(\theta_i)}{\phi} + c(y_i, \phi)\right)y_i dy \\
&= \int exp\left(\frac{y_i\theta_i - \psi(\theta_i)}{\phi} + c(y_i, \phi)\right)y_i dy.
\end{aligned}$$

Because of the likelihood function in equation 6.2, this can be further reduced:

$$\frac{\partial \psi(\theta_i)}{\partial \theta_i} = \int y_i f(y_i) dy.$$

Since the term on the right side exactly is the mean, we conclude that

$$\frac{\partial \psi(\theta_i)}{\partial \theta_i} = \mu_i.$$

The same correspondence can be shown for the discrete case, just by replacing the integral with the summation.

For each of the GLMs the vector of regression coefficients, β, can be estimated by solving the following equation:

$$\mathbf{S}(\beta) = \sum_{i=1}^{m} \left(\frac{\partial \mu_i}{\partial \beta} \right)' \upsilon_i^{-1} \left(Y_i - \mu_i(\beta) \right) = 0. \tag{6.6}$$

Here $\upsilon_i = Var(Y_i)$. The solution to equation 6.6 is the maximum likelihood estimate, and it is named $\hat{\beta}$. It is obtained by iteratively reweighted least squares. [Diggle et al., 2002, p. 345].

For large data sets, the estimate $\hat{\beta}$ follows approximately a Gaussian distribution, where

$$E(\hat{\beta}) = \beta,$$

and

$$V = \left(\sum_{i=1}^{m} \left(\frac{\partial \mu_i}{\partial \beta} \right)' \upsilon_i^{-1} \frac{\partial \mu_i}{\partial \beta} \right)^{-1}. \tag{6.7}$$

By replacing β with $\hat{\beta}$ in equation 6.7 an estimate for the variance, $\hat{V}$, can be found. [Diggle et al., 2002, p. 346].

6.2 The binomial case

This model is also called logistic regression, and it is used for binary variables as for example the absence or presence of the use of medication. The binomial regression assumes a linear relation between the logarithm of the odds and the explanatory variables. This relation is expressed mathematically in the following equation:

$$log \left(\frac{Pr(Y_i = 1)}{Pr(Y_i = 0)} \right) = \mathbf{x}_i' \beta.$$

Here the regression coefficients, β, represent the change of the log odds of the dependent variable per unit change of x. Since $Pr(Y_i = 1) = \mu_i$, we have that

$$log \left(\frac{Pr(Y_i = 1)}{Pr(Y_i = 0)} \right) = log \left(\frac{\mu_i}{1 - \mu_i} \right),$$

and then, by equation 6.1, the link function, h, for logistic regression is

$$h(\mu_i) = log\left(\frac{\mu_i}{1-\mu_i}\right).$$

Furthermore, the binomial case presents the property that the variance is absolute determined by the mean, which is expressed here:

$$Var(Y_i) = E(Y_i) \cdot (1 - E(Yi)) = \frac{exp(\mathbf{x}_i'\beta)}{(1 + exp(\mathbf{x}_i'\beta))^2}.$$

[Diggle et al., 2002, p. 343].

6.3 The Poisson case

Poisson regression is appropriate when exploring count data. They are also called log-linear models, and they are useful in studies where the response variable represents the number of events occurring in a given period of time. Since the number of events is either zero or a positive integer, the nature of count data is discrete and non-negative. Therefore, it can be assumed that the logarithm of the expected response is a linear function of explanatory variables, which is expressed here:

$$log\left(E(Y_i)\right) = \mathbf{x_i'}\beta. \tag{6.8}$$

According to equation 6.1 this means, that the link function, h, for Poisson regression is:

$$h\left(\mu_i\right) = log\left(\mu_i\right).$$

The interpretation of the regression coefficient for a specific explanatory variable could be this: every time the given explanatory variable increases one unit, while all the other explanatory variables are held constant, the expected counts before and after are noted. Then the logarithm of the ratio between these before and after values is the same as the regression coefficient.

The Poisson regression is named after the Poisson distribution for counts. The distribution has this probability function:

$$p(y) = exp(-\mu)\frac{\mu^y}{y!},$$

where $y = 0, 1, \ldots$ The variance of a Poisson distribution equals its mean. We see from equation 6.8, that the mean $E(Y_i) = exp(\mathbf{x_i'}\beta)$, and hence

$$Var(Y_i) = E(Y_i) = exp(\mathbf{x_i'}\beta).$$

The Poisson models have a likelihood function on the form of expression 6.2, where the parameters are as follows:

$$\theta_i = log(\mu_i), \psi(\theta_i) = exp(\theta_i), c(y_i, \phi) = -log(y_i!), \phi = 1.$$

[Diggle et al., 2002, p. 344].

6.4 GLMs for longitudinal data

In this section we will introduce one of three different extensions of generalized linear models for longitudinal data. The three models are called marginal models, random effects models, and transition models. Since random effects models are best suited to this project, we will leave out a further introduction of marginal models and transition models. For the random effects models the underlying idea and domain of application will be presented, followed by a short introduction of the likelihood function.

In a linear random effects model, the response is assumed to be a linear function of explanatory variables, where the regression coefficients might vary from person to person. This variability from one individual to the next reflects the natural heterogeneity due to unmeasured factors. Furthermore it is assumed that repeated observations for one person are dependent. The correlation among repeated observations arises because it is not possible to observe the true regression coefficients. We have only imperfect measurements available. Assuming that the data for a subject are conditionally independent, observations following a GLM extend the idea of regression models for discrete and non-Gaussian continuous responses. In this case the regression coefficients can vary from person to person according to a distribution, F. [Diggle et al., 2002, p. 128-129].

