Modelling the Disease Course of Multiple Sclerosis with Mixed Effects

Claudia Lamina

Modelling the Disease Course of Multiple Sclerosis with Mixed Effects

Diplomarbeit

(Erster Teil der Diplom-Hauptprüfung)
im
Studiengang Statistik
an der
Fakultät für Mathematik, Informatik und Statistik
der
Ludwigs-Maximilians-Universität München

Vorgelegt von:
Claudia Lamina

Betreuer:
Dr. Mariaclelia Di Serio

Referenten:
Prof. Dr. Albrecht Neiß
Prof. Dr. Ludwig Fahrmeir

München, den 19. September 2002

Acknowledgements

I would like to thank Dr. Mariaclelia Di Serio for her constant support, but also motivating criticism. I also thank Prof. Dr. Albrecht Neiß and Prof. Dr. Ludwig Fahrmeir for their support and for ideas and advices on the statistical analysis. I want to express my thanks to the team of the Sylvia Lawry Centre for Multiple Sclerosis Research for many motivating discussions and the Institut für Medizinische Statistik und Epidemiologie for financial support and providing a PC workplace. I especially thank Dr. Philip Young for correcting my "German-style" English and Prof. Dr. Ludwig Kappos, Basel, for his medical expertise.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Multiple sclerosis is a chronic, potentially disabling, disease, that affects the brain and spinal cord (central nervous system). Astonishingly little is known about the causes and potential triggers of the disease. However, major clinical trials in the 80s made a breakthrough in the treatment of MS patients. Today, there are several medications available which intervene in disease progression or manage acute attacks. With this background, efforts had to be taken to enhance research and improve existing methods. New methods for clinical trials, both medical and statistical, had to be found because the standard placebo-controlled study could no longer be conducted. The founding of the Sylvia Lawry Centre for Multiple Sclerosis Research (SLCMSR), with the aim of establishing a large database on data of MS clinical trials, is one of the ways to deal with this new situation.

This thesis is aimed at finding new statistical methods for analyzing and predicting the clinical progression of MS patients within clinical trials. The data available for analysis are those provided by the first releases of the SLCMSR database from spring 2002. Starting from the main goals of the SLC, that is the computation of the possible disease course in a "virtual patient", the focus was centered on finding structures and relationships in the data. However, putting statistical methods to the test was even more important. The question was which models are worth following in the future and on further releases of the database?

The thesis is structured as it follows: the second chapter gives an overview of the disease and provides the essential biomedical background. Aims and philosophy of the Sylvia Lawry Centre and a description of the dataset used are then described in detail.

In the following chapter explanations of the statistical methods can be found. Actually the common statistical tools used in the MS community focuses on "time to event" as endpoints. That means that survival analysis is mainly used. This is due to the characteristics of the main response variable in MS, the Expanded Disability Status Scale (EDSS) which measures the grade of disability. The EDSS has many shortcomings, one of them being nonlinear and discontinuous. Thus, popular outcome measures in clinical trials have been time to reach a certain level in EDSS, or time to worsening, defined by an increase of 1 point in EDSS. Survival models have been used to analyze such kind of data. However, a lot of available information is lost, as measurements between the first observation and the reaching of the event are neglected. To take advantage of the repeated measures on MS patients participating in a clinical study, the dataset has been analyzed longitudinally. Then, the modified variance-covariance structure was to be taken into account by introducing a random effect in addition to the normal error variance. This subject-specific effect can also be thought of explaining unknown or unobserved variables. In the case of MS, where disease progression is highly heterogeneous from one patient to the next, random effects could even evolve as more relevant than in other situations. Neglected unobserved heterogeneity may lead to considerably biased estimates for the remaining effects. Mixed effect models, including both random and fixed effects, are explained in the third chapter. In this thesis, two methods for mixed models were followed:

- random-intercept: a mixed model with subject-specific intercepts
- quadratic random-slope: a mixed model with subject-specific intercept, slope and quadratic slope.

The main interest here lies in the question of how much the disability of a patient decreased or increased during the course of a clinical study. Hence, the change in EDSS has been taken as a response variable. In a first attempt, this variable has been assumed to be metric and a Gaussian model has been analyzed. To test, whether this assumption can be justified, the metric variable was recategorized into five ordered categories and an ordinal threshold model has been estimated.

Several covariates were also available. In particular, the included variables were those recognized by the MS community as prognostic risk factors

such as age at disease onset, gender, the EDSS value at the first observation in the study, duration of the disease from onset to first observation, the disease course, the patients have been categorized in and the time from first observation. To ensure, that all relationships between covariates and the response variable are detected, metric variables were modelled with smooth functions. Thus, it is not necassary to restrict oneself to linearity assumptions or other functional forms. Two different methods of smoothing have been used: Polynomial regression splines in the classical and P-Splines in the Bayesian setting. All of these already mentioned methods, mixed effects, the ordinal threshold model and generalized additve models with smooth functions are presented in the classical setting first. The main characteristics as well as estimation and inference procedures, based on the maximum likelihood principles, are explained.

Next, an introduction on Markov chain Monte Carlo methods (MCMC) is given. Bayesian statistics, based on MCMC, are extremely flexible and many complex methods, that cause problems in other settings, can be incorporated. In this thesis, random effects and smooth P-Spline functions are embedded in the Bayesian setting.

Chapter 4 contains the model formulations and results of the analyses. Following a short description of the dataset and variables, the results are given together with interpretations and conclusions to each model. The ordinal threshold model with mixed effects couldn't be analyzed with Bayesian methods due to numerical problems. Thus, the program MIXOR, a mixed effects ordinal regression tool, was used in this case. Both Gaussian models, the random-intercept as well as the random-slope model, were estimated with the program BayesX.

Finally, chapter 5 gives an overview of the results and problems arising out of the conducted estimations. Model improvements and generalizations are mentioned as well as suggestions for further analyses.

Chapter 2

Medical Background

Statistical modelling on data of Multiple Sclerosis patients is not easy due to the complex and heterogeneous disease course. It is therefore necessary to provide some background information on the disease, its diagnosis and treatment. Since this is just a brief overview, the interested reader is referred to [Kesselring(1996)] to get more detailed information on the disease in general and on recent research.

2.1 What is MS?

Multiple Sclerosis is a chronic inflammatory disease of the central nervous system (brain, spinal cord). With about 3 MS patients in every 10000 people, it is one of the most common neurological diseases in Northern and Middle Europe and Northamerica. Most patients are diagnosed between the ages of 20 and 40, and about two thirds are women. Moreover there is a regional and ethnical concentration in the incidence of MS: MS patients are almost always Caucasian and grew up in colder climates . MS is not a life-threatening disease. Only a very few patients have such a severe type of disease that it shortens life expectancy. The vast majority can be expected to live a normal or near-normal life-span.

Strong evidence suggests that MS results from an autoimmune process. It is not currently known what triggers this reaction. It is possible, that a common virus acts as a trigger for MS. However, there seems to be some genetic predisposition, as there is a slightly increased incidence of MS in close relatives. Yet MS is not directly inherited. Even though the causes of the

disease are not clear, there is a wide consensus between researchers about the underlying mechanism. The blood/brain barrier is responsible for keeping the tissue of the central nervous system (CNS) and the blood vessels apart. But in case of MS, this important safety wall can be breached temporarily by white blood cells. Certain immune cells mistakenly identify the coating of the nerves, the myelin sheath, as foreign substance and attack it. During this demyelinating process, the myelin sheath becomes inflamed and eventually destroyed. Myelin serves as an insulation of the nerve fibers and makes fast and efficient conduction of nerve impulses possible. Therefore, when myelin is destroyed, some impulses are blocked or delayed from traveling to or from the brain. While at first a regeneration of the myelin is possible, later, scar tissue (sclerosis) may be formed at the lesion sites. Ultimately, the demyelination leads to degeneration of the nerves themselves.

As damage to the myelin sheath can occur any time and can affect any part of the brain or spinal cord, two MS patients with the same set of symptoms cannot be found. The multiple occurrence of scarring has given the name of the disease: multiple sclerosis.

2.2 Disease course

Determined by the multiple locations of damage in the CNS, there can be a lot of different symptoms, such as:

- Visual disturbances
- Balance and coordination problems
- weakness
- spasticity
- altered sensation
- pain
- fatigue
- bladder/bowel/sexual dysfunction
- cognitive and emotional disturbances

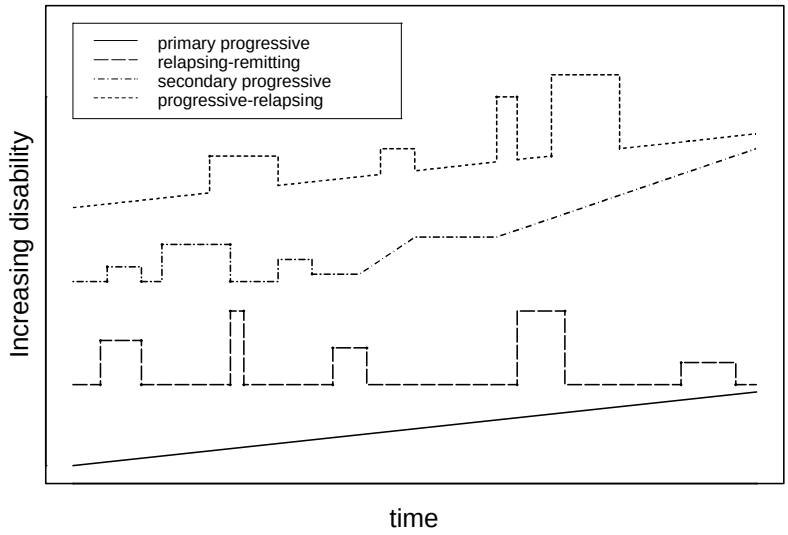

Figure 2.1: Possible disease courses in MS

Periods of active MS symptoms are called attacks or relapses. Initially, attacks are followed by complete or partial remissions. When and how frequent attacks occur, is not predictable. A few weeks, or a few years may go by between two consecutive relapses. This period, which most patients go through in the beginning of the disease is called *relapsing-remitting (rr)*. Other patients may experience a steady worsening of symptoms from onset (*primary-progressive MS (pp)*), but this is very rare. The majority enter this stage of continuous deterioration usually after several years of a relapsing-remitting MS (*secondary-progressive MS (sp)*). One can also observe "mixed" forms like the *progressive-relapsing (pr)* disease course, where patients go through a steady increase of disability from beginning, but with additional relapses. This form is rare and behaves in a manner similar to primary progressive MS. The different types of disease courses are shown in Figure 2.1.One can argue, that these different courses lead to a totally different definition of the disease. But the idea, that these stages represent only different phases of MS is equally accepted. Unfortunately, the knowledge of the natural course of the disease is still incomplete. Magnetic resonance imaging (MRI), which is able to visualize the scars in the CNS, suggests that the disease is active and

deterioration continues even in quiet periods without symptoms. This leads to the assumption that there is a underlying latent process with attacks and accumulating disability as symptoms of this process.

2.3 Diagnosis

Due to the big variety of symptoms over time and from one person to another, MS is not always easily diagnosed. Other diseases of the CNS with the same symptoms have to be ruled out first. A definite diagnosis requires evidence of multiple spots of scar tissue in different parts of the CNS via MRI, and at least two separate attacks of the disease. A neurological examination and an analysis of cerebrospinal fluid is also necessary. In some cases a definite diagnosis can take years. It can also happen, that patients seek medical help during an assumed first attack and realize with hindsight, that they've already had mild symptoms several years ago. Such retrospective information is not very reliable and therefore the onset date of the disease is not always clear.

2.4 Treatment

MS is not curable, but that doesn't mean that it is not treatable. There are many treatments and medications to relieve specific symptoms [Polman et al.(2001)]. *Corticosteroids* may be used to shorten acute attacks. But before the beginning of the nineties there were no treatments to modify the disease course . With the knowledge that MS is an autoimmune disease immune modulating therapies were considered. Many clinical trials have been conducted during the last 15 years. Eventually, this led to regulatory approval of four agents. Their active ingredients *Interferon beta 1a*, *Interferon beta 1b* and *Glatiramer* reduce the severity and frequency of relapses and slow the onset of disability. They have also been shown to reduce the progression rate of disability. Various other therapies are already being put to the test. Clinical experience suggests, that these already approved medications are most effective if taken early in the disease. This is a new challenge for improving the diagnosis at an early level.

2.5 Quantifying disability

The degree of disability is usually measured by the so called *Kurtzke Expanded Disability Status Scale* (*EDSS*)[Kurtzke(1983)]. The EDSS measures MS-related impairments of functional systems (FS). The values of these systems, ranging between 0 (normal function) and 5 (severe disability) sometimes 0 to 6, are used to define the categorical EDSS score (see Table 2.1), ranging from 0 (no MS symptoms) to 10 (death due to MS) in half point increments. The FS have been chosen so that they are independent from each other, yet together they reflect all the neurologic impairments in MS.

The functional systems in MS:

- Pyramidal Function
- Cerebellar Function
- BrainStem Function
- Sensory Function
- Bowel and Bladder Function
- Visual Function
- Cerebral Function
- Other Function

The predecessor of the EDSS, DSS, was used in clinical studies from 1955. It was modified to become the EDSS in 1985 and is to date the most widely used measurement scale. Popular outcome measures in clinical studies have been time to reach EDSS 4.0 or 6.0 or defined worsening, that is increase of EDSS of at least 1 point. Thus, the EDSS is probably the only variable to be found in every clinical study and natural history data base.

Despite this significance of the EDSS, it has many well-known shortcomings:

- the translation from the functional system status scores into the EDSS is somewhat complicated
- it is insensitive to cognitive dysfunction in MS

0	Normal Neurological Exam
1	No disability, minimal signs on 1 FS
1.5	No disability minimal signs on 2 of 7 FS
2	Minimal disability in 1 of 7 FS
2.5	Minimal disability in 2 FS
3	Moderate disability in 1 FS; or mild disability in 3 - 4 FS, though fully ambulatory
3.5	Fully ambulatory but with moderate disability in 1 FS and mild disability in 1 or 2 FS; or moderate disability in 2 FS; or mild disability in 5 FS
4	Fully ambulatory without aid, up and about 12hrs a day despite relatively severe disability. Able to walk without aid 500 meters
4.5	Fully ambulatory without aid, up and about much of day, able to work a full day, may otherwise have some limitations of full activity or require minimal assistance. Relatively severe disability. Able to walk without aid 300 meters
5	Ambulatory without aid for about 200 meters. Disability impairs full daily activities
5.5	Ambulatory for 100 meters, disability precludes full daily activities
6	Intermittent or unilateral constant assistance (cane, crutch or brace) required to walk 100 meters with or without resting
6.5	Constant bilateral support (cane, crutch or braces) required to walk 20 meters without resting
7	Unable to walk beyond 5 meters even with aid, essentially restricted to wheelchair, wheels self, transfers alone; active in wheelchair about 12 hours a day
7.5	Unable to take more than a few steps, restricted to wheelchair, may need aid to transfer; wheels self, but may require motorized chair for full day's activities
8	Essentially restricted to bed, chair, or wheelchair, but may be out of bed much of day; retains self care functions, generally effective use of arms
8.5	Essentially restricted to bed much of day, some effective use of arms, retains some self care functions
9	Helpless bed patient, can communicate and eat
9.5	Unable to communicate effectively or eat/swallow
10	Death

Table 2.1: EDSS values and their definition

- there is a significant intra-rater and inter-rater variability, although there is a special EDSS-rating training in the beginning of a study
- EDSS is an ordinal (nonlinear discontinuous) scale. That is, the meaning of a 1.0-step change varies in different parts of the scale
- Because of its strong emphasis on ambulation in the middle range of the scale, the EDSS is insensitive to changes in other neurological functions

There have been attempts to find new aggregate measures, like the *Multiple Sclerosis Functional Composite* (*MSFC*)[Cutter(1999)]. It was recommended by a task force of the National Multiple Sclerosis Society as a new clinical outcome measure for clinical trials. It is a single score on a continuous scale and contains a test of walking speed, arm skills, and cognitive function. The benefit over the EDSS is improved sensitivity to changes in the middle range of disability and the consideration of cognitive impairment. But, to date, it is not widely used and has not been approved by federal agencies.

Another intuitive idea would be to make use of brain imaging techniques. However, the frequency of scarring in the MRI scans are not well correlated with clinical symptoms.

Therefore, although inadequate, EDSS is still the most recognized measurement of disability.

2.6 Ethical considerations of clinical trials in MS

Double-blind randomized placebo-controlled clinical trials have so far been the gold standard for clinical trials in MS. But due to the availability of therapies, the use of placebo groups in further clinical trials is questionable. The Declaration of Helsinki [World Medical Association(2000)], a summary of ethical principles in medical research involving human subjects, stated that "the benefits, risks, burdens and effectiveness of a new method should be tested against those of the best current ... therapeutic methods." The use of placebo is only allowed if and only if there is "no proven ... therapeutic method". To conform to these guidelines, an international Task force on Placebo-controlled clinical trials in MS [Lublin and Reingold(2001)] constituted several scenarios, in which it is nonetheless ethical for patients

to participate in clinical trials: Fully informed patients who refuse to take available medications and patients for whom available therapies have failed. Unfortunately, this compromise comes with several shortcomings: It is less generalizable, it may include more non-responders than usual and the drop-out rate will be higher. It could therefore be foreseen that future trials are likely to compare active treatment groups. Due to the variable disease course of MS, equivalence studies are not approved by authorities. Thus, trials with active control groups may not provide adequate evidence of the effectiveness of a new agent. Other disadvantages are increased sample sizes because of smaller between-group differences and lengthened duration and costs. Thus, new ways of designing clinical trials have to be found.

.