Now consider a logistic model for an Indonesian children's health study looking at the dependence of respiratory infection on vitamin A deficiency. Let x_{ij} indicate whether or not child i is vitamin A deficient (1: yes; 0: no) at visit j, and let Y_{ij} denote whether the child has respiratory infection (1: yes; 0: no). It is plausible to assume that the propensity for respiratory infection varying among children according to their different genetic predispositions and unmeasured influences of environmental factors. A simple model would assume that the effect of vitamin A deficiency on the probability for a respiratory disease is the same for every child, but each child still has its own propensity for respiratory infection. Such a model would take this form:

$$logit \Pr(Y_{ij} = 1|U_i) = (\beta_0^* + U_i) + \beta_1^* x_{ij},$$

where the U_i are the random effects. Here x_{ij} is 1 when child number i is vitamin A deficient at visit j, and 0 if the child is not vitamin A deficient. It is assumed, that given U_i, the repeated observations for each child are independent of one another. The model requires that we make assumptions about the distribution of the U_i. Like each child has it own set of measurements, it also has a value for U_i, and a typical used parametric model to describe the distribution of these is the Gaussian with mean zero and unknown variance, v^2.

For a typical child with random effect $U_i = 0$, the parameter β_0^* in the model is the log-odds of respiratory infection. The log-odds ratio for respiratory infection

when a child is vitamin A deficient relative to when that same child is not is denoted by the parameter β_1^*. Lastly the degree of heterogeneity across children in the propensity for infection, not caused by x, is represented by the variance v^2.

The specifications for a general random effects generalized linear model are:

" 1. Given $\mathbf{U}_i$, the responses $Y_{i1}, \ldots, Y_{in_i}$ are mutually independent and follow a GLM with density

$$f(y_{ij}|\mathbf{U}_i) = \exp\left(\frac{(y_{ij}\theta_{ij} - \psi(\theta_{ij}))}{\phi} + c(y_{ij}, \phi)\right).$$

The conditional moments,

$$\mu_{ij} = E(Y_{ij}|\mathbf{U}_i) = \psi'(\theta_{ij}) \qquad \text{and} \qquad V_{ij} = Var(Y_{ij}|\mathbf{U}_i)\psi''(\theta_{ij})\phi,$$

satisfy

$$h(\mu_{ij}) = \mathbf{x}'_{ij}\beta^* + \mathbf{d}'_{ij}\mathbf{U}_i \qquad \text{and} \qquad v_{ij} = v(\mu_{ij})\phi,$$

where h and v are known link and variance functions, respectively, and $\mathbf{d}_{ij}$ is a subset of $\mathbf{x}_{ij}$.

2. The random effects, $\mathbf{U}_i$, $i = 1, \ldots, m$, are mutually independent with a common underlying multivariate distribution, F. " [Diggle et al., 2002, p. 129].

Now it is clear that the basic idea behind random effects models is that there is a natural heterogeneity across the subjects, and that this is represented by a probability distribution. Furthermore, from their sharing unmeasurable variables, $\mathbf{U}_i$, a correlation among observations for one subject arises. Sometimes a model like this is called a *latent variable* model. In this model the effects of the explanatory variables on an individual are represented by the regression coefficients, β^*. The model is definitely most useful when the aim is to make inference about individuals rather than about the population average, which exactly matches the aim of this project. [Diggle et al., 2002, p. 127-130].

It is of course interesting to estimate the unknown parameters of the model which in this model is also possible using the traditional maximum likelihood estimation. Expressed as a function of the unknown parameters, the likelihood of the data is given by

$$L(\beta^*, \alpha; \mathbf{y}) = \prod_{i=1}^{m} \int \prod_{j=1}^{n_i} f(y_{ij}|\mathbf{U_i}) f(\mathbf{U_i}; \alpha) d\mathbf{U}_i.$$

Here α denotes the parameters of the random effects distribution. The term $f(y_{ij}|\mathbf{U_i})$ is the joint distribution of the unobserved random effects of the data, and $f(\mathbf{U_i}; \alpha)$ denotes the random effects. The integral has a closed form when the data are Gaussian. In this case a method for maximizing the likelihood or the restricted likelihood exist. But when the data are non-linear, it is often necessary to use numerical integration techniques in order to evaluate the likelihood. [Diggle et al., 2002, p. 137-138].

Chapter 7

Random effects models

In a random effects model the regression coefficients measure the influence of explanatory variables on the responses for heterogeneous subjects. The basic property for this model is that there is a natural heterogeneity among individuals in a subset of the regression coefficient. An example of this is the intercepts. When working in the framework of GLM, it is assumed that on condition of unobservable variables $\mathbf{U}_i$ we have independent responses from a distribution in the exponential family expressed in equation 6.2. This means that the link function is

$$h\left(E\left(Y_{ij}|\mathbf{U}_i\right)\right) = \mathbf{x}_{ij}^T\beta + \mathbf{d}_{ij}^T\mathbf{U}_i.$$

Here $\mathbf{x}_{ij}$ is a covariate vector of length p, while $\mathbf{d}_{ij}$ is a covariate vector of length q. Two different methods of making inference about random effects models exist. Only one of them is suitable when the subject-specific coefficients themselves are of interest, which is the case in this project. Therefore that approach is at focus here. The strategy is to operate as if the $\mathbf{U}_i$ are an independent sample from some distribution. Under this assumption longitudinal and cross-sectional information is combined. And then both β, the fixed effects, and $\mathbf{U}_i$, the random effects, are estimated under this working model. The analysis should weight the longitudinal information more heavily, when there is a large variability across subject. This is because comparisons among subjects are not as precise as comparisons within subjects. One of the fundamental assumptions of a random effects model is that the random effects, $\mathbf{U}_i$, are independent of the explanatory variables. This assumption can be checked using a so-called specification test. [Diggle et al., 2002, p. 169-171]

In the following section estimation of β in the random effects GLM is described. Both conditional and full maximum likelihood estimation are used. Later in the chapter we look at the specific cases of logistic models for binary data and for count data Poisson regression models.