2.7 The Sylvia Lawry Centre for Multiple Sclerosis Research

2.7.1 Purpose

The Sylvia Lawry Centre for Multiple Sclerosis Research (SLCMSR) was founded 2001 at the Technical University of Munich. Its primary purpose is to advance the understanding of the course of Multiple Sclerosis in order to make future developments of therapies for MS patients faster and less costly. To overcome the ethical problems with placebo groups, past data will be used "to predict how a placebo group of patients would react under certain conditions, thus creating a 'virtual' placebo group" [McDonald(2000)]. A database has been created, that contains data on untreated patients from natural history studies and on placebo patients from major therapeutic studies conducted worldwide by academic research groups and pharmaceutical industry. A careful analysis of factors contributing to, or predicting, the clinical course of a patient should then achieve the aims mentioned above.

2.7.2 SLCMSR database

To date, the Sylvia Lawry Centre contains data from 17 placebo-controlled clinical trials and 5 natural history data bases. The combined data from 2996 patients with 22512 observations are available for analysis. So far, several

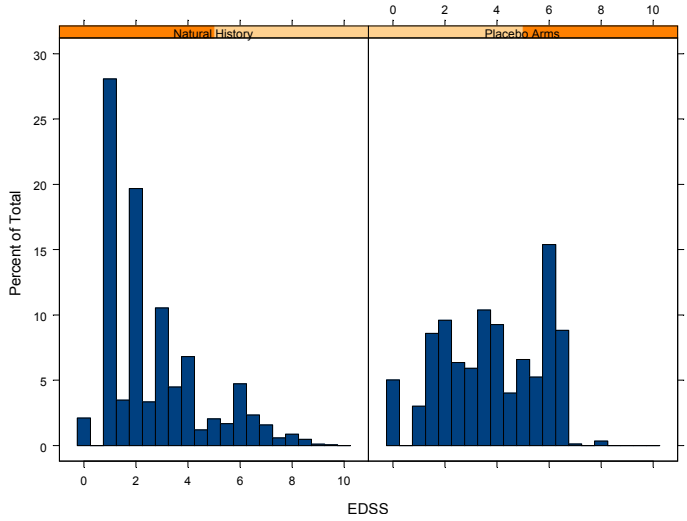

Figure 2.2: EDSS values at the first observation

demographical variables, the functional scores and EDSS-values for each time point have been pooled into one joint database. The natural question is then, are the data homogeneous enough to justify pooling? A first point to consider is whether clinical trials data should be pooled with natural history data. Patients can only participate in clinical trials, if they fulfill several inclusion criterias. These are different from study to study, but all have restrictions on age (e.g from 18 to 55), EDSS value on entry of the study (e.g. 1-2 or 3-6.5) and duration of the disease (e.g diagnosis at least 2 years before entry of the study). On the contrary, natural history databases are often regional registries, that get their data from practising physicians or evolve out of data management systems of hospitals. Naturally, many patients are observed and followed from the date of diagnosis. Thus, the majority of patients entering a natural history database, have a low disability level, measured by EDSS (Figure 2.2), were just diagnosed (Figure 2.3), and are on the average younger than patients entering a clinical trial (Figure 2.4). Moreover, poor monitoring or at least highly varying observation dates are typical for these kind of data. One can argue that clinical trials and natural history data are different subsamples of MS patients.

Clinical trials data may not represent the typical MS patient due to the

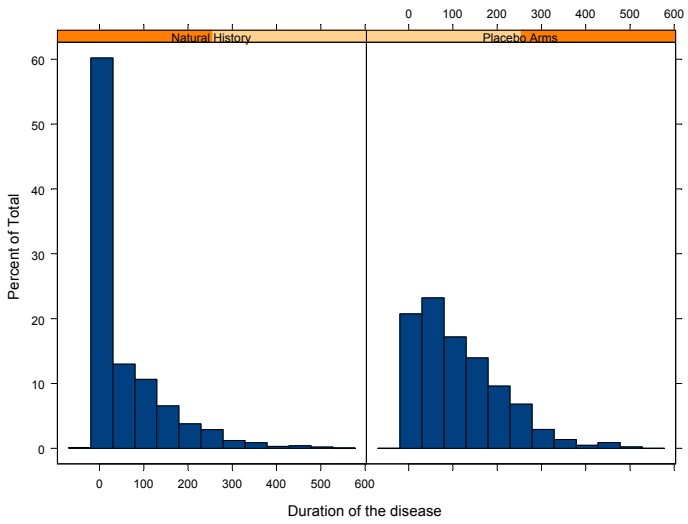

Figure 2.3: Duration of MS from onset to first observation

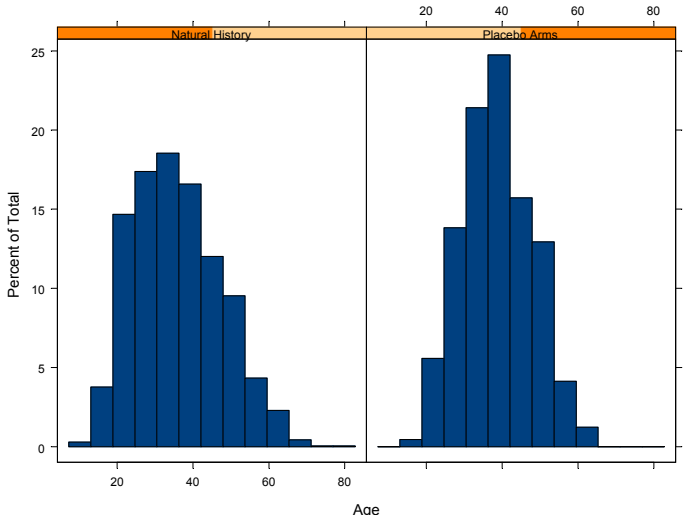

Figure 2.4: Age at the first observation

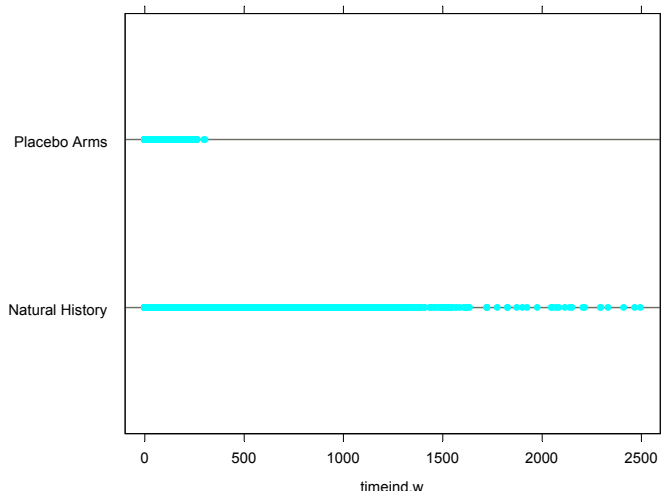

Figure 2.5: Distribution of observation dates, seperated by study type

entry restrictions. Furthermore, the observation times are much shorter than in natural history databases (Figure **??**). But good monitoring and comparable time spans (1 to 4 months) between two subsequent observations (Figure 2.6), makes them much easier to handle. As a consequence, the analysis was restricted to the data from clinical trials only. The general goal of the centre, to facilitate clinical trials in MS studies, is in line with this decision. Unfortunately, this limits the number of patients to 897 and the number of observations to 8716. Still the question of whether these data can be pooled, remains. Of course, a study effect can never be ruled out. But this has been an important question, since multicentre studies were first conducted. Work is already in progress to assess such effects. But as the data analysts are not allowed to identify the different studies due to data safety arrangements with the data donors, including a category for each study is not an option to consider in this work. Differences in inclusion and exclusion criteria can be assessed by the available baseline variables. Furthermore, much effort has been put on training of examining neurologists in MS studys, so that comparability of neurologic measurements can be asssumed.

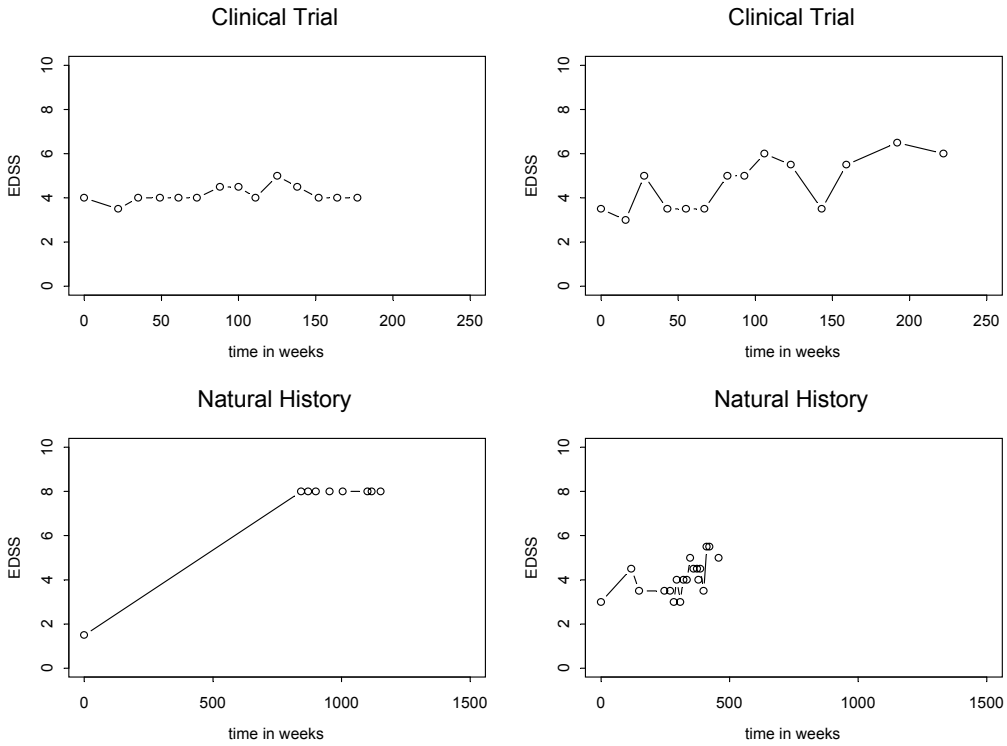

Figure 2.6: Examples for typical patient profiles, seperated for study types

Chapter 3

Statistical Methodology

Throughout the analysis of the SLC dataset, a wide range of statistical methods has been used. Due to the longitudinal stucture of the data, mixed models have been a natural choice. The basic theory of linear mixed models will be explained first, and will then be extended to models with an ordinal response. Both of these models, the mixed model and ordinal threshold model, are based on the concept of generalized linear models (GLM). To introduce the notation, some terms need to be clarified.

Classical linear models have been the standard tools to assess the dependence of a normally distributed response variable Y on covariates $X_1, ..., X_p$. Generalized linear models extend those models to allow also for non-normally distributed response variables. As in linear regression, the effect of the covariates is explained by the linear predictor

$$\eta = \alpha + X_1\beta_1 + ... + X_p\beta_p \tag{3.1}$$

where $\beta = (\beta_1, ..., \beta_p)'$ is the unknown parameter vector and the distribution of the response belongs to an exponential family. By taking the normal distribution, the familiar linear regression is returned as a special case. Some other distributions, that belong to an exponential familiy, are Binomial, Poisson, Exponential, Gamma and Inverse Gaussian distributions.

The second step of generalization is done by the so called link-function: The linear predictor η is related to the mean $\mu = E(Y)$ of the response variable Y of interest by a monotone one-to-one link-function g and its unique inverse function h by

$$g(\mu) = \eta \Longleftrightarrow \mu = h(\eta). \tag{3.2}$$

Thus, in contrast to linear regression, the mean itself is not modelled, but rather a transformation of the mean, introduced by the link-function and its inverse function.

The assumptions of the GLM are loose enough to cover a wide class of statistical models, but nevertheless tight enough to allow the development of estimation and inference methodologies. The explanation of these methods goes beyond the scope of this thesis. It is assumed that the reader is either familiar with basic concepts of GLM's and estimation and inference methods applied to it, or is referred to [McCullough and Nelder(1989)], where full details on this subject can be found.

Other models desribed in this chapter are generalized additive models to allow for non-linear relationship of the predictors and the response variable. These concept makes modelling more flexible, as the data itself determine the form of the relationship, without too restrictive constraints.

Combining all the methods mentioned above that is a mixed model with an ordinal response and a nonlinear relationship, is not straightforward. Putting it all together within a Bayesian framework was the main idea of this thesis. Section 3.4 contains a description of Bayesian methods and MCMC simulation.

3.1 Mixed models

Clinical studies often focus on the investigation of the change of a specific parameter over time. This is normally accomplished by repeated measures of the interesting variable in the included patients. These data are often heterogeneous in their structure: the time intervals between measurements can vary among individuals and not even the number of observations has to be the same due to skipped examinations or withdrawals from the study. Furthermore, the correlation between repeated measurements within subjects has to be included. These complications are difficult to account for in a conventional analysis. In a classical linear model one could take a dummy variable for the patients into the model. But then these effects are assumed to take constant values and inference can only be applied to the individuals observed. But as patients in a study are assumed to represent the whole population of patients, this approach doesn't make sense. So called random effects are therefore introduced. These introduce another source of variation in addition to the residual variance. They are assumed to be random and have

a probability distribution function. With this approach one is able to account for natural heterogeneity among individuals. It can be thought of accounting for unknown variables that have not been observed. In practice, one not only wants to include subject-specific effects, but also population-specific effects that are constant among all individuals, the so called fixed effects. The two kind of effects are defined in the following (see [Milliken and Johnson(1992)]):

Definition 1 *Random effects*
A factor is random if its levels consist of a random sample of levels from a population of possible levels. Interaction effects with random effects are also random

Definition 2 *Fixed effects*
A factor is fixed if its levels are selected by a non-random process or if its levels consist of the entire population of possible levels

Definition 3 *Mixed model*
A model is called mixed or a mixed effect model if some of the factors are fixed effects and some are random effects.

The model formulation in the next sections follows the notation in [Verbeke and Molenberghs(2000)]. Detailed description can also been found in [Fahrmeir and Tutz(2001)] and [Brown and Prescott(1999)], where extensions to generalized and nonlinear mixed effect models are explained as well. This introductory chapter is restricted to linear mixed models only. A short extension to mixed models in an ordinal threshold setting is given in section 3.2.3.

3.1.1 The linear mixed model

Linear mixed models assume that the normal response y_i for the i-th subject depends linearly on population-specific effects β and subject-specific effects b_i :

$$Y_i = X_i\beta + Z_i b_i + \epsilon_i \qquad \text{for } i = 1, ...N \text{ individuals} \tag{3.3}$$

The response vector $Y_i = (Y_{i1}, ..., Y_{iT_i})'$consists of repeated observations for subject i at time points $t = 1, ...T_i$. The p fixed effects β describe average

trends, whereas the q subject-specific parameters b_i describe how the evolution of the i-th subject deviates from the average evolution in the population. X_i and Z_i are (T_i x p)- and (T_i x q) design matrices that contain the known covariates.

Finally, the (T_i x 1)-vector ε_i is the usual vector of residuals and it is assumed to be independent $N(0, \sigma^2 R_i)$ distributed with R_i being the covariance matrix containing the temporal correlations between ε_{it}'s. Very often, R_i is chosen to be equal to the identity matrix I_{T_i} of dimension T_i, so that $cov(\varepsilon_i) = \sigma^2 I_{T_i}$. In that case, the T_i responses on individual i are independent, conditional on b_i and β. For simplicity, we will consider this conditional independence model in the following.

The basic idea underlying a random effects model is that the unobservable heterogeneity among subjects can be expressed by a random variable with a distribution function. This so called mixing distribution is often assumed to be a (multivariate) normal distribution, so that $b_i \sim N(0, Q)$. Since the mean of this distribution can be incorporated in $X_i\beta$, it is useful to set the mean vector of b_i to 0. The covariance matrix is a positive semi-definite (q x q)- matrix and it is usually assumed to be unstructered, so that.

$$Q = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \cdots & \sigma_{1q} \\ \sigma_{12} & \ddots & & & \vdots \\ \vdots & & \sigma_{ij} & & \vdots \\ \vdots & & & \ddots & \vdots \\ \sigma_{1q} & \cdots & \cdots & \cdots & \sigma_q^2 \end{pmatrix}. \tag{3.4}$$

Another assumption that has to be made is the mutual independence of the sequences $\{\varepsilon_i\}$ and $\{b_i\}$.

The N subject-specific regression models of the form (3.3) can also be combined to the following matrix-notation

$$Y = \mathbf{X}\beta + \mathbf{Z}b + \epsilon \tag{3.5}$$

where the vectors Y_i, b_i and ϵ_i and the matrices X_i have been stacked to obtain the vectors $\mathbf{Y} = (Y_1, ..., Y_i, ..., Y_N)'$, $b = (b_1, ..., b_i, ..., b_N)'$, $\varepsilon = (\varepsilon_1, ..., \varepsilon_i, ..., \varepsilon_N)'$ and $\mathbf{X} = (X_1, ..., X_i, ..., X_N)'$. Likewise, Z results in a

block-diagonal matrix with the blocks Z_i for $i = 1, ..N$:

$$\mathbf{Z} = \begin{pmatrix} Z_1 & 0 & .. & .. & 0 \\ 0 & \ddots & & & \vdots \\ \vdots & & Z_i & & \vdots \\ \vdots & & & \ddots & 0 \\ 0 & .. & .. & 0 & Z_N \end{pmatrix}. \tag{3.6}$$

Such linear mixed models do not apply to longitudinal data only, but can be used in any kind of grouped data. In multi-centre trials, where the interest doesn't lie on the effects of the centres itself, but more on the portion of variation that is attributable to the centre, the centre and centre*treatment effects can be incorporated into the model as random terms. In any other case were there are repeated measures on one group or cluster, like split-plot designs in classical designed experiments, a mixed model is an appropiate concept.

As already mentioned before, any number of random effects can be specified in a mixed model. Though, introducing too many random components can not only lead to numerical problems but make identifyability difficult. In the case of a longitudinal medical study, it may be useful to allow for both the intercept as well as the slope of evolution profiles of each patient to vary randomly. The resulting random-intercept and random-slope concepts [Fahrmeir and Tutz(2001)] are typical applications of linear mixed models and will be explained in the following examples.

Example 4 *Random-intercepts model*

Consider, that there is a cluster- or subject effect with constant slope coefficient γ, so that

$$y_i = \tau_i + \gamma w_i + \varepsilon_i \quad with \;\; \varepsilon_i \sim N(0, \sigma^2) \tag{3.7}$$

The intercepts are assumed to be iid $N(\tau, \sigma^2)$ distributed. That means graphically, that the effect for each subject is parallel to the population trend (see Figure 3.1). With $\beta' = (\tau, \gamma')$, $X_i' = (1, w_i')$, $Z_i' = 1$ and $b_i = (\tau_i - \tau) \sim N(0, \sigma_\tau^2)$ a linear random effects model of the form (3.3) is obtained. In practice, a random-intercepts model is achieved by taking the subject identification number as a random variable.

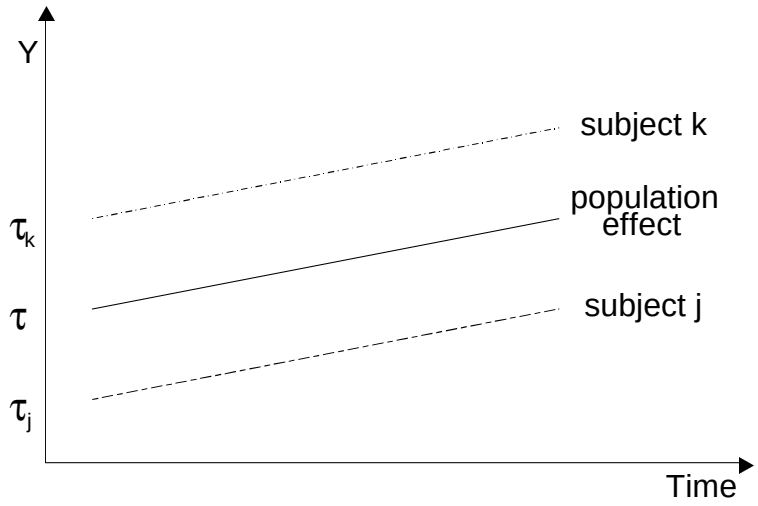

Figure 3.1: The random-intercepts model

Example 5 *Random-slopes model*

Random-intercepts models are restrictive in that they require the slope coefficients to be equal for each subject. A random slope model also allows other coefficients to vary between subjects. Thus, in a longitudinal setting, the evolution profiles for each subject have specific intercepts and and slopes (see Figure 3.2), following the model formulation

$$y_i = \tau_i + \gamma_i w_i + \varepsilon_i \quad with \;\; \varepsilon_i \sim N(0, \sigma^2) \tag{3.8}$$

Suppose, that the regression coefficients $\beta_i' = (\tau_i, \gamma_i')$ vary independently across subjects according to a normal density with $\beta_i \sim N(\beta, Q)$,where $E(\beta_i) = \beta$ can be interpreted as the population effect. That means that the subject-specific effect can be written as $b_i = (\beta_i - \beta)$ with $b_i \sim N(0, Q)$. Rewriting the regressors as $X_i' = Z_i' = (1, w_i')$ returns the linear random effects model of the form (3.3).

So far, fixed effects haven't been considered in this example of a random-slopes model. For such a mixed model, where some coefficients vary across subjects and others are constant, the parameter vector β_i has to be partitioned into a fixed and random part: $\beta_i' = (\beta_{i1}', \beta_{i2}')$ with $\beta_{i2} = \beta_2$ for all i being the fixed part. $X_i' = (X_{i1}', X_{i2}')$ has to be split, too. The parameter β_i then

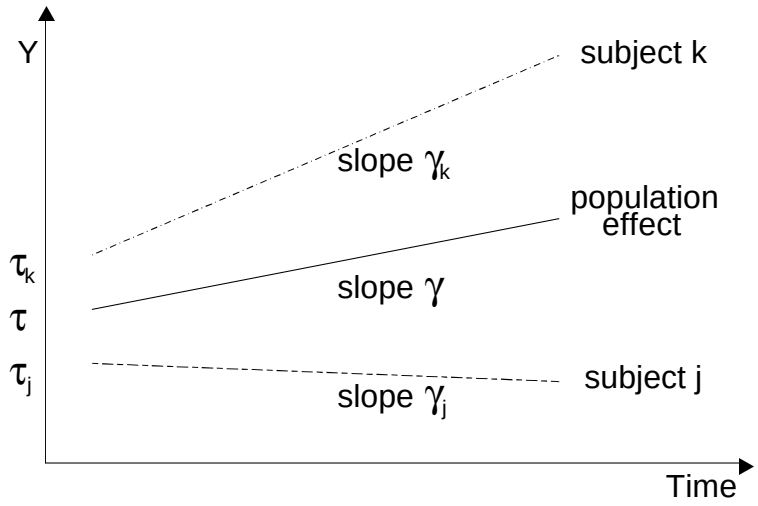

Figure 3.2: The random slopes model

follows the multivariate normal distribution:

$$\beta_i = \begin{pmatrix} \beta_{i1} \\ \beta_{i2} \end{pmatrix} \sim N\left(\begin{pmatrix} \beta_{i1} \\ \beta_{i2} \end{pmatrix}, \begin{pmatrix} Q & 0 \\ 0 & 0 \end{pmatrix} \right) \tag{3.9}$$

The coefficient β_i has to be rewritten as $\beta_i = \beta + a_i$ with $a_i' = (b_i', 0)$ and $b_i \sim N(0, Q)$ to get the familiar formula (3.3) for a mixed effects model with $X_i' = (X_{i1}', X_{i2}')$ and $Z_i = X_{i1}$.

In practice, one gets a random-slopes model by taking the subject identification number as well as the interaction of this subject id with time (in a longitudinal setting) as random.

Obviously, choosing the appropiate fixed and random effects in a model is not an easy task. The second step in modelling is the estimation of the parameters. Several methods have been used to estimate the parameters for the fixed effects. Two of these, the maximum likelihood (ML) and restricted maximum likelihood (REML) will be explained in the following.

3.1.2 Estimation and Inference for fixed effects

Estimation of fixed effects does not explicitely assume the presence of random effects. Thus, it is based on the so called *marginal model*:

Since both variance components ε_i and b_i are Gaussian, the model (3.3) can be rewritten as a multivariate heteroscedastic linear regression model:

$$Y_i = X_i\beta + \epsilon_i^*. \tag{3.10}$$

$\epsilon_i^* = Z_i b_i + \epsilon_i$ consists of the random component and the residual variance and is independent and normally $N(0, V_i)$ distributed with

$$V_i = Z_i Q Z_i' + \sigma^2 R_i. \tag{3.11}$$

This marginal version of the linear mixed model is not conditioned on the subject-specific effects b_i anymore.

The corresponding marginal version of the matrix form (3.5) can be written as

$$Y = X\beta + \epsilon^* \quad \text{with} \quad \epsilon^* = Zb + \epsilon. \tag{3.12}$$

The marginal distribution for Y is normal with $E(Y|X,\beta) = X\beta$ and covariance matrix V equal to a block-diagonal matrix with blocks V_i, as given in (3.11):

$$V = \begin{pmatrix} V_1 & 0 & \cdots & \cdots & 0 \\ 0 & \ddots & & & \vdots \\ \vdots & & V_i & & \vdots \\ \vdots & & & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & V_N \end{pmatrix} \tag{3.13}$$

The marginal distribution of the response is therefore given by $Y_i \sim N(X\beta, V_i)$.

As the variance components are normally distributed, the best approach for estimation is to use likelihood-based methods. The classical method, Maximum Likelihood Estimation (MLE), is obtained by setting up a joint likelihood of the parameters to estimate and maximize it with respect to these parameters. In the case of a mixed model, a vector α has to be defined first, that contains all variance and covariance parameters in Q and R_i.

If α is known, one gets the following likelihood-function:

$$\begin{aligned} L_{ML}(Y_i|\beta,\alpha) &= f(Y_1,...Y_N|\beta,\alpha) \\ &= \prod_{i=1}^{N} f(Y_i|\beta,\alpha) \\ &= \prod_{i=1}^{N} \Big\{ (2\pi)^{\frac{-T_i}{2}} \, |V_i(\alpha)|^{-\frac{1}{2}} \\ &\quad \times \exp\left(-\frac{1}{2}(Y_i - X_i\beta)' V_i^{-1}(\alpha)(Y_i - X_i\beta) \right) \Big\} \end{aligned} \tag{3.14}$$

It is recommended not to maximize the likelihood itself, but the so called log-likelihood, as it is easier to handle numerically. The logarithm is a strict monotone function and therefore, maximizing $L_{ML}(\beta,\alpha)$ and $\ln L_{ML}(\beta,\alpha)$ gives the same results. The log-likelihood then has the form

$$\ln L_{ML}(\beta,\alpha) = \sum_{i=1}^{N} \ln f(Y_i|\beta,\alpha). \tag{3.15}$$

Maximizing (3.15) conditional on α gives the following maximum likelihood estimator for β:

$$\widehat{\beta}(\alpha) = \left(\sum_{i=1}^{N} X_i V_i^{-1} X_i \right)^{-1} \sum_{i=1}^{N} X_i V_i^{-1} y_i. \tag{3.16}$$

This estimator $\widehat{\beta}(\alpha)$ corresponds to a weighted least squares estimator in a multiple linear regression with the matrix $W = V^{-\frac{1}{2}}$ containing the weights.

If the assumption of known variance components doesn't hold, α has to be replaced by a consistent estimate $\widehat{\alpha}$. V_i is set to $\widehat{V}_i = V_i(\alpha)$ respectively. Simultaneous maximization is possible, but it is often done hierarchically. First, α is set fixed and β is estimated by its ML-estimate as in (3.16). Replacing β by $\widehat{\beta}(\alpha)$ in (3.14) gives a profile likelihood, that has to be maximized with respect to α to obtain the estimate $\widehat{\alpha}$. This estimate is then set into (3.16) again to get $\widehat{\beta}$.

The disadvantage of MLE is, that it produces biased estimates for the variance components in α. The more fixed effects there are to estimate,

the larger the bias of the MLE for α is. Restricted Maximum Likelihood-Estimators (REML) take the loss of degrees of freedom resulting from estimating β into account . Thus, the bias of REMLs is generally smaller. The idea of this approach is, that it eliminates the parameters β from the likelihood, so that it only depends on α. The REML estimator $\widehat{\alpha}_{REML}$ for the variance components can be obtained by maximizing

$$L_{REML}(\alpha) = C \left| \sum_{i=1}^{N} X_i V_i^{-1}(\alpha) X_i \right|^{-\frac{1}{2}} L_{ML}(\widehat{\beta}(\alpha), \alpha) \tag{3.17}$$

with C being a constant that does not depend on α. As the term $\left|\sum_{i=1}^{N} X_i V_i^{-1}(\alpha) X_i\right|$ doesn't depend on β, the joint restricted likelihood for α and β can be written as

$$L_{REML}(\beta, \alpha) = \left| \sum_{i=1}^{N} X_i V_i^{-1}(\alpha) X_i \right|^{-\frac{1}{2}} L_{ML}(\beta, \alpha). \tag{3.18}$$

Maximizing the REML likelihood function with respect to α and β simultaneously, gives the REML estimators for all parameters. Note, that the REML estimates for the fixed effects are not identical to those obtained by the ML method, although the REML estimation only depends on the variance components.

After obtaining estimates for the interesting parameters, interest usually lies in testing hypothesis on those parameters. Approximative Wald-, t-and F-Tests are available to answer the questions corresponding to the statistical tests. See [Verbeke and Molenberghs(2000)] for a detailed description of the tests.

3.1.3 Estimation and Inference for random effects

Although interest often lies only in the fixed effects and the variance components, estimation of the random effects may prove useful. These show deviation of subject-specific profiles from the population average and thus make detection of outlying subjects, or even a group of outlying subjects possible. Moreover, estimation of random effects is necessary, if prediction of $\widehat{Y}_i$ for the ith profile is of interest.

The marginal model, that has been used for estimation of the fixed effects, isn't conditioned on the subject-specific effects b_i anymore. Thus, the original

random-effects model (3.3) has to be used for inference. As defined before, random effects represent unobservable heterogeneity between subjects. They are assumed to be a normally distributed random variable with $E(b_i) = 0$ and $cov(b_i) = Q$. A technique, that considers all parameters as random, is Bayesian statistics. It is therefore natural to use Bayesian methods for estimation of the random effects. Only an outline of Bayesian statistics will be shown here, as a detailed desription follows later in 3.4.

As a first step, a distribution independent of any observation has to be specified. Such a distribution is called the prior distribution or "a priori". Combining the likelihood function together with the prior distribution gives the posterior distribution ("a posteriori") of the parameters given the observed data y_i. Inference is based on this posterior distribution. In the case of random effects, the mixing distribution $N(0, Q)$ is independent of the data and can be interpreted as the prior distribution. The prior density function is denoted by $f(b_i)$ and the density function of Y_i conditional on b_i by $f(y_i|b_i)$. The posterior density function of b_i given $Y_i = y_i$ can then be calculated using Bayes' theorem:

$$f(b_i|y_i) = \frac{f(y_i|b_i)\ f(b_i)}{\int\ f(y_i|b_i)\ f(b_i)\ db_i} \tag{3.19}$$

The "best" empirical Bayesian point estimator with respect to the mean square error is the posterior mean, which is taken as the estimator for the random effects. It is given by

$$\begin{aligned} \widehat{b_i} &= E(b_i|Y_i = y_i) \\ &= \int\ b_i\ f(b_i|y_i)\ db_i \\ &= QZ_iV_i^{-1}(\alpha)(y_i - X_i\beta) \end{aligned} \tag{3.20}$$

with the corresponding covariance matrix

$$var(\widehat{b_i}) = QZ_i' \left\{ V_i^{-1} - V_i^{-1}X_i \left(\sum_{i=1}^{N} X_i'V_i^{-1}X_i \right)^{-1} X_iV_i^{-1} \right\} Z_iQ. \tag{3.21}$$

Naturally, the posterior mean and covariance matrix depend highly on their assumed prior distribution. However, estimation of fixed effects and

variance components are robust with respect to misspecification of the prior distribution, as their estimation is based on the marginal model (3.10). Yet, the calculations of (3.20) and (3.21) were performed conditionally on all the parameters of the marginal model. It is assumed that they have been estimated before with ML or REML-methods and have been replaced by their estimates. The estimates obtained by (3.20) are called "*Empirical Bayes" (EB) estimates*. Note, that the calculation of the posterior mean and covariance matrix involves integrals, that cannot be solved analytically for most situations. It is therefore necessary to carry out approximation procedures like Monte Carlo integration (see 3.4).

3.1.4 Shrinkage

Suppose, interest lies in predicting $\widehat{Y}_i$ of subject i and the random and fixed effects have already been estimated. It follows from (3.20), that

$$\begin{aligned}
\widehat{Y}_i &= X_i\widehat{\beta} + Z_i\widehat{b}_i \\
&= X_i\widehat{\beta} + Z_iQZ_iV_i^{-1}(y_i - X_i\beta) \\
&= \left(I_{n_i} - Z_iQZ_iV_i^{-1}\right) X_i\widehat{\beta} + Z_iQZ_iV_i^{-1}y_i \\
&\quad \sum_i V_i^{-1}X_i\widehat{\beta} + \left(I_{n_i} - \sum_i V_i^{-1}\right) y_i. \qquad (3.22)
\end{aligned}$$

That means, that the individual predicted profile is a weighted average of the population-profile $X_i\widehat{\beta}$ and the observed data y_i. The degree of *shrinkage*, how such methods are usually called, depends on the variance-covariance structure of the data. If the residual variance is large in comparison to the between-subject variance expressed by the random effects, more weight is given to the overall average profile. If the opposite is true, more weight will be given to the observed data. That also implies that estimates of the random effects are shrunken towards the marginal mean compared with their fixed effects counterparts. Potential problems of extreme parameter estimates occuring by chance when based on small numbers, can then be avoided.

3.2 The ordinal threshold model

3.2.1 The model formulation

Categorized variables can often be assumed to have an ordinal structure, i.e. category "good" is better than "average", whereas "average" is better than "bad". But there is no way to quantify this quality variable or measure the difference between for example "good" and "average". If such a variable is taken as a response in a regression model, there are different ways to tackle the problem. Often, the data are assumed to be measured on an interval scale and thus, classical regression models are used. This might be a possibility, if there are many categories, but often the assumption of a metric response is not justified. As an alternative, nominal models can also be used to analyze these kind of data. This approach is acceptable, but the ordinal structure is ignored and important information lost. The model usually used for ordinal regression is the cumulative threshold model.

In this case, the response variable exists of the ordered categories 1,..k, whereas there is no useful interpretation of the differences. The threshold model is based on the idea, that there is a latent non-observable metric variable and that the observed variable merely is a categorized version of this latent variable. In the example of the three ordered categories introduced above, the underlying latent variable could be interpreted as a continuous goodness scale. If a certain boundary is reached, the observable variable takes the value "average" or "good" for a higher boundary. The relationship between the latent variable $\widetilde{Y}$ and the observable variable Y can then be described as follows:

$$Y = r \Longleftrightarrow \theta_{r-1} < \widetilde{Y} \leq \theta_r \tag{3.23}$$

on the latent continuum $-\infty = \theta_0 < \theta_1 < ... < \theta_k = \infty$ for $r = 0, ...k$ categories.