7.1 Estimation for generalized linear mixed models

In the random effects GLM the following assumptions are made:

" 1. the conditional distribution of Y_{ij} given $\mathbf{U}_i$ follows a distribution from the exponential family with density $f(y_{ij}|\mathbf{U}_i);\beta$.

2. given $\mathbf{U}_i$, the repeated measurements, $Y_{i1},\ldots,Y_{in_i}$, are independent.

3. the $\mathbf{U}_i$ are independent and identically distributed with density function $f(\mathbf{U}_i;G)$.
" [Diggle et al., 2002, p. 171]

As usual $i=1,\ldots,m$. In the case with conditional likelihood the random effects are treated as if they where fixed parameters to be removed from the problem, so the third assumption above is actually not necessary. But in the case with maximum likelihood estimation, $\mathbf{U}=(U_1,\ldots,U_m)$ is treated as a set of unobserved variables, which is then integrated out of the likelihood. Hereby it is assumed that the random effects follows a Gaussian distribution with mean zero and variance matrix G. [Diggle et al., 2002, p. 171]

Conditional likelihood estimation
When estimating β using conditional maximum likelihood, the basic idea is to see the random effects as a set of nuisance parameters. Because then β can be estimated with the conditional likelihood of the data given the sufficient statistics for the $\mathbf{U}_i$. If $\theta_{ij}=\theta_{ij}(\beta,\mathbf{U})$, the likelihood function for β and $\mathbf{U}$ is:

$$\prod_{i=1}^{m}\prod_{j=1}^{n_i} f\left(y_{ij}|\beta,\mathbf{U}_i\right) \propto \prod_{i=1}^{m}\prod_{j=1}^{n_i} \exp\left(\theta_{ij}y_{ij}-\psi(\theta_{ij})\right).$$

In order to simplify the focus is now restricted to the canonical link function. Here $\theta_{ij}=\mathbf{x}_{ij}^T\beta+\mathbf{d}_{ij}^T\mathbf{U}_i)$, so the likelihood above is rewritten into

$$\exp\left(\beta^T\sum_{i,j}\mathbf{x}_{ij}y_{ij}+\sum_i\mathbf{U}_i^T\sum_j\mathbf{d}_{ij}y_{ij}-\sum_{i,j}\psi(\theta_{ij}).\right).$$

From this we see that $\sum_{i,j}\mathbf{x}_{ij}y_{ij}$ is a sufficient statistic for β, and that $\sum_j\mathbf{d}_{ij}y_{ij}$ is a sufficient statistic for $\mathbf{U}_i$. Since the conditional likelihood is proportional to the conditional distribution of the data given the sufficient statistics for the $\mathbf{U}_i$, we have the following contribution from subject i:

$$\begin{aligned} f\left(\mathbf{y}_i|\sum_j\mathbf{d}_{ij}y_{ij}=\mathbf{b}_i;\beta\right) &= \frac{f\left(\mathbf{y}_i;\beta,\mathbf{U}_i\right)}{f\left(\sum_j\mathbf{d}_{ij}y_{ij}=\mathbf{b}_i;\beta,\mathbf{U}_i\right)} \\ &\propto \frac{f\left(\sum_j\mathbf{x}_{ij}y_{ij}=\mathbf{a}_i,\sum_j\mathbf{d}_{ij}y_{ij}=\mathbf{b}_i;\beta,\mathbf{U}_i\right)}{f\left(\sum_j\mathbf{d}_{ij}y_{ij}=\mathbf{b}_i;\beta,\mathbf{U}_i\right)}. \end{aligned}$$

If the GLM is discrete, the expression above can be rewritten into:

$$\frac{\sum_{R_{i1}} \exp\left(\beta^T \mathbf{a}_i + \mathbf{U}_i^T \mathbf{b}_i\right)}{\sum_{R_{i2}} \exp\left(\beta^T \sum_j \mathbf{x}_{ij} y_{ij} + \mathbf{U}_i^T \mathbf{b}_i\right)}.$$

Here R_{i1} denotes the set of possible values for $\mathbf{y}_i$ in which $\sum_j \mathbf{x}_{ij} y_{ij} = \mathbf{a}_i$, and R_{i2} is the set of values for $\mathbf{y}_i$ such that $\sum_j \mathbf{d}_{ij} y_{ij} = \mathbf{b}_i$. For β the conditional likelihood given the data for all m subjects becomes:

$$\prod_{i=1}^{m} \frac{\sum_{R_{i1}} \exp\left(\beta^T \mathbf{a}_i\right)}{\sum_{R_{i2}} \exp\left(\beta^T \sum_j \mathbf{x}_{ij} y_{ij}\right)}. \tag{7.1}$$

[Diggle et al., 2002, p. 171-172]