That means, if the latent variable lies between the boundaries $\theta_{r-1}$and θ_r, the observable variable takes the value r. $\widetilde{Y}$ is explained by the regressor variables in the linear form

$$\widetilde{Y} = -X'\delta + \varepsilon \tag{3.24}$$

where the nuisance parameter has the distribution function F with $E(\varepsilon) = 0$. Thus, the conditional mean of $\widetilde{Y}$ can be written as $E(\widetilde{Y}\,|X) = -X'\delta$.

These assumptions result in the cumulative model

$$P(Y \leq r|X) = F(\theta_r + X'\delta) \tag{3.25}$$

The cumulativity refers to the following property:

$$P(Y \leq r|X) = P(Y = 1|X) + P(Y = 2|X) + ... + P(Y = r|X) \tag{3.26}$$

or, respectively

$$P(Y = r|X) = P(Y \leq r|X) - P(Y \leq (r-1)|X). \tag{3.27}$$

So far, no intercept has been included in the model formulation (3.25). Including an intercept would make identifyability of the thresholds $\theta_0 < ... < \theta_{r-1}$impossible, that is ensured by setting the intercept equal to 0.

Note, that the negative sign "$-$" in (3.24) is only included to obtain a simpler form of the model (3.25). Reparametrization of the parameter vector to $\xi = (-\delta)$ would lead to the cumulative model

$$P(Y \leq r|X) = F(\theta_r - X'\xi). \tag{3.28}$$

The choice of the distribution function F naturally influences the appearence of the model. Common choices are the logistic or the normal distribution. In the following, the normal distribution will be used. A detailed description of the cumulative threshold model and more information on other modelling strategies within the framework of a threshold approach can be found in [Tutz(2000)].

Choosing the normal distribution gives the so called *ordinal probit model*:

$$P(Y \leq r|X) = \phi(\theta_r + X'\delta) \tag{3.29}$$

with

$$\phi(\nu) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\nu} e^{-\frac{x^2}{2}} dx. \tag{3.30}$$

Considering the density $F' = f$ of the latent variable, the model can be illustrated in the following way: By means of the relationship (3.23) between the observable and underlying variable the density is cut into k parts. The areas defined by the density curve and the thresholds θ_r and θ_{r+1} can then be interpreted as the probability of being in category r.

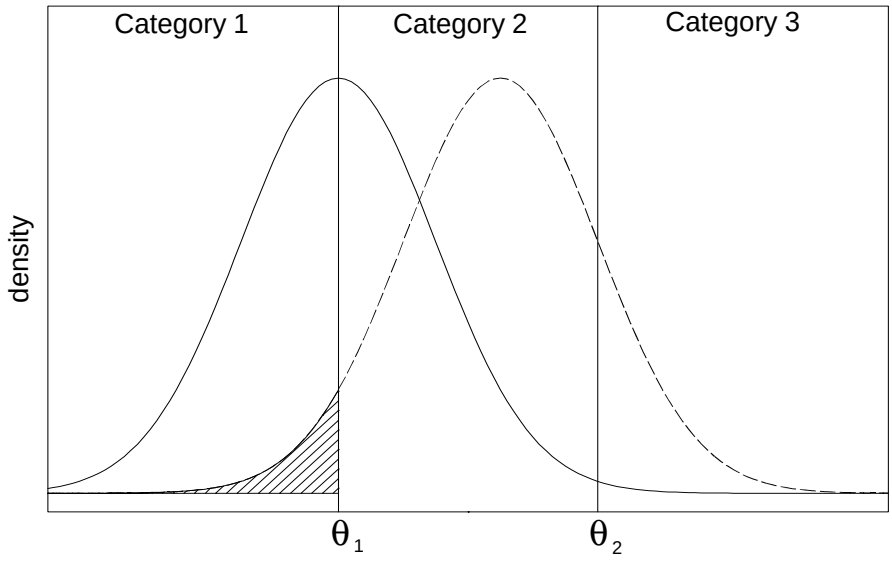

Figure 3.3: The densities of population 1 (dashed line) and population 2 (full line), cut into 3 parts by thresholds θ_1 and θ_2

Figure 3.3 illustrates the following situation: The thresholds θ_1 and θ_2 divide $\widetilde{Y}$ into 3 categories. The observable variable then belongs to one of the three ordered levels $r = 1, 2$ and 3. The regressor term $-X'\delta$ causes a shift of the non-observable variable according to the parameter-values of the regressor variables. In this example, the probability to reach a higher category is higher for population 1 compared to population 2. The hatched area shows the probability for a subject from population 2 to reach category 1.

3.2.2 The ordinal model as a GLM

The threshold model is a special case of categorical regression models. Methods that have been developed for categorical models, or more general, for generalized linear models, can also be used for ordinal regression. For this purpose, cumulative threshold models have to be embedded in the framework of generalized linear models [Fahrmeir and Tutz(2001)]. The link-function g_r is immediately given from (3.25) or (3.29) by

$$g_r(\pi_1, ...\pi_k) = F^{-1}(\pi_1 + ... + \pi_r), \quad r = 1, ..., k \tag{3.31}$$

where $\pi_r = P(Y = r|X)$ for any distribution function F and

$$g_r(\pi_1, ...\pi_k) = \phi^{-1}(\pi_1 + ... + \pi_r) = \theta_r + X'\delta = W\beta \tag{3.32}$$

for the normal distribution given in (3.30). The linear term $W\beta$ is determined by the design-matrix

$$W = \begin{pmatrix} 1 & & & x' \\ & 1 & & x' \\ & & \ddots & \vdots \\ & & 1 & x' \end{pmatrix} \tag{3.33}$$

and the parameter vector $\beta = (\theta_1, ..., \theta_k, \delta)$. The restriction to the ordinal structure $\theta_1 < ... < \theta_k$ may cause problems, if the thresholds are not seperated very well. An alternative formulation can avoid numerical flaws. The vector β is reparameterized, so that the resulting parameters are not restricted any more:

$$\alpha_1 = \theta_1, \qquad \alpha_r = \log(\theta_r - \theta_{r-1}), \quad r = 2, ...k \tag{3.34}$$

The parameter vector is now given by $\beta = (\alpha_1, ...\alpha_k, \delta)$ and the design matrix has the form

$$W = \begin{pmatrix} 1 & & & x' \\ & 1 & & 0 \\ & & \ddots & \vdots \\ & & 1 & 0 \end{pmatrix}. \tag{3.35}$$

This reparametrization returns the following form of the model:

$$F^{-1}(P(Y = 1|X) = \alpha_1 + X'\delta \tag{3.36}$$

and

$$\log(F^{-1}(P(Y \leq r|X)) - F^{-1}(P(Y \leq r - 1|X) = \alpha_r, \quad r = 2, ..., k, \tag{3.37}$$

where the parameters $\alpha_1, ...\alpha_k$ are no longer restricted.
Hence, the link-function is determined by

$$g_1(\pi_1, ...\pi_k) = F^{-1}(\pi_1) \tag{3.38}$$

and

$$g_r(\pi_1, ...\pi_k) = \log(F^{-1}(\pi_1 + ... + \pi_r) - F^{-1}(\pi_1 + ... + \pi_{r-1})). \tag{3.39}$$

Due to the representation of the cumulative threshold model as a generalized linear model it is possible to use known and validated estimation and inference procedures. For example, estimation can be done via maximum likelihood. The likelihood derived out of the latter reparametrization in this section has the form

$$L_{ML}(Y_i|\theta, \delta) = \prod_{r=1}^{k}(F(\theta_r + x_i\delta) - F(\theta_{r-1} + x_i\delta))^{d_{ir}} \quad \text{for } i = 1, ...N, \tag{3.40}$$

where

$$d_{ir} = \begin{cases} 1 & \text{if } Y_i = r \\ 0 & \text{if } Y_i \neq r \end{cases} .$$

As usual, the maximum likelihood principle for constructing an estimation function is based on maximizing the likelihood, or equivalently the corresponding log-likelihood, by the parameters.

3.2.3 Mixed effects in the threshold model

Recall, that estimation in the mixed-effects model is mainly peformed in the marginal model, but for non-normal data it is not possible to pack the mixed model (3.3) into the closed form defined in (3.10). Thus, consider the original model formulation $Y_i = X_i\beta + Z_ib_i + \epsilon_i \quad \text{for } i = 1, ..., N$ individuals hierarchically with 2 levels. In this new representation, i denotes the level-2 units ($i = 1, ..., N$ subjects in the longitudinal context) and t denotes the level-1 units ($t = 1, ...T_i$ repeated observations), that are nested within each level-2 unit. Hence, the mixed-effects model can be rewritten as

$$Y_{it} = X'_{it}\beta + Z'_{it}b_i + \epsilon_{it}. \tag{3.41}$$

Then, within a given level-2 unit i, the term (3.41) can be reduced to

$$Y_t = X'_t\beta + Z'_tb + \epsilon_t. \tag{3.42}$$

Setting $-w_t = X_t'\beta + Z_t'b$, together with the likelihood function of the ordinal threshold model (3.40), immediately leads to the so called *Maximum Marginal Likelihood Estimation*, introduced by Hedeker and Gibbons [1994]. The probability of Y_i given the parameters of the ordinal model θ and δ and the paramters of the mixed-effects part β and α, is equal to the product of the probabilites of the level-1 responses:

$$L_{ML}(Y_i|\zeta) = L_{ML}(Y_i|\theta,\delta,\beta,\alpha) = \prod_{t=1}^{T_i}\prod_{r=1}^{k}(F(\theta_r + w_{it}) - F(\theta_{r-1} + w_{it}))^{d_{itr}} \tag{3.43}$$

where

$$d_{itr} = \begin{cases} 1 & \text{if } Y_{it} = r \\ 0 & \text{if } Y_{it} \neq r \end{cases}.$$

Denote the distribution of the random component b_i as $g(b)$. Then the marginal density of Y_i in the population is expressed as

$$h(Y_i) = \int_b L_{ML}(Y_i|\zeta)g(b)db. \tag{3.44}$$

For estimation, the marginal log-likelihood

$$\log L_{ML} = \sum_{i=1}^{N} \log h(Y_i) \tag{3.45}$$

for the patterns from the N level-2 units has to be maximized. Computational details and supporting examples can be found in [Hedeker and Gibbons (1994)]. In the following, an example is given for data with an ordinal response structure and a random intercept.

Example 6 *Let x_{it}' be the design vector for the fixed effects, θ_r the rth threshold and $b_i \sim N(0,\sigma^2)$ the subject-specific effects. Then, the linear predictor has the form*

$$\eta_{itr} = \theta_r \ + b_i + x_{it}'\delta. \tag{3.46}$$

This can be interpreted as subject-specific shifting of the thresholds where

$$\theta_{ir} \ = \theta_r \ + b_i. \tag{3.47}$$

Thus, the conditional response probabilities are given by

$$\pi_{it1} = F(\theta_{i1} + x_{it}'\delta), \quad \pi_{itr} = F(\theta_{ir} + x_{it}'\delta) - \pi_{it,r-1}, \quad \textit{for } r = 2,...k. \tag{3.48}$$

3.3 Generalized additive models

For the models considered so far, there is one underlying principle: The values of the explanatory variables influence the response in a linear way. In the framework of generalized linear models the linear predictor is usually written as

$$\eta = \alpha + X_1\beta_1 + ... + X_p\beta_p, \tag{3.49}$$

where $X_{1,...}X_p$ denote the explanatory variables and $\alpha, \beta_1, ...\beta_p$ the corresponding coefficients. As usual, the linear predictor η is related to the mean with the link-function by $g(\mu) = \eta$.

In practice, this model is extremly useful and convenient. It is able to determine the importance of each predictor with a single coefficient, that is easy to interpret. But, as already mentioned, it restricts it to cases were the dependence of $E(Y)$ on $X_1, ..., X_p$ is linear in each of the predictors. By replacing the constant estimators β with smooth functions $f(x)$ data can be modelled, that don't hold the strong linearity assumption. In this so called generalized additive model (GAM), introduced by Hastie and Tibshirani [1990], the linear predictor is then assumed to be a sum of smooth functions and has the form

$$\eta = \alpha + f_1(X_1) + ... + f_p(X_p). \tag{3.50}$$

There are many different approaches for modelling the functions $f_1, ...f_p$. In principle, any known smoother can be used to estimate the function, such as LOESS, k-nearest neighbor method, polynomial smoothing splines or regression splines. "A smoother is a tool for summarizing the trend of a response measurement Y as a function of one or more predictor measurements $X_1, ...X_p$. It produces an estimate of the trend that is less variable than Y itself; hence the name *smoother*. An important property of a smoother is its nonparametric nature: it doesn't assume a rigid form of the dependence of Y on $X_1, ..., X_p$." [Hastie and Tibshirani(1990)]. That is, the shape of each of the covariate effects is data-driven: The data themselves determine the form of the relationship between the predictors and the response variable. Thus, a too restrictive approach by assuming a functional relationship can be avoided. Essentially, splines are very helpful to detect unknown trends in the data. For example, if a spline function on one predictor "looks like" a quadratic function, then at a next step a quadratic function can be used

to model this variable. In this example, a linear model would have ignored the strong relationship between the predictor and the response. Splines can therefore be used as an explorative tool to detect trends and use the gained information to formulate a more parameter sparse model.

Note, that the generalized additive model consists of a sum of such smooth functions. That is, additivity of effects is assumed. This concept retains the interpretability of the familiar linear model and allows, that some predictors can be modelled with smooth functions $f(x)$, and others with constant estimators. Out of the big set of splines, only three, the Polynomial-, B- and P-Splines are picked out, starting with the simplest, the polynomial regression spline. To indicate, that spline curves act as the interesting smooth functions $f(x)$, they will be adressed as $s(x)$ in the following.

3.3.1 Polynomial Splines

In general, splines offer a compromise between an interpolation of the data and a global smooth fit by representing the fit as a piecewise defined function. The pieces on the interval $[\xi_0, \xi_m]$ are seperated by a sequence of knots $\xi_0 < \xi_1 < ... < \xi_m$. The partial functions B_i , called basis functions, are fitted to the data in the range of two subsequent knots. They are resticted to follow a set of smoothing conditions with the neighboring basis functions at the breakpoints.

Definition 7 *Polynomial Spline*

A function s: $[\xi_0, \xi_m] \rightarrow \Re$ *is called a polynomial spline of degree k, if the following properties are satisfied:*

- *s is a polynomial of degree k in any subinterval* $[\xi_i, \xi_{i+1})$
- *s has k-1 piecewise continuous derivatives*
- *s has a derivative of degree k that is step function with jumps at* $\xi_0, \xi_1, ..., \xi_m$

All functions satisfying this definition, can be illustrated uniquely by a linear combination of $m + k - 1$ basis functions $B_i(x)$ [Hämmerlin and Hoffman(1992)]. The simplest set of basis functions has the following form (see Figure 3.4 a)):

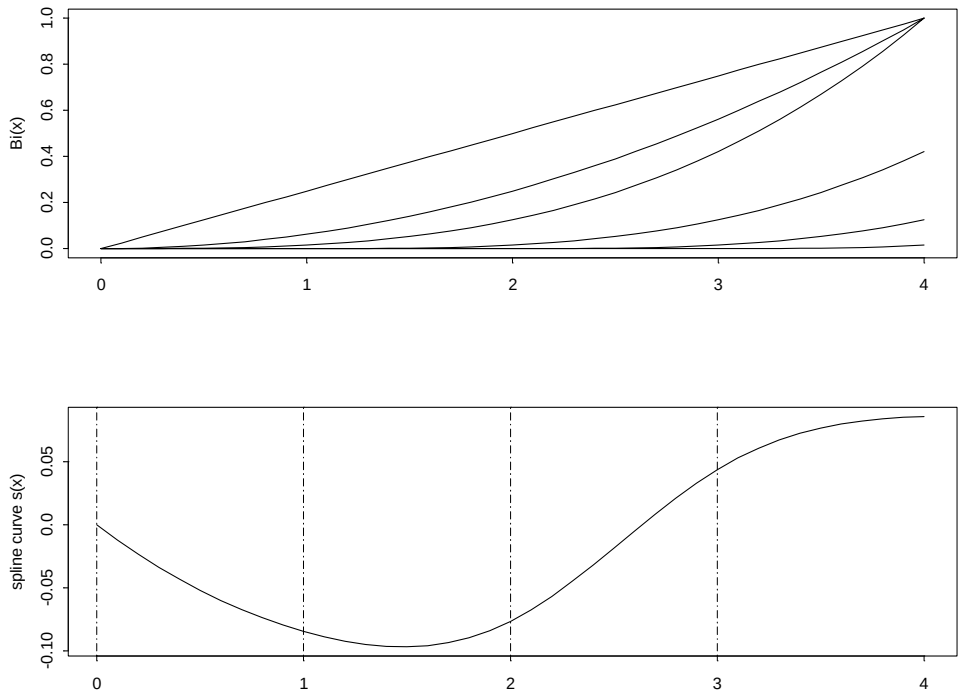

Figure 3.4: a) top: cubic spline basis functions with knots at x=0, 1, 2, 3 b) bottom: a spline function obtained by a linear combination of the basis functions

$$\begin{aligned} B_i(x) &= x^i \qquad\qquad \text{for } i = 0, 1, ..., k \\ B_i(x) &= (x - x_{(i-k)})_+^k \quad \text{for } i = k+1, k+2, ..., m+k-1, \end{aligned} \tag{3.51}$$

where

$$(x - x_{(i-k)})_+^k = \begin{cases} (x - x_{(i-k)})^k & \text{for } x > x_i \\ 0 & \text{for } x \leq x_i \end{cases} \tag{3.52}$$

denotes a truncated polynom. Hence, the name *truncated power series* for the basis functions defined in (3.51). Then, the smooth functions $s(x)$

can be written as

$$s(x) = \beta_0 + \sum_{i=1}^{k} \beta_i x^i + \sum_{i=k+1}^{m+k-1} \beta_i (x - x_{(i-k)})^k, \tag{3.53}$$

where the coefficients β_i have to be estimated. One advantage of illustrating basis functions in this form is the simple interpretation of the coefficients.