Maximum likelihood estimation

When estimating β using this method, the assumption about the random effects allows us to learn about one individual's coefficient by understanding the variability in coefficients across the whole population. If there is little variability, the estimate for an individual should be based on the average of the population, but if there is substantial variation, we should rely more on the data from each individual, when estimating their coefficients. The unknown parameter δ is defined to include both β and the elements of G. The likelihood function for this parameter δ is

$$L(\delta; \mathbf{y}) = \prod_{i=1}^{m} \int \prod_{j=1}^{n_i} f(y_{ij}|\mathbf{U}_i; \beta) f(\mathbf{U}_i; G) \, d\mathbf{U}_i.$$

This is actually the marginal distribution of $\mathbf{Y}$, which is found by integrating the joint distribution of $\mathbf{Y}$ and $\mathbf{U}$ with respect to $\mathbf{U}$. The maximum likelihood estimate for δ can be found solving the score equations. These are obtained by taking the derivative with respect to δ of the log likelihood and setting it equal to zero. Under the assumption that the complete data for an individual is represented by $(\mathbf{y}_i, \mathbf{U}_i)$, and if the focus is restricted to the canonical link function, we have the complete data score function for β:

$$\mathbf{S}_\beta(\delta|\mathbf{y}, \mathbf{U}) = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \mathbf{x}_{ij} (y_{ij} - \mu_{ij}(\mathbf{U}_i)) = 0,$$

where

$$\mu_{ij}(\mathbf{U}_i) = E(y_{ij}|\mathbf{U}_i) = h^{-1}\left(\mathbf{x}_{ij}^T \beta + \mathbf{d}_{ij}^T \mathbf{U}_i\right).$$

The score equations for the observed data are found by taking the expectation of the complete data equations with respect to the conditional distribution of the unobserved random effects given the data. The observed data score functions are labeled $\mathbf{S}_\beta(\delta|\mathbf{y})$, and it is defined as the expectations of $\mathbf{S}_\beta(\delta|\mathbf{y}, \mathbf{U})$, which is the complete

data score functions, with respect to the conditional distribution of $\mathbf{U}$ given $\mathbf{y}$. Thus we have:

$$\mathbf{S}_\beta(\delta|\mathbf{y}) = \sum_{i=1}^{m}\sum_{j=1}^{n_i}\mathbf{x}_{ij}\left(y_{ij} - E\left(\mu_{ij}(\mathbf{U}_i)|\mathbf{y}_i\right)\right) = 0.$$

Similarly the score equations for G are obtained as:

$$\mathbf{S}_G(\delta|\mathbf{y}) = \frac{1}{2}G^{-1}\left(\sum_{i=1}^{m}E(\mathbf{U}_i\mathbf{U}_i^T|\mathbf{y}_i)\right)G^{-1} - \frac{m}{2}G^{-1} = 0.$$

[Diggle et al., 2002, p. 172-173]
When there is no restrictions on G, it can be calculated like this:

$$\begin{aligned}
\frac{m}{2}G^{-1} &= \frac{1}{2}G^{-1}\left(\sum_{i=1}^{m}E(\mathbf{U}_i\mathbf{U}_i^T|\mathbf{y}_i)\right)G^{-1} \\
&\Updownarrow \\
mG &= \sum_{i=1}^{m}E(\mathbf{U}_i\mathbf{U}_i^T|\mathbf{y}_i) \\
&\Updownarrow \\
G &= \frac{1}{m}\sum_{i=1}^{m}E(\mathbf{U}_i\mathbf{U}_i^T|\mathbf{y}_i).
\end{aligned}$$

Here the remaining question is, how the $\mathbf{U}_i$ can be calculated.

7.2 Random effects model for binary data

Again we distinguish between the conditional likelihood approach and the maximum likelihood approach. In this section the conditional likelihood approach for binary responses is described first, followed by a review of the maximum likelihood method.

Conditional likelihood estimation
The random intercept logistic model is given by

$$\operatorname{logit}\Pr(Y_{ij} = 1|U_i) = \beta_0 + U_i + \mathbf{x}_{ij}^T\beta. \tag{7.2}$$

In order to simplify it is assumed that $\mathbf{x}_{ij}$ does not include an intercept term. Further more $\gamma_i = \beta_0 + U_i$. For β and γ_i the joint likelihood function is proportional to this expression:

$$\prod_{i=1}^{m}\exp\left(\gamma_i\sum_{j=1}^{n_i}y_{ij} + \left(\sum_{j=1}^{n_i}y_{ij}\mathbf{x}_{ij}^T\right)\beta - \sum_{j=1}^{n_i}\log\left(1+\exp(\gamma_i + \mathbf{x}_{ij}^T\beta)\right)\right).$$

Now given the sufficient statistics for the γ_i the conditional likelihood for β has the form as the expression:

$$\prod_{i=1}^{m} \frac{\exp\left(\sum_{j=1}^{n_i} y_{ij}\mathbf{x}_{ij}^T\beta\right)}{\sum_{a\subseteq\{1,\ldots,n_i\}} \exp\left(\sum_{l\in a}\mathbf{x}_{il}\beta\right)}.$$

Here $|a| = y_{i.}$. Even though this holds in theory, that the calculations can only be performed in practice, when $y_{i.}$ and n_i is rather small. For example if $n_i = 2$ and $y_{i.} = 4$. Otherwise there is way too many combinations of a, even for your excellent computers today. [Diggle et al., 2002, p. 175]

Now as an example a 2×2 crossover trial is considered. The first group is called AB, and it contains the subjects who received the active treatment (A) first and then placebo (B). In the second group, which is named BA, we have the subjects who received placebo (B) followed by active treatment (A). There is four different outcomes for each; $(1,1)$, $(0,1)$, $(1,0)$ and $(0,0)$. Here $Y = 1$ denotes a normal response, while $Y = 0$ is an abnormal response. Each subject has two visits. If a, b, c and d correspond to the numbers of response pair for each possible combination of outcomes, we can write the information as in table 7.2. From this table it is clear that c_1 is the number of subjects in the first group with normal response at the first visit and an abnormal response at the second one.