Apparantly, the form of the resulting spline is highly dependent on the number and location of knots and the degree of the polynomials. In practice, quadratic or cubic splines are used (see Figure 3.4 b)) to get smooth curves at the knots. In any case, higher degrees than three are usually not necessary, as higher order discontinuities are not detectable visually. Nevertheless, polynomial regression splines are not always satifactory: The basis functions are very heterogeneous in their values as well as in the length of their nonzero intervals. This makes them numerically unstable and sensitive towards the choice of the number and location of knots. B-Splines and P-Splines, which will be explained in the next two sections, are better choices for basis functions.

3.3.2 B-Splines

Obviously, the choice of appropiate basis functions is crucial to the final regression spline. Basic Spline Curves (B-Splines) are a popular choice for basis functions due to their numerical stable behaviour.

Definition 8 *B-Splines*

Let $\Psi = \{\xi_i\}, i \in Z$ be a set of knots with $\xi_i < \xi_{i+1}, \xi_i \longrightarrow -\infty$ for $i \longrightarrow -\infty$ and $\xi_i \longrightarrow \infty$ for $i \longrightarrow \infty$. The set of all possible splines of degree k to Ψ is called space of splines $S_k(\Psi)$.

The B-Spline of degree k to the knot ξ_i of Ψ is defined recursively:

$$k = 1: \qquad B_i^1(x) = \begin{cases} \frac{x-\xi_i}{\xi_{i+1}-\xi_i} & : \xi_i \leqq x < \xi_{i+1} \\ \frac{\xi_{i+2}-x}{\xi_{i+2}-\xi_{i+1}} & : \xi_{i+1} \leqq x < \xi_{i+2} \\ 0 & otherwise \end{cases}$$

$$k > 1: \qquad B_i^k(x) = \frac{x-\xi_i}{\xi_{i+1}-\xi_i} B_i^{k-1}(x) + \frac{\xi_{i+k+1}-x}{\xi_{i+k+1}-\xi_{i+1}} B_{i+1}^{k-1}(x)$$

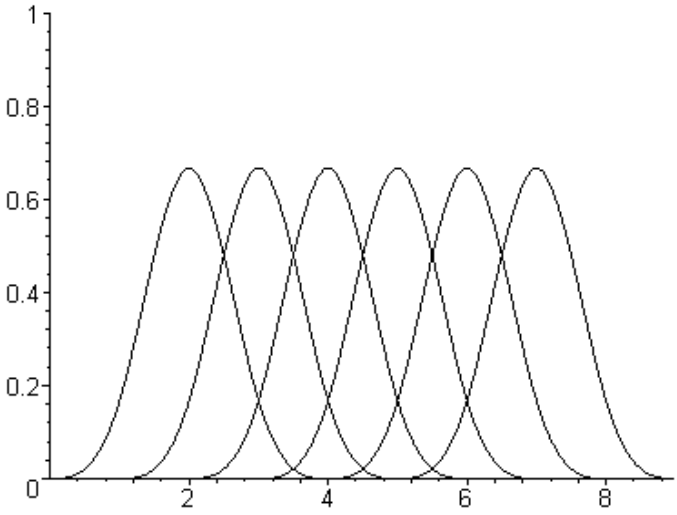

Figure 3.5: Cubic B-Spline basis functions

$\Longrightarrow$ *The Spline curve* $s \in S_k(\Psi)$ *can then be described as a linear combination of the seperate B-Splines and their coefficients* β_i :

$$s(x) = \sum_{i=-k+1}^{m-1} \beta_i B_i^k(x), x \in [\xi_1, \xi_m] \tag{3.54}$$

B-Splines depend only on the degree k and the values of Ψ. They are local, which means non-zero in a defined interval and zero outside of this interval. This then makes them numerically superior to other basis functions.

In the following example, some characteristics of B-Splines will be explained for a B-Spline of degree 3.

Example 9 *Cubic B-Splines*

Let s be a B-Spline of degree 3 and an equidistant knot vector $u = (0, 1, ...9)$ *(see Figure 3.5). Each single B-Spline function is based on 5 knots and is zero outside of the range of the knots. It consists of 4 cubic parts, joined together at the knots with continuous second order derivatives. 5 adjacent basis functions overlap to form an appropiate basis for a smooth*

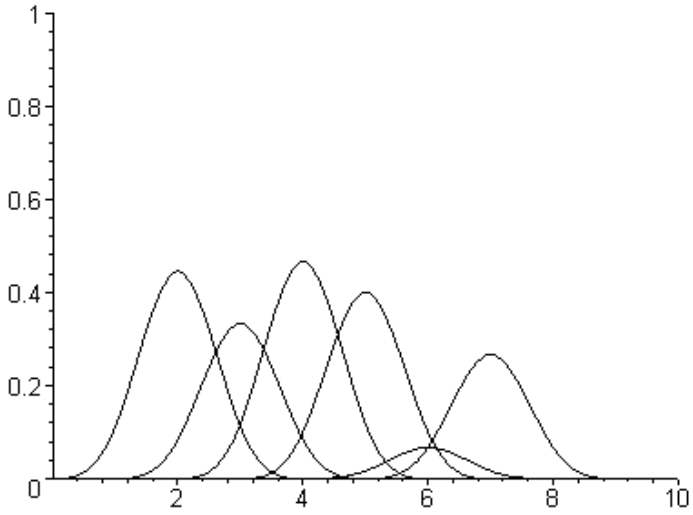

Figure 3.6: B-Spline basis functions scaled by their parameter

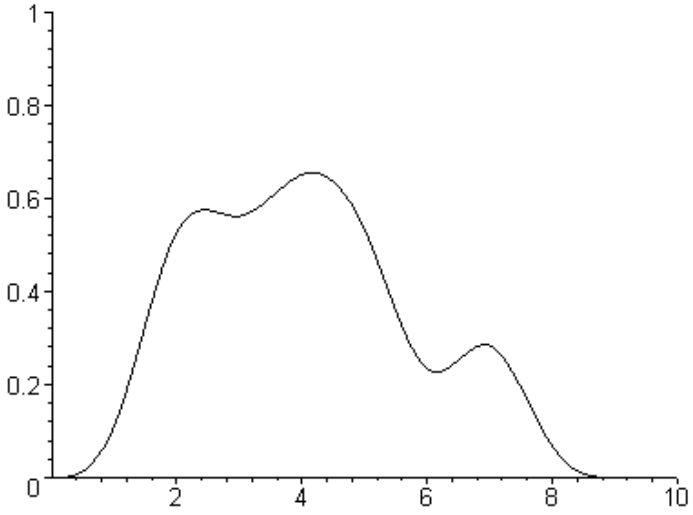

Figure 3.7: Resulting B-Spline curve as sum of the scaled basis functions

function. To ensure that each point, within the range of the data, is overlapped by 4 B-Splines, 3 invisible knots and their corresponding B-Splines are added beyond the margin of the data.

In general, for a spline curve of degree k, k invisible knots have to be added. Thus, if m knots span the range of the data, there are m+2k knots altogether, on which m+k-1 B-Splines are based on. An equal number of parameters then has to be estimated. Given a response vector $y = (y_1, ...y_N)$ and covariate $x = (x_1, ...x_N)$, $\sum_{i=1}^{N}(y_i - s(x_i))^2$ has to be minimized. From (3.54), one gets:

$$\sum_{i=1}^{N}(y_i - \sum_{i=1}^{m+k-1} \beta_i B_i^k(x))^2 \longrightarrow \min \tag{3.55}$$

The coefficients β_i can be interpreted as the amplitutes of the single B-Splines.

Example 10 *Figure 3.6 shows the individual B-Splines already scaled by their corresponding coefficients. The final spline curve is then generated as the sum of the scaled B-Splines (Figure 3.7).*

Obviously, the final shape of a spline curve highly depends on the number and position of the knots. A small number of knots leads to a smooth curve, but the data may be underfitted, whereas too many knots lead to a overfitted, rough curve. But since the choice on the number of knots is restricted to a set of whole numbers, the smoothness of the function can only be controlled in a limited way. It is common, to choose equidistant knots over the range of the data or simply to take the quantiles. To some extent the choice is arbitrary in a kind.

Naturally, the coefficients themselves also regulate the roughness.The more adjacent β_i differ from each other, the rougher the curve is. This leads to the idea of P-Splines.

3.3.3 Penalized B-Splines

Eilers and Marx [1996] proposed to overfit the data with a relatively large number of knots but to restrict the high variation of the curve by using a difference penalty on coefficients of adjacent B-Splines. Consider the regression of N data points (x_j, y_j) on a set of $r = m + k - 1$ B-Splines B_i . As a penalty term, the integral of the squared second derivative of the form

$$\lambda \int_{x_{\min}}^{x_{\max}} \left[\sum_{i=1}^{r} \beta_i B_j'' (x) \right]^2 dx \tag{3.56}$$

can be used. The parameter λ controlls the smoothness of the function continuously, therefore called smoothing parameter.

Since the minimalization with this term is numerically complicated, it is approximated by a simple difference penalty based on the lth differences (Δ^l) of adjacent B-Spline-coefficients:

$$\lambda \sum_{i=l+1}^{r} \left(\Delta^l \beta_i \right)^2 \tag{3.57}$$

Thus, the following term should be minimized:

$$\sum_{i=1}^{N} (y_i - \sum_{i=1}^{r} \beta_i B_i^k(x))^2 + \lambda \sum_{i=l+1}^{r} \left(\Delta^l \beta_i \right)^2 \tag{3.58}$$

The substitution of the integrated square of the l-th derivative with the corresponding difference reduces the dimensionality of this problem from the number of observations N to the number of B-Splines r. Other features of Penalized B-Splines, also called P-Splines are:

- P-Splines can fit polynomial data exactly. If the response variable y is a polynomial in x of degree k, then P-Splines of degree k or higher will exactly fit the data for any λ.
- P-Splines conserve the moments of the data
- The limit of a P-Splines fit of degree k or higher with a large smoothing parameter is a polynomial of degree k-1

3.3.4 Estimation of P-Splines in GAM's

In generalized additive models, the effects of a covariate x on the linear predictor η is modelled with a smooth function $f(x)$. By substituting $f(x)$ with a penalized P-Spline $s(x)$ one gets the following model equation [Lang and Brezger(2001)] with (3.54):

$$g(\mu_j) = \eta_j = f(x_j) = \sum_{i=1}^{r} \beta_i B_i^k(x_j) \tag{3.59}$$

Estimation in this model is done via maximum likelihood method. The penalization term (3.57) is subtracted from the log-likelihood

$$l(y, \beta) = \prod_{j=1}^{N} \left(\frac{y_j \theta_j - b(\theta_j)}{\phi} \right), \tag{3.60}$$

so that the Penalized Likelihood PL is given by

$$PL = l(y, \beta) - \lambda \sum_{i=l+1}^{r} \left(\Delta^l \beta_i \right)^2. \tag{3.61}$$

Fisher-Scoring-Algorithm is used to conduct the maximization on the Penalized Likelihood with respect to the unknown regression coefficients. An extension to more than one smoothing spline is straightforward, i.e.

$$PL = l(y, \beta_1, ...\beta_p) - \lambda_1 \sum_{i=l+1}^{r} \left(\Delta^l \beta_{1i} \right)^2 - ... - \lambda_p \sum_{i=l+1}^{r} \left(\Delta^l \beta_{pi} \right)^2. \tag{3.62}$$

The crucial point is to find a compromise between flexibility and smoothness, regulated by the smoothing parameters $\lambda_j, j = 1, ...p$.

3.3.5 Choosing the smoothing parameter

As the smoothing parameter λ influences the smoothness of the function, a way has to be found to choose an optimal value for λ. The method recommended by Eilers and Marx is to minimizing the Akaike information criterion

(AIC). The deviance is corrected for the effective number of parameters. In this case, the AIC is defined as

$$AIC(\lambda) = dev(y; \beta, \lambda) + 2 * \dim(\beta, \lambda) \tag{3.63}$$

where $\dim(\beta, \lambda)$ is the dimension of the vector of parameters β. More information on the deviance and dimension can be found in [Eilers and Marx(1996)] and [Hastie and Tibshirani(1990)]. The computation of AIC's for many values of λ is very time-consuming and becomes very unpracticable in higher dimensions. Furthermore, the function $AIC(\lambda)$ doesn't need to have a global minimum. On the contrary, it often has several local minima, which makes it difficult to decide on one optimal λ value. It has been shown [Brezger(2000)], that even in cases of a unique minimum, the choice of λ is not optimal, but produces a curve, that is too rough. Alternatives to AIC are cross-validation methods [Fahrmeir and Tutz(2001)], defined as

$$GVC = \frac{1}{N} \sum_{j=1}^{N} \left\{ y_j - \widehat{f_\lambda^{-j}}(x_j) \right\}^2 \tag{3.64}$$

where $\widehat{f_\lambda^{-j}}(x_j)$ is the solution of the penalized least-squares estimation problem (3.61) after leaving out the jth data point. But like the AIC's, computation is not practicable and becomes intractable for a high number of smooth functions. Methods, which can avoid the costly and sometimes arbitrary hunt for an optimal λ would be highly desirable.

The Bayesian version of P-Splines is such a method, since the smoothing parameters are considered as random and estimated simultaneously together with other model parameters.

3.4 MCMC methods

The Bayesian principle has already been mentioned in (3.1.3). Estimates for fixed effects have been calculated by ML- or REML methods and then, empirical Bayes methods were used to get estimates for the random effects. In general, classical frequentistic statistics assumes parameters to be unknown but fixed, whereas in the Bayesian context, all parameters are specified as

random variables with a prior distribution. Since the calculation of the posterior distribution obtained by using Bayes' theorem is computationally infeasible in most cases, Markov chain Monte Carlo (MCMC) methods are used for simulation. The popularity of Bayesian methods originates from the flexibility of Bayesian modelling, e.g. extending GLM's to more complex problems, such as ordinal mixed models or semiparametric GLM's. Modern computer techniques have made it possible to apply such methods to a brought variety of models. It has led to a burst of activity in the last decade. In the following, the basics of Bayesian statistics, Markov chains and MCMC simulation methods are described. For more details refer to [Brezger(2000)] and [Gilks et al.(1996)]. The initial motivation of introducing MCMC methods was the flexibility of this method with respect to extensions like Mixed models, ordinal response variables and/or generalized additive models. The embedding of all these models into the Bayesian framework closes this chapter on statistical methods.

3.4.1 Bayes-statistics

Let θ be a vector of p model parameters $\theta = (\beta_1, ...\beta_p)$. In Bayesian statistics, these parameters are assumed to be random variables, that have to be provided with a distribution assumption. Previous knowledge of the parameters, if available, can be used to define this prior distribution. Thus, a Bayesian model contains two components:

1. Priori assumption about the unknown parameters θ, indicated by $P(\theta)$
2. The likelihood of the data given the parameters, denoted by $P(y|\theta)$

In the case of a GLM, where the structural assumption is given by

$$\eta_i = X_{i1}\beta_1 + ... + X_{ip}\beta_p \tag{3.65}$$

and

$$E(Y_i) = \mu_i = h(\eta_i), \tag{3.66}$$

these components can be written as in the following:

1. Often, the prior assumption for parameters is given by the multivariate normal distribution

$$\beta \sim N(b_0, \Sigma_0) = N(b_0, K_0^{-1}), \tag{3.67}$$

where K_0 is called *precision matrix*. In the limiting case $\Sigma_0 \longrightarrow \infty$, a so called *diffuse priori*

$$\beta \propto const. \tag{3.68}$$

is obtained, where each value has the same probability. This priori is used in cases, where there is no previous knowledge about the parameters.

2. The distributional assumption of the response y conditional on the parameters is given by

$$P(y|\beta) = \prod_{i=1}^{N} P(y_i|\beta) \propto \prod_{i=1}^{N} L_i(\beta), \tag{3.69}$$

where $L_i(\beta)$ denotes the likelihood. The log-likelihood for an exponential family was already given in (3.60), so that

$$P(y|\beta) \propto \prod_{j=1}^{N} \exp\left(\frac{y_j\theta_j - b(\theta_j)}{\phi}\right). \tag{3.70}$$

If the priori distribution of the parameters and the likelihood is given, the posterior distribution can be obtained via Bayes' Rule as

$$P(\theta|y) = \frac{P(y|\theta)\ P(\theta)}{\int\ P(y|\theta)\ P(\theta)\ d\theta}. \tag{3.71}$$

Note, that the posterior is proportional to the prior times the likelihood:

$$P(\theta|y) \propto P(y|\theta)P(\theta). \tag{3.72}$$

The posterior can therefore be interpreted as a weighted average of the likelihood and the prior. But even in cases, where no previous information on

the parameters is available, a Bayesian approach is justifiable, when using a diffuse priori as in (3.68). Then, (3.72) reduces to

$$P(\theta|y) \propto P(y|\theta). \tag{3.73}$$

The posterior distribution is used for estimation and inference about the parameters. The idea is to draw random numbers from the distribution and calculate all summary statistics of interest out of the random sample. For example, the mean can be estimated by the sample mean, modus or median. It follows from (3.73), that the estimate obtained by using the modus from the posterior distribution, corresponds to a classical ML-estimator in case of a diffuse prior.

If the posterior distribution was given, then Bayesian inference would be easy and straightforward. But in most cases it is intractable and MCMC methods have to be used to draw random samples out of the posterior. For the statistical user, there are two basic methods of MCMC: The *Metropolis-Hastings-Algorithm* and, as a special case of it, the *Gibbs sampler* . The basic idea is to construct a Markov chain, whose realisations can be taken as random samples of the interesting distribution. Hence, in the following, the basic terms of Markov chains will be clarified.

3.4.2 Markov chains

A Markov chain is a sequence of random variables $\left\{X^{(0)}, X^{(1)}, X^{(2)}...\right\}$ with state space A, having the property that, given the present, the future is conditionally independent of the past. In other words:

$$\begin{aligned} P(X^{(n+1)} &\in A|X^{(n)} = x, X^{(n-1)} \in A^{(n-1)}, ..., X^{(0)} \in A^{(0)}) \\ &= P(X^{(n+1)} \in A|X^{(n)} = x). \end{aligned} \tag{3.74}$$

$P(.|.)$ is called the transition kernel of the chain and will be assumed to be homogeneous: that is, $P(.|.)$ does not depend on n. Hence, the chain will gradually "forget" its initial state and $P(.|.)$ will converge to a steady state, which is known as the *stationary distribution*. The rate at which a Markov chain converges to a stationary distribution is called the *mixing rate*. Denote the stationary distribution by $\pi^*(.)$. For increasing n, the realisations $\left\{X^{(n)}\right\}$ will look more and more like samples from $\pi^*(.)$. Thus, one way of drawing

samples from a posterior distribution is to construct a Markov chain having the posterior as its stationary distribution $\pi^*(.)$. After a so called *burn-in period*, the realisations of the Markov chain can be accepted as samples of the interesting distribution.

3.4.3 The Metropolis-Hastings-Algorithm

If the interesting stationary distribution $\pi^*(\theta)$ is not easily accessible, a so called *proposal density* $q(\theta^{(n-1)}, \theta^{(n)})$ has to be defined, where $\theta^{(n-1)}$ denotes the current state of the chain and $\theta^{(n)}$ a new one. How to find an appropiate proposal density is not be of interest here, but can be found in [Brezger(2000)]. At each iteration n, the next state $\theta^{(n)}$ is chosen by sampling a canditate point θ^{cand} from the proposal distribution and accepted with probability α. The acceptance probability is defined as follows:

$$\alpha(\theta^{(n-1)}, \theta^{(n)}) = \min\left(1, \frac{\pi(\theta^{(n)})q(\theta^{(n)}, \theta^{(n-1)})}{\pi(\theta^{(n-1)})q(\theta^{(n-1)}, \theta^{(n)})}\right), \tag{3.75}$$

where $\pi(\theta)$ denotes the density of $\pi^*(\theta)$.
Thus, the Algorithm can be implemented in the following steps:

1. *Set* $n = 0$*; Initialize starting value* $\theta^{(0)}$*;*
2. *Repeat as long* $n < N$ *:*
   a) *Sample a point* θ^{cand} *from* $q(\theta^{(n-1)}, \theta^{cand})$
   b) *Sample a variable* U*, that is distributed uniformly on* $(0, 1)$
   c) *If* $U \leq \alpha(\theta^{(n-1)}, \theta^{cand})$*, set* $\theta^{(n)} = \theta^{cand}$*,*
      otherwise, set $\theta^{(n)} = \theta^{(n-1)}$
   d) *Set* $n = n + 1$

In many cases, it is not advisable, to update all interesting parameters in one step, but to reduce it to a partial vector $\theta_k \in \theta$, $k = 1, ...p$. In GLM's or GAM's, for example, it might be necessary to assume different prior distributions for parameters. Therefore, θ is decomposed into the subvectors $\theta_1, ...\theta_p$. Random samples can be drawn from the so called *full conditionals* $\pi(\theta_k|\theta_{j\neq k})$, $j = 1, ...p$ with means of the proposal densities $q(\theta_k^{(n-1)}, \theta_k^{(n)}|\theta_{j\neq k})$. The vector $\theta_{j\neq k}$ contains all parameters, with the exception of θ_k, that is $\theta_{j\neq k} = \{\theta_1, ..., \theta_{k-1}, \theta_{k+1}, ..., \theta_p\}$. The steps of this *single-component Metropolis-Hastings-Algorithm* can be implemented just like before, but for each θ_k separately and conditional on $\theta_{j\neq k}$. In many cases, a

blockwise updating scheme is preferable to a single step algorithm. In this case, blocks of parameters updated jointly instead of single parameters.

A special case of the single-component or block-component Metropolis-Hastings-Algorithm is the *Gibbs sampler.* It is used, if the full conditional posterior distributions are known and random variables can be drawn easily from them. Thus, the proposal density is replaced by the full conditional itself:

$$q(\theta_k^{(n-1)}, \theta_k^{cand} | \theta_{j \neq k}^{(n)}) = \pi(\theta_k^{cand} | \theta_{j \neq k}^{(n)}). \tag{3.76}$$

Substituting (3.76) into (3.75) gives an acceptance probability of 1. That is, Gibbs sampler estimates are always accepted. Samples are drawn out of the full conditional distributions similar to the Metropolis-Hastings-Algorithm to obtain estimates for the interesting parameters.

3.4.4 Exploration of Markov chains

As already mentioned, the realisations of a Markov chain can be accepted as samples of the stationary distribution after a number of iterations. By looking at trace plots of the samples, it is possible to assess the width of the *burn-in period.* Naturally, the first iterations still show a trend or oscillate, before convergence is reached. But too few variation in the samples is not good, either, as all regions of the stationary distribution should be reached. Ideal samples are scattered around the expected value of the distribution with irregular variation. Figure 3.8 illustrates the following situation: The trend of the samples ascends in the beginning, until convergence is reached. The samples are scattered widely around the expected value. An estimated sample mean would come very close to the expected mean, provided, the samples at the beginning are omitted. The decision about the burn-in phase is done visually. In this case, convergence is reached around the 100th iteration. In practice, the burn-in phase is chosen much higher, just to be on the safe side.

Another consideration to be made is whether all samples are to be used for estimation. It follows from the construction of Markov chains, that the values are not independent. Estimation from a dependent sample requires a higher number of samples, as if one had an independent sample. To avoid memory problems, it is recommendable, not to safe all samples, but only every 10th,

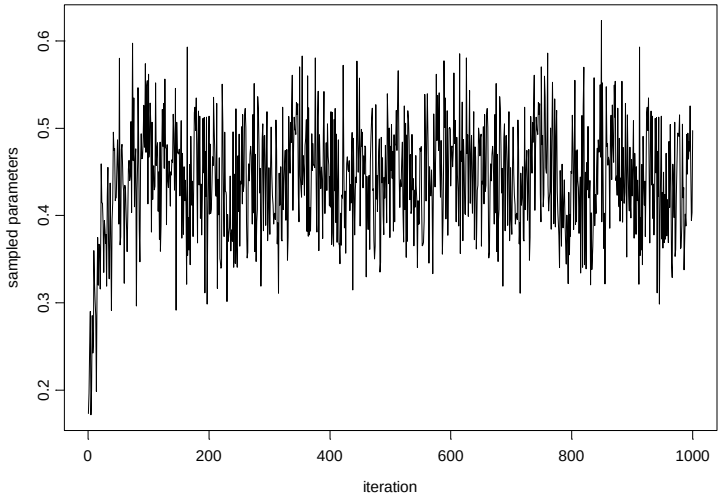

Figure 3.8: Trace plot of samples to assess burn-in period

20th or 100th. The choice of the increment depends on the autocorrelation of the sampled values. An autocorrelation close to 0 is desirable. The difference between two increment values, also called step width, is shown in Figures 3.9 and 3.10: The first one shows the autocorrelation for a sample, where each value is used. The line descends under the critical line indicating a significant autocorrelation at the 30th iteration and reaches another dip at iteration 50. That means, that autocorrelation between pairs of values, that are 50 iterations apart, is not significant. Hence, 50 has been chosen as the step width and thus, only every 50th sampled value is stored and used for estimation. Figure 3.10 shows the autocorrelation on the same data with the same prior distribution but using a step width of 50.

In practice, a small number of iterations is used to assess the burn-in phase by means of trace plots of interesting parameters. The decision about the increment is done by choosing the step width equal to 0 in the beginning and assess the ideal step width visually by looking at autocorrelation plots. Then, using the information obtained before, a new simulation with a higher number of iterations can be started, as in Example 11.

Example 11 *A dataset with 2 covariates* $x_1 \sim N(0, 0.1)$ *and* $x_2 \sim U(-1, 1)$ *is created, that define the response variable* y *according to*

$$y = 2 * x_1 - 0.5 * x_2.$$

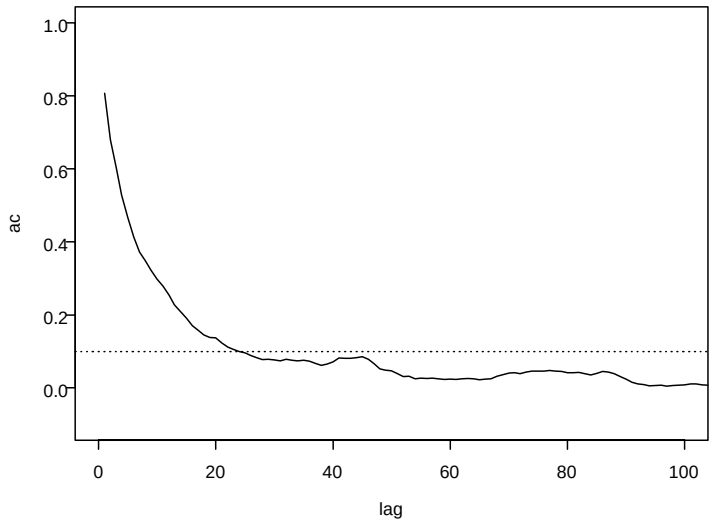

Figure 3.9: Autocorrelation between samples (step width=0)

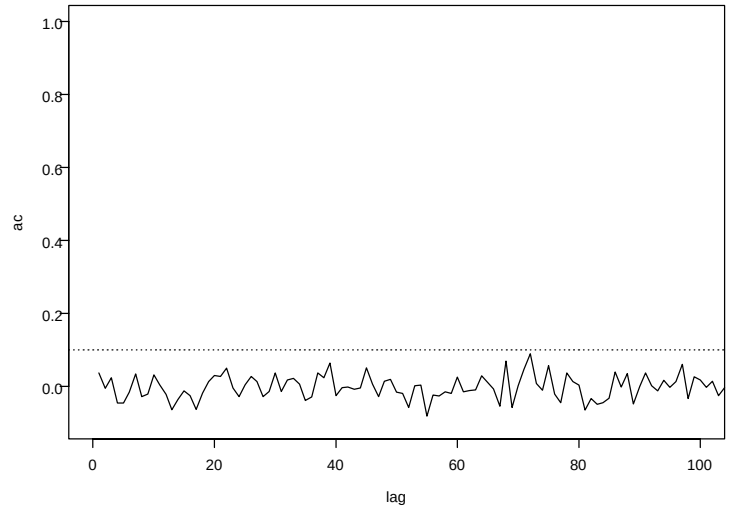

Figure 3.10: Autocorrelation between samples (step width=50)

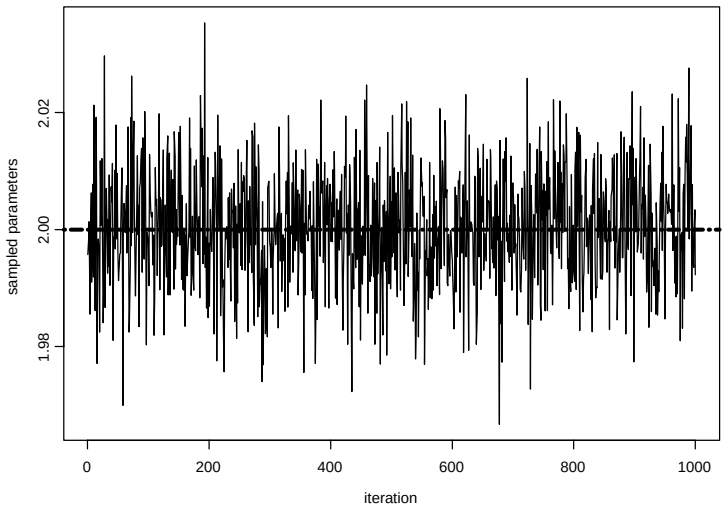

Figure 3.11: Trace plots for β_1 (Example 11)

Consider the parameters β_1 and β_2 as unknown, so that the linear regression $y = \beta_1 * x_1 + \beta_2 * x_2$ is obtained. Prior distributions for the parameters are diffuse. Visual diagnostic of trace and autocorrelation plots suggested a burn-in of 500 and step width of 20. At the next step, a simulation with 2500 iterations was started, so that, after subtracting the burn-in phase and considering, that only each 20th value is used, 1000 sampes are left for estimation. Then, the samples are scattered widely around the true parameter values of 2 (see Figure 3.11) and -0.5 (see Figure 3.12). Furthermore, autocorrelation is close to 0 (see Figure 3.13).

3.4.5 Prior assumptions for more complex models

The final motivation of introducing MCMC methods, was the simplicity of including complex structures, like random effects, regression splines and an ordinal response variable. This is due to the property, that everythying, that can be provided with a distribution can be estimated. In the following, the models described in 3.1, 3.2 and 3.3 are carried forward into the Bayesian context. See [Lang and Brezger(2001)] and [Fahrmeir and Lang(2001)] for reference on the following paragraphs.

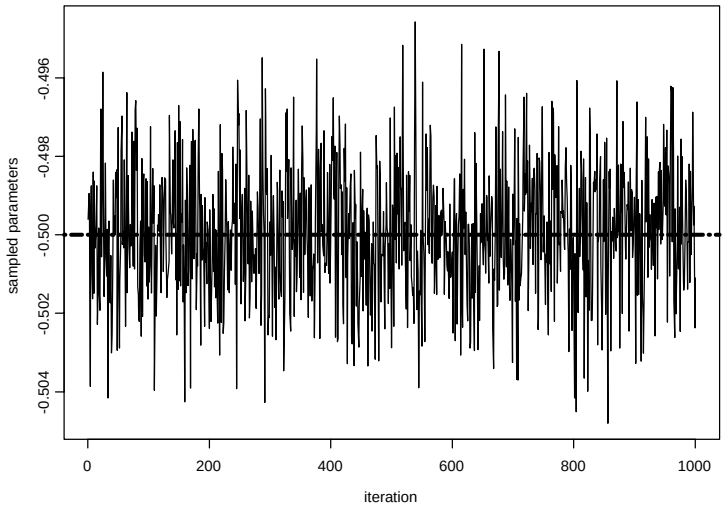

Figure 3.12: Trace plots for β_2 (Example 11)

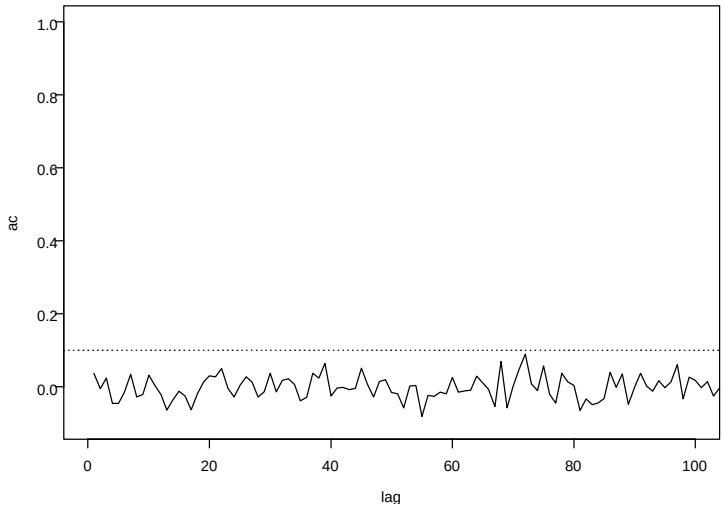

Figure 3.13: Autocorrelation for Example 11

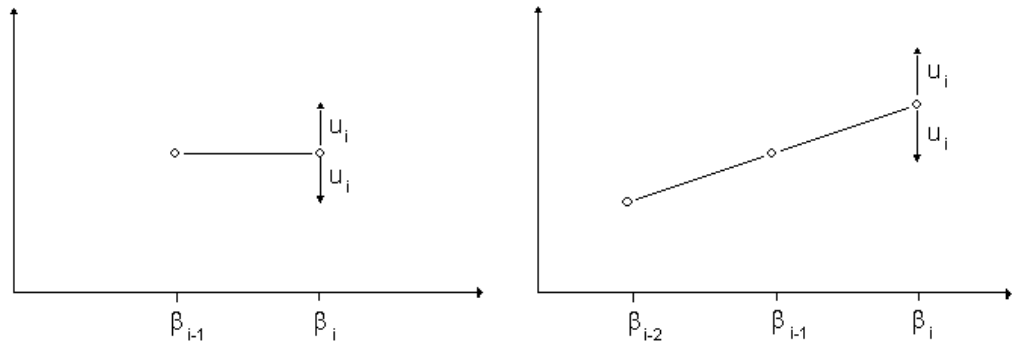

Figure 3.14: Prior distribution for RW(1) (left) and RW(2) (right)

Bayesian P-Splines

For simplicity reasons, it is assumed, that there is only one P-Spline ($p = 1$). Extensions to $p > 1$ are straightforward. Recall, that classical P-Splines are based on the lth differences (Δ^l) of adjacent B-Spline-coefficients $\lambda \sum_{i=l+1}^{r} \left(\Delta^l \beta_i\right)^2$. The unknown parameters $\beta_i, i = 1, ...r$ are considered as random variables and therefore, a distribution has to be specified. The difference penalties are replaced by their stochastic analogues, that is, random walks. First order random walks (RW(1)) correspond to first differences and are specified as follows:

$$\begin{aligned} \text{RW(1)} \quad : \quad & \beta_i = \beta_{i-1} + u_i, \\ u_i \quad \sim \quad & N(0, \tau^2) \text{ and } \beta_1 \propto const. \end{aligned} \tag{3.77}$$

Another possibility are second order random walks (RW(2)). Likewise, RW(2) corresponds to second differences in classical P-Splines. Note, that the prioris for the initial values, β_1 for RW(1), and $\beta_1$and β_2 for RW(2) respectively, are diffuse, that is:

$$\begin{aligned} \text{RW(2)} \quad : \quad & \beta_i = 2 * \beta_{i-1} + \beta_{j-2} + u_i, \\ u_i \quad \sim \quad & N(0, \tau^2), \beta_1 \propto const \text{ and } \beta_2 \propto const. \end{aligned} \tag{3.78}$$

The illustration of this concept in Figure 3.14 shows that u_i, or equivalently the variance parameter $Var(u_i) = \tau^2$, regulates the smoothness of the function. The coefficient β_i is restricted to deviate at most by u_i from the preceding coefficient β_{i-1}, or alternatively from the interpolating line between β_{i-2} and β_{i-1}, in the case of a second order random walk.