Group	$(1,1)$	$(0,1)$	$(1,0)$	$(0,0)$
AB	a_1	b_1	c_1	d_1
BA	a_2	b_2	c_2	d_2

Consider the logistic model in equation 7.2 including only the treatment variable x_1. For this model the conditional likelihood from equation 7.1 becomes:

$$\left\{\frac{\exp(\beta_1)}{1+\exp(\beta_1)}\right\}^{b_1+b_2}\left\{\frac{1}{1+\exp(\beta_1)}\right\}^{c_1+c_2}.$$

Here the estimate for β_1 that maximizes the conditional likelihood is

$$\hat{\beta}_1 = log\left(\frac{b_1+b_2}{c_1+c_2}\right),$$

and the variance estimate is

$$\hat{Var}(\hat{\beta}_1) = (b_1+b_2)^{-1}(c_1+c_2)^{-1}.$$

When the period effect, x_2, is added to the model, the conditional likelihood function becomes

$$\frac{\exp((\beta_1+\beta_2)b_1)}{(1+\exp(\beta_1+\beta_2))^{b_1+c_1}}\frac{\exp((\beta_1+\beta_2)b_2)}{(1+\exp(\beta_1+\beta_2))^{b_2+c_2}}.$$

Here the maximum conditional likelihood estimate of β_1 is

$$\hat{\beta}_1 = \frac{1}{2} log\left(\frac{b_1 b_2}{c_1 c_2}\right),$$

while the corresponding variance estimate is

$$\hat{Var}(\hat{\beta}_1) = \frac{1}{4}(b_1^{-1} + c_1^{-1} + b_2^{-1} + c_2^{-1}).$$

One of the advantages of this conditional likelihood approach is that the random effects are removed from the likelihood used to estimate β. This means that the assumption that the random effects are sampled from a particular probability function is avoided. On the other hand we rely entirely on within-subject comparisons, which may be a disadvantage. [Diggle et al., 2002, p. 176-177]

Maximum likelihood estimation
The motivation for random effects models arises naturally when working with binary data. More precisely the variability among clustered binary responses, and the observation that this variability exceeds what is expected due to binomial variation alone. The first random effects model is called the beta-binomial distribution. In the following the term cluster translate to an individual in a longitudinal study, but in the context of genetic studies it could refer to a family or household. We let the n_i binary responses from cluster i be denoted by $\{Y_{i1}, \ldots, Y_{in_i}\}$. The beta-binomial assumes first of all that $Y_{i1}, \ldots, Y_{in_i}$, conditional on μ_i, are independent with common probability μ_i, and secondly that the μ_i follow a beta distribution with mean μ and variance $\delta\mu(1-\mu)$. The total number of positive responses for a cluster has, unconditionally, a beta-binomial distribution, where

$$E(Y_{i.}) = n_i \mu_i$$

and

$$Var(Y_{i.}) = n_i \mu_i (1-\mu_i)(1 + (n_i - 1)\delta).$$

The parameter δ is either called the correlation coefficient or the over-dispersion parameter and it represents the correlation for each pair of binary responses from the same cluster. At first it was thought, that this parameter needed to be positive, but later it was pointed out, that in the beta-binomial model δ has the following lower bound:

$$\delta_0 = max\left\{\frac{-\mu}{n-\mu-1}, \frac{-(1-\mu)}{n+\mu}\right\}.$$

Later on the beta-binomial model was extended to allow variation of the covariates within clusters.

Now we focus on the case, where the random effects are Gaussian. The logistic

model has the following likelihood function for β and G:

$$\begin{aligned} L(\beta, G; \mathbf{y}) &= \prod_{i=1}^{m} \int \prod_{j=1}^{n_i} \left(\mu_{ij}(\beta, \mathbf{U}_i)\right)^{y_{ij}} \left(1 - \mu_{ij}(\beta, \mathbf{U}_i)\right)^{1-y_{ij}} f(\mathbf{U}_i; G) d\mathbf{U}_i \\ &= \prod_{i=1}^{m} \int \exp\left(\beta^T \sum_{j=1}^{n_i} \mathbf{x}_{ij} y_{ij} + \mathbf{U}_i^T \sum_{j=1}^{n_i} \mathbf{d}_{ij} y_{ij} - \sum_j \log\left(1 + \exp(\mathbf{x}_{ij}^T \beta + \mathbf{d}_{ij}^T \mathbf{U}_i)\right)\right) \\ &\quad \times (2\pi)^{-\frac{q}{2}} |G|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} \mathbf{U}_i^T G^{-1} \mathbf{U}_i\right) d\mathbf{U}_i. \end{aligned}$$

In the first line $\mu_{ij}(\beta, \mathbf{U}_i) = E(Y_{ij}|\mathbf{U}_i; \beta)$. The reduction is possible because of the logit link and the Gaussian assumption on the random effects. In the reduced expression G is a variance matrix of each $\mathbf{U}_i$ with the dimensions $q \times q$. When $\mathbf{U}_i$ is integrated out, the contribution from subject i to the likelihood is obtained. So to take a closer look at this contribution, we need to determine the integral of

$$(2\pi)^{-\frac{q}{2}} |G|^{-\frac{1}{2}} \exp\left(\frac{1}{2} \mathbf{U}_i^T G^{-1} \mathbf{U}_i\right) \prod_{j=1}^{n_i} \left(1 + \exp(\mathbf{x}_{ij}^T \beta + \mathbf{d}_{ij}^T \mathbf{U}_i)\right).$$

with respect to $\mathbf{U}_i$. The last term can be rewritten as

$$\begin{aligned} \prod_{j=1}^{n_i} \left(1 + \exp(\mathbf{x}_{ij}^T \beta + \mathbf{d}_{ij}^T \mathbf{U}_i)\right) &= \sum_{a \subseteq \{1,\ldots,n_i\}} \prod_{j \in a} \exp(\mathbf{x}_{ij}^T \beta + \mathbf{d}_{ij}^T \mathbf{U}_i) \\ &= \sum_{a \subseteq \{1,\ldots,n_i\}} \exp(\mathbf{x}_{ia}^T \beta + \mathbf{d}_{ia}^T \mathbf{U}_i), \end{aligned}$$

where $\mathbf{x}_{ia} = \sum_{j \in a} \mathbf{x}_{ij}$ and $\mathbf{d}_{ia} = \sum_{j \in a} \mathbf{d}_{ij}$. This means that we need integration with respect to $\mathbf{U}_i$ of

$$\sum_{a \subseteq \{1,\ldots,n_i\}} (2\pi)^{-\frac{q}{2}} |G|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} \mathbf{U}_i^T G^{-1} \mathbf{U}_i + \mathbf{x}_{ia}^T \beta + \mathbf{d}_{ia}^T \mathbf{U}_i\right).$$

We wish to rewrite the expression as a Gaussian. Inside the exp-bracket we let the term $\mathbf{x}_{ia}^T \beta$ stand and take a closer look at

$$-\frac{1}{2} \mathbf{U}_i^T G^{-1} \mathbf{U}_i + \mathbf{d}_{ia}^T \mathbf{U}_i. \tag{7.3}$$

We wish to decompose the expression in equation 7.3, so it becomes:

$$-\frac{1}{2} \left(\mathbf{U}_i - G\mathbf{d}_{ia}\right)^T G^{-1} \left(\mathbf{U}_i - G\mathbf{d}_{ia}\right).$$

This expression is rewritten first by extending the transposition followed by multiplication of the brackets:

$$\begin{aligned} -\frac{1}{2} & \left(\mathbf{U}_i - G\mathbf{d}_{ia}\right)^T G^{-1} \left(\mathbf{U}_i - G\mathbf{d}_{ia}\right) \\ = & -\frac{1}{2}\left(\mathbf{U}_i^T - \mathbf{d}_{ia}^T G^T\right) G^{-1} \left(\mathbf{U}_i - G\mathbf{d}_{ia}\right) \\ = & -\frac{1}{2}\left(\mathbf{U}_i^T G^{-1}\mathbf{U}_i - \mathbf{U}_i^T G^{-1} G\mathbf{d}_{ia} - \mathbf{d}_{ia}^T G^T G^{-1}\mathbf{U}_i + \mathbf{d}_{ia}^T G^T G^{-1} G\mathbf{d}_{ia}\right) \\ = & -\frac{1}{2}\mathbf{U}_i^T G^{-1}\mathbf{U}_i + \frac{1}{2}\mathbf{U}_i^T\mathbf{d}_{ia} + \frac{1}{2}\mathbf{d}_{ia}^T\mathbf{U}_i - \frac{1}{2}\mathbf{d}_{ia}^T G\mathbf{d}_{ia}. \end{aligned}$$

Since $\frac{1}{2}\mathbf{U}_i^T\mathbf{d}_{ia}$ and $\frac{1}{2}\mathbf{d}_{ia}^T\mathbf{U}_i$ are scalars, they can be added and become one $\mathbf{d}_{ia}^T\mathbf{U}_i$. Hence we have that

$$-\frac{1}{2}\left(\mathbf{U}_i - G\mathbf{d}_{ia}\right)^T G^{-1}\left(\mathbf{U}_i - G\mathbf{d}_{ia}\right) = -\frac{1}{2}\mathbf{U}_i^T G^{-1}\mathbf{U}_i + \mathbf{d}_{ia}^T\mathbf{U}_i - \frac{1}{2}\mathbf{d}_{ia}^T G\mathbf{d}_{ia}. \quad (7.4)$$

So in order to make the right side of equation 7.4 equal to the expression in equation 7.3, we need to add $\frac{1}{2}\mathbf{d}_{ia}^T G\mathbf{d}_{ia}$ to the right side of equation 7.4. Thus we have that

$$-\frac{1}{2}\mathbf{U}_i^T G^{-1}\mathbf{U}_i + \mathbf{d}_{ia}^T\mathbf{U}_i = -\frac{1}{2}\mathbf{U}_i^T G^{-1}\mathbf{U}_i + \mathbf{d}_{ia}^T\mathbf{U}_i - \frac{1}{2}\mathbf{d}_{ia}^T G\mathbf{d}_{ia} + \frac{1}{2}\mathbf{d}_{ia}^T G\mathbf{d}_{ia}.$$