It can be shown [Brezger(2000)], that the joint distribution of the prior is given by

$$\beta \propto \exp(-\frac{1}{2\tau^2}\beta' K \beta) \tag{3.79}$$

with the symmetric penalty matrix K. In the case of a first and second order random walk, K has the form

$$\text{RW(1):} \quad K = \begin{pmatrix} 1 & -1 & 0 & \cdots & \cdots & \cdots & 0 \\ -1 & 2 & -1 & 0 & & & \vdots \\ 0 & -1 & 2 & -1 & 0 & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & 0 & -1 & 2 & -1 & 0 \\ \vdots & & & 0 & -1 & 2 & -1 \\ 0 & \cdots & \cdots & \cdots & 0 & -1 & 1 \end{pmatrix}, \tag{3.80}$$

$$\text{RW(2):} \quad K = \begin{pmatrix} 1 & -2 & 1 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ -2 & 5 & -4 & 1 & 0 & & & & \vdots \\ 1 & -4 & 6 & -4 & 1 & 0 & & & \vdots \\ 0 & 1 & -4 & 6 & -4 & 1 & 0 & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & 0 & 1 & -4 & 6 & -4 & 1 & 0 \\ \vdots & & & 0 & 1 & -4 & 6 & -4 & 1 \\ \vdots & & & & 0 & 1 & -4 & 5 & -2 \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 & 1 & -2 & 1 \end{pmatrix}. \tag{3.81}$$

In addition to the coefficients, the variance parameter τ_i has to be supplemented with a prior distribution as well. Thus, this parameter is also assumed to be random and is estimated simultaneously with the coefficients

by means of the single-component Metropolis-Hastings-Algorithm. The advantage of this procedure is, that the problem of choosing a smoothing parameter is avoided. The variance parameter τ_i corresponds to the smoothing paramter λ in the classical approach of P-Splines, but it is data-driven and therefore more reliable than λ. The priori for the variance parameter, also called hyperparameter, is given by the inverse Gamma distribution

$$\tau^2 \sim IG(a, b). \tag{3.82}$$

By choosing $a = 1$ and $b = 0.005$ a flat distribution is obtained, that is approximately non-informative.

Unobserved heterogeneity

Suppose, that repeated measurements have been taken on n individuals and a mixed effects model is used. Thus, the random effects $b_i, i = 1, ..., n$ have to be submitted with an appropriate prior distribution. Assume, that the b_i´s are i.i.d. Gaussian, i.e.

$$b_i \sim N(0, \tau_{ra}^2). \tag{3.83}$$

Just like the hyperparamter in the random walk approach, the variance parameter $Var(b_i) = \tau_{ra}^2$ is assumed to be random. In this case, the Inverse Gamma distribution is taken as well, so that

$$\tau_{ra}^2 \sim IG(a_{ra}, b_{ra}) \text{ with } a_{ra} = 1 \text{ and } b_{ra} = 0.005. \tag{3.84}$$

Gaussian response

In the case of a Gaussian response variable $y_i \sim N(\mu_i, \sigma^2)$ an additional scale parameter σ^2 has to be estimated. In analogy to the variance parameters considered before an inverse Gamma distribution

$$\sigma^2 \sim IG(a_\sigma, b_\sigma) \tag{3.85}$$

is assigned. Like before, a_σ is set to 1 and b_σ to 0.005, to obtain an approximately non-informative distribution.

Ordinal response

Consider an ordinal response variable Y, that is assumed to be a categorized version of a latent variable $\widetilde{Y}$ with thresholds $-\infty = \theta_0 < \theta_1 < ... < \theta_k = \infty$. Thus, there are $k-1$ parameters to estimate in addition to the unknown coefficient parameters. The thresholds $\theta = (\theta_1, ..., \theta_{k-1})'$ are considered as random. Like fixed effects (see (3.68)), they are supplemented with diffuse priors, i.e.

$$p(\theta) \propto const. \tag{3.86}$$

3.4.6 Inference for more complex models

For inference in the Bayesian model, the following conditional independence assumptions have to be fulfilled:

1. The observations of the response y_{it} are conditionally independent for given covariates and parameters.
2. The prior distributions for fixed effects, random effects, function evaluations and for the variance parameters or hyperpriors are mutually independent.

Given these assumptions and the prior distributions, the posterior can be specified, on which estimation and inference is based on. Note, that it is very likely, that constant effects will be included into the model in addition to random effects and the parameters for a P-Spline. In the following, such a constant effect, supplemented with a diffuse priori (see (3.68)) is denoted with γ. For the ease of notation, only one constant effect, respectively P-Spline is assumed.

Gaussian response

It has already been noted in (3.72), that the posterior is proportional to the product of the likelihood and the prior. In the case of a Gaussian response variable, the joint distribution of all parameters is given by

$$\begin{aligned} P(y, \beta, \tau^2, \gamma, b, \tau_{ra}, \sigma^2) &= P(y|\beta, \tau^2, \gamma, b, \tau_{ra}^2, \sigma^2) P(\beta|\tau^2) \\ & \quad P(\tau^2) P(b|\tau_{ra}^2) P(\tau_{ra}^2) P(\gamma) P(\sigma^2). \end{aligned} \tag{3.87}$$

The full conditionals, that are of primary interest, can be derived out of this illustration easily, i.e.

$$\begin{aligned} P(\beta|\cdot) &\propto P(y|\beta,\tau^2,\gamma,b,\tau^2_{ra},\sigma^2)P(\beta|\tau^2) \\ P(\tau^2|\cdot) &\propto P(\beta|\tau^2)P(\tau^2) \\ P(\sigma^2|\cdot) &\propto P(y|\beta,\tau^2,\gamma,b,\tau^2_{ra},\sigma^2)P(\sigma^2) \\ P(\tau^2_{ra}|\cdot) &\propto P(b|\tau^2_{ra})P(\tau^2_{ra}) \\ P(b|\cdot) &\propto P(y|\beta,\tau^2,\gamma,b,\tau^2_{ra},\sigma^2)P(b|\tau^2_{ra}) \\ P(\gamma|\cdot) &\propto P(y|\beta,\tau^2,\gamma,b,\tau^2_{ra},\sigma^2)P(\gamma). \end{aligned} \tag{3.88}$$

It can be shown, that the full conditional distributions of β, b and γ are multivariate Gaussian, whereas the full conditionals of τ^2, τ^2_{ra} and σ^2 are all inverse Gamma distributions. Since all distributions are known, a simple Gibbs sampler can be used to update the parameters of the model either in single component steps or blockwise. It is reasonable to update the parameters β of function evaluations and the random effects jointly by block moves. A detailed updating algorithm and mean and variance parameters of the full conditionals can be found in [Lang and Brezger(2001)].

Ordinal response

The posterior of the model obtained by the ordinal threshold concept depends on the additional parameters $\theta = (\theta_1, ..., \theta_{k-1})'$. Defining this model by a latent variable $\widetilde{Y}$ has computational advantages. Set $\alpha = (\gamma, \theta)$ as the vector of fixed effects. Then, posterior analysis is based on

$$P(\widetilde{y},\beta,\tau^2,\alpha,b,\tau^2_{ra}|y) \propto P(y|\widetilde{y})P(\widetilde{y}|\beta,\alpha,b)P(\beta|\tau^2)P(\tau^2)P(\alpha)P(b|\tau^2_{ra})P(\tau^2_{ra}). \tag{3.89}$$

As $\widetilde{y}$ has to obey the constraint (3.23), the full conditionals of Y and $\widetilde{Y}$ are given by

$$P(Y|\widetilde{Y}) = \sum_{r=1}^{k} I(\theta_{r-1} < \widetilde{Y} \leq \theta_r) I(Y = r) \tag{3.90}$$

and

$$P(\widetilde{Y}|\cdot) = P(Y|\widetilde{Y})P(\widetilde{y}|\beta,\alpha,b) \tag{3.91}$$

The full conditional distribution of the latent variable is a truncated standard normal distribution, with truncation points determined by the restriction (3.90). Drawing out of a truncated normal distribution evolves as numerically difficult and almost not solvable together with random effects. Thus, the interested reader is referred to [Fahrmeir and Lang(2001)] for details on sampling schemes on ordinal data. Reparametrization strategies to overcome the numerical problems are also provided.

Chapter 4

Analysis of SLCMSR dataset

In the previous chapter, the theoretical background for estimating classical as well as Bayesian mixed-effect models has been provided, including extensions to generalized additive models and the possibility to use an ordinal response variable. Existing software for estimating such complex models is rare. Analysis has been carried out with BayesX, a tool for Bayesian inference based on MCMC techniques and MIXOR, a program for mixed-effects ordinal regression analysis.

4.1 Description of the data

The longitudinal dataset, available for analysis in spring 2002, contained data from 897 placebo patients with altogether 8716 observations. The number of observations ranged from 1 to 23 for each patient (see Figure 4.1). The observation dates are not equidistant and differ from patient to patient. This is due to the fact that the patients are recruited from different clinical trials and therefore present a high degree of heterogeneity. However, study protocols are responsible for regulating the intervals between two subsequent observations. It can be assumed, that observation intervals of 1 to 3 months have been set up for the data included in the dataset. Thus, there seems to be a rather homogeneous observation structure (see Figure 2.6). Nevertheless, the observation dates of all patients together are scattered over the whole range of the timeframe. In this analysis, time in weeks ("**time**") from the first observation has been used as a metric covariate.

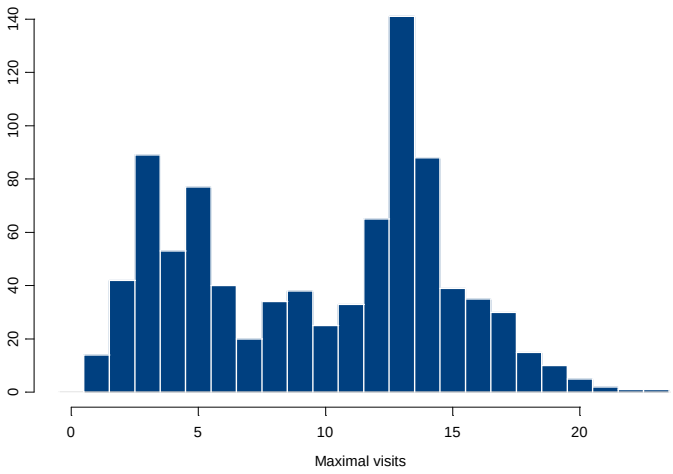

Figure 4.1: Number of maximal visits

Defining the response variable is crucial to each analysis. Since interest lies in the question, of how much a patients disability increases during the study period, the focus was centered on the change in EDSS-values. The absolute change from the first to the last observation, although this ignores all information in between, gives a brief overview of how much the disability of MS-patients increased or decreased within the period of the clinical trial (see Figure 4.2). The level of disability, measured by the EDSS, doesn't change for 34% of the patients. On the average (see Table 4.1), the EDSS increases by less than half a point. Thus, only a small change of disability can be expected, although the range lies between -3.5 and 7. Recalling, that MS is a chronic demyelinating disease it is surprising that decreases of disability were detected. This could be due to measurements during relapses that don't reflect the underlying level of disability. Theoretically, these cases should be ruled out, according to study protocol and relapses should be identifiable. Furthermore, changes in living conditions and climate can be responsible for fluctuations. Due to these heterogeneity, it is not advisable to take the absolute change from first to last observation as an outcome measure, but to look at the vector of changes over time. The decision to take the change in EDSS at all may be criticized, but measures have been taken to ensure that the variable "**change**" means the same thing for each patient. There is a broad concensus within experts, that, when higher values of EDSS

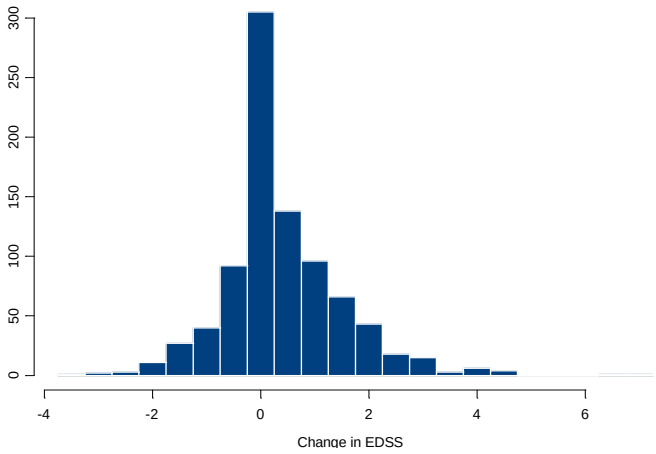

Figure 4.2: Histogram of change in EDSS

are already reached, changes have higher clinical significance. Since higher EDSS-values are dominated by ambulation, a change of 1 from e.g. EDSS 7 to 8 is more severe than from 1 to 2, where only a slight increase in one of the functional scores is needed. The European Agency for the Evaluation of Medicinal Products stated that "Based on EDSS values, treatment failure or progression should be predefined e.g. as the achievement of a specified degree of disability or of a sustained worsening of relevant magnitude (1 point when EDSS scores $\leq$ 0.5; 0.5 points if baseline score is $>$ 5.5)." [EMEA(2001)]. Corresponding to this regulation, changes in EDSS-values higher than 5.5 have been weighted twice as much than changes below this level. Note, that this weighted change ("**changew**") is a measure of severeness in changes of disability and cannot be related to the original EDSS-values any more. At a first attempt, this weighted change, although an ordinal variable, has been assumed to be metric, since there are 25 ordered categories, ranging from -3.5 to 9.5.

min	1. Qu.	median	mean	3.Qu.	max	std
-3.5	0.0	0.0	0.39	1.0	7.0	1.117

Table 4.1: Descriptive statistics for change in EDSS

Variable	Coding
i	i=1,...N: index variable identifying the patient
changew	weigthed change in EDSS from first observation
t	time in weeks from first observation
edss	EDSS at first observation
age	age at disease onset
dur	duration in months from onset to first observation
gender	$= \begin{cases} 0 \text{ for female} \\ 1 \text{ for male} \end{cases}$
course	$\text{course}^{(1)} = \begin{cases} 1 \text{ if course} = \text{pp or pr} \\ 0 \text{ otherwise} \end{cases}$ $\text{course}^{(2)} = \begin{cases} 1 \text{ if course} = \text{sp} \\ 0 \text{ otherwise} \end{cases}$ Reference category is relapsing-remitting

Table 4.2: List of variables used

In addition to time from first observation, several baseline variables were available to include as covariates. Table 4.2 gives an overview of all variables used and describes the coding, if necessary. The EDSS-values at the first observation ("**edssentry**") range from 0 to 8 (see Figure 2.2). It can be assumed, that the EDSS-value at first observation, also called baseline EDSS, reflects the goal of the clinical trial. That is, in studies designed for relapsing-remitting patients, the entry EDSS is rather low, whereas it is higher for patients with a progressive course (see Figure 4.3).

To account for the different disease courses ("**course**"), this categorization was also included in the analysis. As there are only 9 primary-progressive patients, they were put together with the progressive-relapsing (n=115). Relapsing-remitting (n=368) and secondary-progressive patients (n=338) make up the majority. The underrepresentation of primary-progressive patients ($\sim$ 15% in the MS-population) is due to the fact, that most treatments tested in clinical studies so far, were aimed at relapsing-remitting or secondary-progressive patients.

The distribution of women (n=612) to men (n=285) ("**gender**") is approximately 2:1 and is therefore consistent with observations in the whole population of MS-patients.

The average age of onset is known from former studies to be about

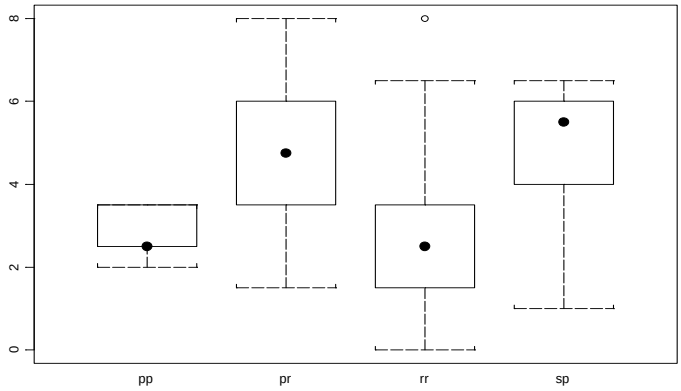

Figure 4.3: Distribution of baseline EDSS in the disease courses primary-progressiv (pp), progressive-relapsing (pr), relapsing-remitting (rr) and secondary-progressive (sp)

34 years. Thus, the SLCMSR dataset represents a younger patient collective with mean of 29.1, median of 28.6 and standard deviation of 7.77 ("**ageonset**"; see Figure 4.4).