According to equation 7.4 this means that

$$\begin{aligned} -\frac{1}{2} & \mathbf{U}_i^T G^{-1}\mathbf{U}_i + \mathbf{d}_{ia}^T\mathbf{U}_i - \frac{1}{2}\mathbf{d}_{ia}^T G\mathbf{d}_{ia} + \frac{1}{2}\mathbf{d}_{ia}^T G\mathbf{d}_{ia} \\ = & -\frac{1}{2}\left(\mathbf{U}_i - G\mathbf{d}_{ia}\right)^T G^{-1}\left(\mathbf{U}_i - G\mathbf{d}_{ia}\right) + \frac{1}{2}\mathbf{d}_{ia}^T G\mathbf{d}_{ia}. \end{aligned}$$

Now the contribution from subject i before integrating out $\mathbf{U}_i$ is proportional to

$$\sum_{a\subseteq\{1,\ldots,n_i\}} (2\pi)^{-\frac{q}{2}}|G|^{-\frac{1}{2}} \exp\left(\mathbf{x}_{ia}^T\beta - \frac{1}{2}\left(\mathbf{U}_i - G\mathbf{d}_{ia}\right)^T G^{-1}\left(\mathbf{U}_i - G\mathbf{d}_{ia}\right) + \frac{1}{2}\mathbf{d}_{ia}^T G\mathbf{d}_{ia}\right),$$

which tells us that it is a mixture of $N(G\mathbf{d}_{ia}, G)$ with weights $\exp(\mathbf{x}_{ia}^T\beta + \frac{1}{2}\mathbf{d}_{ia}^T G\mathbf{d}_{ia})$. Integrating out $\mathbf{U}_i$ then yields the sum of the weights. So if $n_i = 2$ it is doable, but if fx $n_i \geq 10$, then the sum is prohibitive. [Diggle et al., 2002, p. 178-180]

In relation to the data in this project, the models for binary responses are relevant in order to analyze if the patients er so-called high-dosage-users or not. WHO specifies a defined daily dosage (DDD) for each drug, and in the data we can see how many of these each redeemed prescription contains. We are working under the assumption that up till 12 weeks treatment a year is acceptable. This reduces to 3 weeks per quarter, so more than 21 DDD's per quarter defines a high-dosage-user. For all the patients we can make a series of 0's and 1's according to whether or not they are high-dosage-users each quarter in the period of the study, and these strings can be analyzed with the models described in this section.

7.3 Random effects model for count data

Once again we start with a description of the conditional likelihood method, followed by a review of the random effects model for count data.

Conditional likelihood method
Here the random intercept log-linear model for count data is used to conditional likelihood estimation. Conditional on $\gamma_i = \beta_0 + U_i$, the following assumptions are made:

" 1. Y_{ij} follows a Poisson distribution such that

$$\log(E(Y_{ij}|\gamma_i)) = \gamma_i + \mathbf{x}_{ij}^T\beta + \log(t_{ij}), \qquad j = 1, \ldots, n_i,$$

2. $Y_{i1}, \ldots, Y_{in_i}$ are independent. " [Diggle et al., 2002, p. 184]

The likelihood function for β and $\gamma_1, \ldots, \gamma_m$ then becomes proportional to

$$\prod_{i=1}^{m} \exp\left(\gamma_i \sum_{j=1}^{n_i} y_{ij} + \beta^T \sum_{j=1}^{n_i} y_{ij}\mathbf{x}_{ij} + \sum_{j=1}^{n_i} y_{ij}\log(t_{ij}) - \sum_{j=1}^{n_i} t_{ij}\exp(\gamma_i + \mathbf{x}_{ij}^T\beta)\right).$$

When conditioning on $y_{i\cdot} = \sum_{j=1}^{n_i} y_{ij}$, a conditional likelihood, which depends on β only, is obtained:

$$\prod_{i=1}^{m} \binom{y_i}{y_{i1}, \ldots, y_{in_i}} \prod_{j=1}^{n_i} \left(\frac{t_{ij}\exp(\mathbf{x}_{ij}^T\beta)}{\sum_{l=1}^{n_i} t_{il}\exp(\mathbf{x}_{il}^T\beta)}\right)^{y_{ij}}.$$

In this expression the share for subject number i is a multinomial probability, where

$$\pi_{ij} = \frac{t_{ij}\exp(\mathbf{x}_{ij}^T\beta)}{\sum_{l=1}^{n_i} t_{il}\exp(\mathbf{x}_{il}^T\beta)}$$

account for the probability that each of the $y_{i\cdot}$ events will fall into category j. [Diggle et al., 2002, p. 184]

Random effects model for counts
Historically there is a tradition for using the Poisson distribution as a model for count data. In biomedical applications however, it is rarely the case that the variance equals the mean. Usually the variance is bigger than the mean in those cases. Assuming there is a natural heterogeneity among the expected responses across observations can explain that over-dispersion. When the means follow a gamma distribution, the marginal distribution of the responses is the negative binomial distribution, which emerge from two assumptions:

" 1. Conditional on μ_i, the response variable Y_{ij} has a Poisson distribution with mean μ_i;

2. The μ_i are independent gamma random variables with mean μ and variance $\phi\mu^2$. " [Diggle et al., 2002, p. 186-187]

This means that unconditionally the responses are distributed as a negative binomial distribution, where

$$E(Y_{ij}) = \mu \qquad \text{and} \qquad Var(Y_{ij}) = \mu + \phi\mu^2.$$

The simplest extension of this model is when the μ_i depend on covariates through some parametric function. Most common is the use of the log-linear model in which

$$\log(\mu_i) = \mathbf{x}_i^T \beta.$$

Here the explanatory variables do not vary within subjects, which must be regarded as an important limitation. Therefore the model is extended, so the random effects are added on the same scale as the fixed effects. Hereby the following assumptions are made:

" 1'. $\log(E(Y_{ij}|\mathbf{U}_i)) = \mathbf{x}_{ij}^T\beta + \mathbf{d}_{ij}^T\mathbf{U}_i$;

2'. given $\mathbf{U}_i$, the responses $Y_{i1}, \ldots, Y_{in_i}$ are independent Poisson variables with mean $E(Y_{ij}|U_i)$;

3'. the $\mathbf{U}_i$ are independent realizations from a distribution with density function $f(\mathbf{U}_i; G)$. " [Diggle et al., 2002, p. 187]

This approach makes it possible to take into account that the random effects vary within subjects, which means that $\mathbf{d}_{ij}$ does not need to be constant for a given subject. [Diggle et al., 2002, p. 186-187]

The Poisson model is useful in this project, when we wish to analyze *how* big the patients over-use of anti psychotics are. We can count with how many DDD the patients exceeds the recommended 12 weeks a year.

Chapter 8

Discussion

In this report we have described methods and models that fit the purpose of the project, that is to perform a thorough analysis of the use of antipsychotics in elderly patients with dementia. Approaches to modeling repeated measurements taking the within subjects dependency into consideration have been described. Combining GLM with random effects led to generalized linear mixed models, and both the conditional likelihood estimation and the full maximum likelihood estimation methods are reasonable alternatives for the analysis. Furthermore, logistic models for binary data and Poisson models for count data have been explored. These methods will allow to analyse the number of so-called high-dosage users and just how big the potential over-use is, respectively.

When we proceed with the data analysis in the next project, several aspects should be considered. First, investigating the theory of longitudinal data has provided different approaches and models. Second, different covariates need to be taken into account.

Regarding the different methods of estimation; when performing the analysis it could be interesting to compare some of them. For example, a comparison between conditional likelihood estimation and full maximum likelihood estimation seems relevant. The conditional likelihood method is more robust than the full maximum likelihood. On the other hand, full maximum likelihood estimation allows us to learn about one individual's coefficient by understanding the variability in coefficients across the whole population. In a comparison between the two approaches we can see if there is substantial difference in the estimates, and in the distribution of the variances.

Another option would be to compare the estimates found in the longitudinal analysis with those from a regular model with person-fixed intercept. In that way we can explore if the extra work connected with the longitudinal study pays off.

The most obvious explanatory variables are age and sex. When looking at the effect

of age, we should consider the possibility that the older a patient, the more severe the syndrome. In addition to age and sex, we find it relevant to use the recent issue of a warning from the Danish Health and Medicines Authority as a covariate. The use of antipsychotics should be decreasing after a warning has been issued. Furthermore it would be relevant to use information about whether or not the patient lives at home or in a nursing home. Again, we have in mind that when the patients live in a nursing home, the illness is probably more severe than when they still lives at home. A previous study showed regional differences across Denmark [Sundhedsstyrelsen, 2005, p. 1]. Therefore, it may be of interest to see if these differences still exist, and if so if they are somehow related to whether or not the region's center for diagnostic of dementia is located in a somatic or a psychiatric department.

All in all there is lot of work to be done, and it will be interesting to perform the analysis using the theory of longitudinal data.

Bibliography

Azzalini, A. (1996). *Statistical Inference Based on the likelihood.* Chapman & Hall, 1. ed. edition.

Ballard, C. and Corbett, A. (2013). Agitation and aggression in people with Alzheimer's disease. *Current Opinion Psychiatry; 26: p. 252-259.*

Diggle, P. J., Heagerty, P., Liang, K.-Y., and Zeger, S. L. (2002). *Analysis of Longitudinal Data.* Oxford University Press, 2. ed. edition.

for Rationel Farmakoterapi, I. (2006). Psykofarmaka til ældre med delir eller demens.

Hasselbach, S. (2013). Demens. www.netdoktor.dk/sygdomme/fakta/demens.htm.

Lynge, E., Sandegaard, J. L., and Rebolj, M. (2011). The Danish National Patient Register. *Scandinavian Journal of Public Health; 39(Suppl 7): p. 30-33.*

Mors, O., Perto, G. P., and Mortensen, P. B. (2011). The Danish Psychiatric Central Research Register. *Scandinavian Journal of Public Health; 39(Suppl 7): p. 54-57.*

Olofsson, P. (2005). *Probability, Statistics, and Stochastic Processes.* Wiley, 2. ed. edition.

Phung, T. K. T., Andersen, B. B., Høgh, P., Kessing, L. V., Mortensen, P. B., and Waldemar, G. (2007a). Validity of Dementia Diagnoses in the Danish Hospital Registers. *Dementia and Geriatric Cognitive Disorders; 24: p. 220-228.*

Phung, T. K. T., Waltoft, B. L., Kessing, L. V., Mortensen, P. B., and Waldemar, G. (2007b). Time trend in Diagnosing Dementia in Secondary Care. *Dementia and Geriatric Cognitive Disorders; 29: p. 146-153.*

Socialministeriet (2010). Kortlægning af demensområdet i Danmark 2010.

Statistik, D. (2013). Håndbog til data i Lægemiddelstatistikregistret. www.dst.dk/~/media/Kontorer/13.../LMDB%20Håndbog%202014.pdf.

Sundhedsstyrelsen (2005). Forbruget af anti-psykotiske lægemidler blandt ældre.

Sundhedsstyrelsen (2013). National klinisk retningslinje for udredning og behandling af demens.

WHO (2015). Dementia. http://www.who.int/mediacentre/factsheets/fs362/en/.