Furthermore, the duration of the disease from onset to entry of the study is known ("**duration**"). This variable is given in months and ranges from 0 to 492 (see Figure 2.3), although a duration of 0 months is questionable. Most study protocols require a duration of at least 1 or 2 years. Naturally, for most patients with a duration less than one year, the course of the disease is not known, since this can only be classified retrospectively. Notice, that the date of disease onset is not always clear, as some studies take the first symptom as onset and others the first diagnosis.

4.2 Analyzing a Gaussian random intercept model

4.2.1 Formulation of the random intercept model

In the following, the influence of the covariates on the change in EDSS is going to be estimated with Bayesian techniques. For the metric variables, P-Splines of degree 3 and a second order random walk penalty were considered.

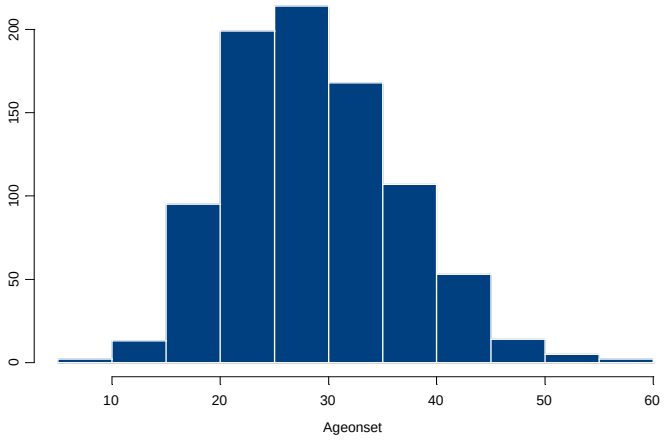

Figure 4.4: Histogram of age at onset

Thus, the model can be specified by the formula

$$\begin{aligned} changew_{it} &= f_1(t_i) + f_2(age_i) + f_3(edss_i) + f_4(dur_i) + \beta_1 * course_i^{(1)} \\ &\quad + \beta_2 * course_i^{(2)} + \beta_3 * gender_i + b_i + \varepsilon_{it}, \end{aligned} \tag{4.1}$$

where $b_i, i = 1, ...N$ is the random intercept, identified by an unique index variable. The functions f_i, $i = 1, ...4$, denote the P-Splines defined in section 3.3 with their Bayesian extension from 3.4.5. For the benefit of estimating a smooth function for time, a random slope term has been left out. Thus, possible non-linear effects of time may be detected. The introduction of a random slope would require a linear term for time. The model could therefore be displayed in the framework of a random intercept model, as in example 4 (section 3.1.1).

The prior distributions were chosen in the usual way, i.e. diffuse priors for the fixed effects β_1, β_2 and β_3 and Inverse Gamma distribution for the variance component of the random effect and the residual variance with $a = 1$ and $b = 0.005$ each. For the P-Splines, 20 equidistant knotes were chosen. The priori for the hyperparameter, that regulates the smoothness is also Inverse Gamma distributed with $\tau^2 \sim IG(1, 0.005)$.

Before looking at the parameter estimates, the convergence and mixing behavior of the MCMC procedure is of interest. Test runs with a small

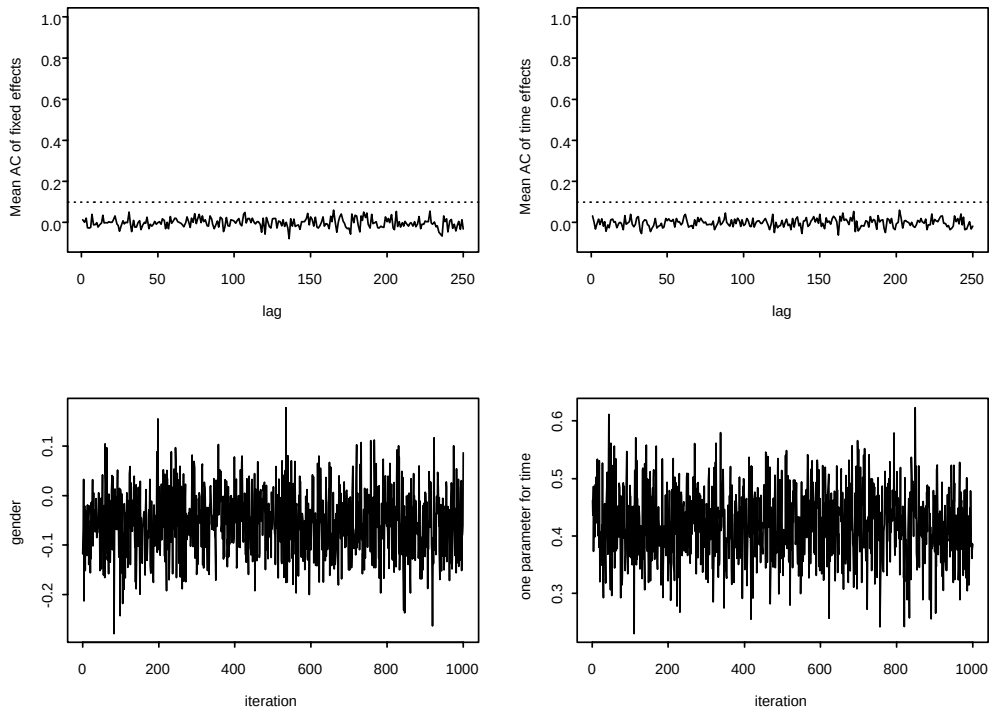

Figure 4.5: Autocorrelations of fixed effects and parameters for time (top) and mixing behavior of the estimate for gender and one time parameter (bottom)

number of iterations suggested taking a burn-in period of 20000 and step width 500. The number of iterations was therefore set to 520000, so that 1000 samples were stored. With these parameters, a good behavior of the chain was obtained. Figure 4.5 shows the sampling and autocorrelation plots of the constant effects and gender, and of one parameter for the time effect. All other autocorrelation and sampling plots are comparable to the examples showed.

source of variation	Mean	Std.Dev.	10% Qu.	50% Qu.	90% Qu.
scale	0.593373	0.009933	0.580887	0.593407	0.606655
intercept	0.536405	0.000929	0.498304	0.535726	0.573733

Table 4.3: Estimates of variance components

4.2.2 Results of the random intercept model

The estimates for the variance components, constant effects and P-Splines, obtained by the program BayesX, will be explained in the following. Since the response variable is the weighted change in EDSS, positive parameter estimates stand for a higher increase in disability, i.e. a worsening disease course. For all parameter estimates, 10 and 90% quantiles are provided. The 80% credible interval, that is defined by the quantiles, gives information on the accuracy of the estimation.

Variance components and random effect

The results for the variance components, i.e. the scale parameter, representing the "within-patients" variance, and the random effect, representing the "between-patients" variance, are shown in Table 4.3. It is remarkable that the random effect variance component is comparable to the scale parameter. Hence, introducing the random effect reduces the magnitude of the residual variance by a substantial part. That means, there is a high variation between the patients, that cannot be explained by the covariates. Ignoring the random-effects variance component could have led to significant results in the remaining effects that are not representative of the population.

It has been assumed, that the subject-specific random effect is normally distributed. Hence, the random effects have been provided with a Gaussian prior distribution. The posterior distribution of the estimates, representing a mixture of the prior distribution and the likelihood, is shown in Figure 4.6. It doesn't show a strong evidence of departures from normality, save for a couple of outliers.

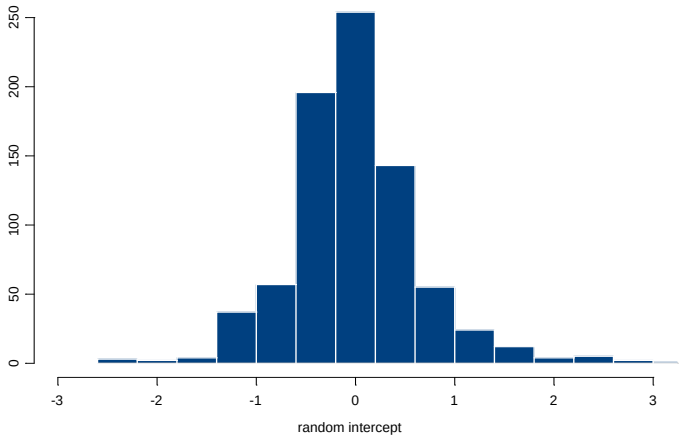

Figure 4.6: Histogram of random effects

Constant effects

Previous studies gave controversial answers on the question, whether gender is a prognostic factor on the disease course of MS (see [Compston et al.(1998)]). However, studies indicating an effect of gender, showed a benefit for women. Recall, that the reference category is defined as "female", so that a slight negative effect for the male patients can be shown in this model (see Table 4.4), indicating a better disease development as for women. But this effect can be neglected, since the magnitude is too small, compared to the variance.

In contrast to gender, an effect of the variable "course" can be seen. The effects of the primary- and relapsing-progressive courses (course$^{(1)}$), as well as the secondary-progressive course (course$^{(2)}$) are comparable, but higher than for the reference category, defined as relapsing-remitting. That is, patients categorized into one of the progressive forms of the disease, show a

Variable	Mean	Std.Var.	10% Qu.	50% Qu.	90% Qu.
gender	-0.053879	0.065302	-0.140182	-0.054003	0.031402
course$^{(1)}$	0.323480	0.104238	0.185648	0.321651	0.460414
course$^{(2)}$	0.339837	0.098472	0.210783	0.339599	0.469657

Table 4.4: Estimates of constant effects

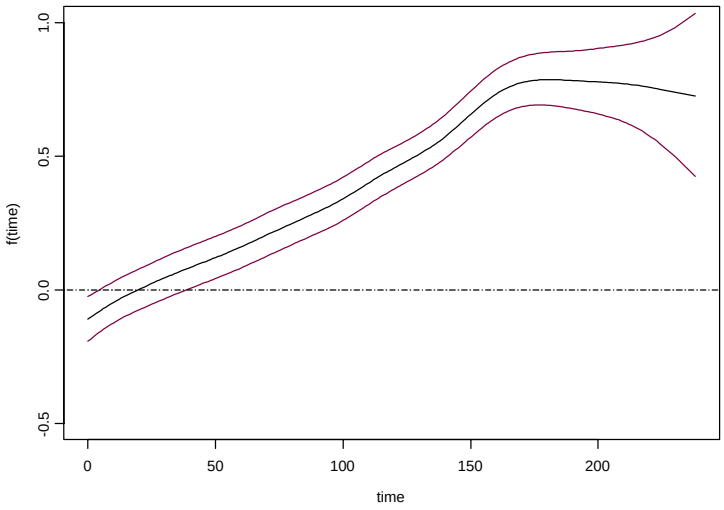

Figure 4.7: P-Spline for time (in weeks)

more severe course within the timeframe of a clinical study. Any other result would have been a big surprise, as a relapsing-remitting disease course is defined to be non-increasing in disability. But patients can only be categorized retrospectively by a physician. The course can change during the duration of the disease, which is almost always the case for relapsing-remitting patients. However, the categorization into one of the disease forms, seems to be predictive for the further short-term disease course. The interpretation of this variable may change, when looking at a longer timeframe.

Effect of time

The estimated effect of time on the change of EDSS is shown in Figure 4.7. Note, that the dashed line (as in the following spline curves) indicates "no effect". The plots always show the posterior mean together with the posterior 10 and 90 percent quantiles. In this case, a linear increasing trend is detectable up to approximately week 160 or 170. Beyond, the trend flattens. But there are not many observations left at the upper tail of the time distribution, so that the credible interval widens. It can also be assumed, that the behavior at the right tail is dominated by a minority of the clinical trials, since most studies were terminated after 2 or 3 years, which corresponds

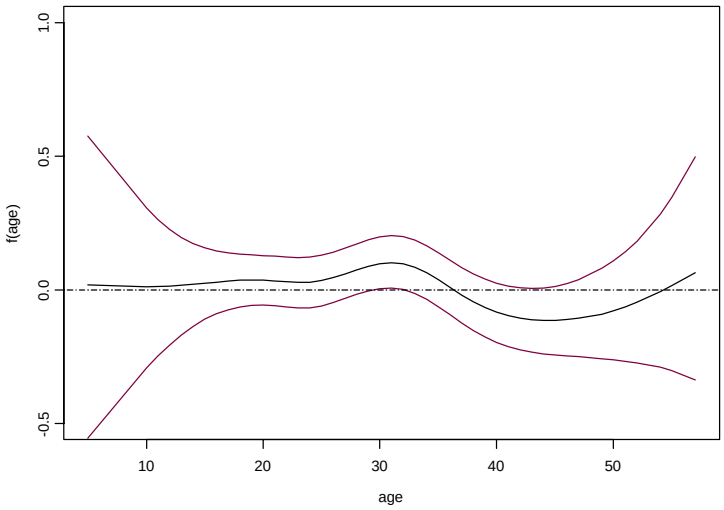

Figure 4.8: P-Spline for age at onset

to approximately 160 weeks. Overall you can say that, the more time that passes since the beginning of the studies, the more the disability increases.

Effect of age at onset

An influence of age on the change of disability couldn't be detected (see Figure 4.8). There seems to be an increasing trend around the age of 30, whereas there is a negative effect of higher ages. But this trend goes in the opposite direction again beyond approximately age 45. However, the credible interval includes the zero effect line over the whole range of the variable. It has been shown several times [Compston et al.(1998)], that age at onset indicates an important predictive factor. For patients, who are diagnosed at a higher age, the disease seems to develop more severe. In most cases a cut point was found to be at about 40 years. An explanation of the different behavior in this case could possibly be found by noticing that the meaning of "onset" is not homogeneous in different clinical trials. It is therefore possible, that the variable "age at onset" is defined slightly different depending on the study. Note, that there is not much information available for patients with very low or high onset age, causing the credible interval to widen at the tails of the P-Spline.

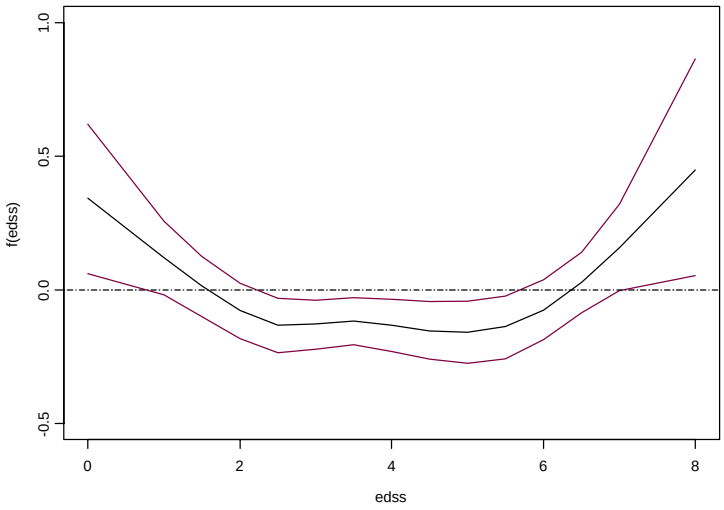

Figure 4.9: P-Spline for baseline EDSS

Effect of baseline EDSS

It has already been observed, that patients, who are in the middle range of disease stage, as indicated by the EDSS, are stable within a certain level of disability. This is due to the fact that the EDSS is not sensitive to other changes than ambulatory ones beyond EDSS 4. The P-Spline estimated in this model (Figure 4.9) is then consistent with previous results. Patients with an entry EDSS of 0 or 1 tend to increase more in their level of disability, than patients with EDSS 2 to 6. Minor changes in the functional systems can already cause substantial changes in the lower part of the EDSS. For EDSS bigger than 6.5, the spline curve increases again, meaning that changes in disability are lower. It has to be taken into account, that the maximum value of the EDSS is 10. Thus, changes from higher EDSS values are restricted by the characteristics of the score. But this result is based on a small number of patients only, so that credible intervals spread out for higher levels of the EDSS.

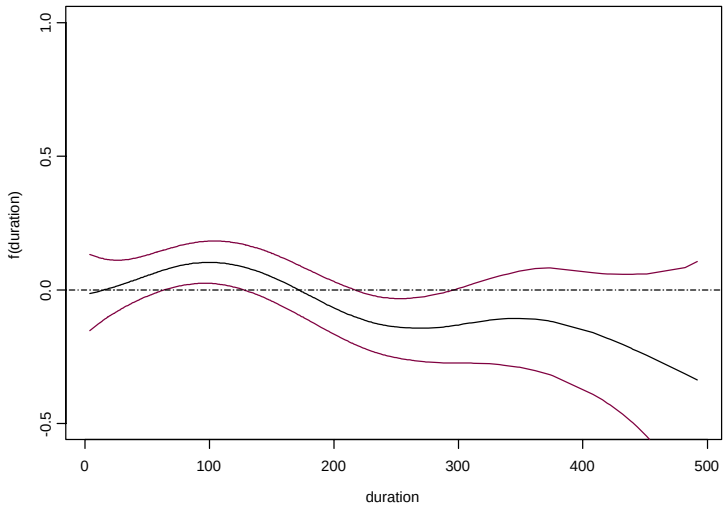

Figure 4.10: P-Spline for duration since onset

Effect of duration

By extrapolating the results of time from trial entry into the past, a positive linear trend could be expected, i.e. the longer the duration, the higher the change in EDSS. This trend couldn't be found in the P-Spline, displayed in Figure 4.10. The effect is positive up to a duration of 100 months, but it decreases from then on. However, in almost every part of the range of the variable, the credible interval overlaps the zero effect line. Furthermore, the wide credible interval at the upper tail of the duration distribution portend an unstable estimation due to a small number of observations. Recall, that the effects are assumed to be additive. That is, the disease duration cannot add much information to the variables already given before.

Conclusion

Overall, it has to be noted, that not all included effects influence the response variable significantly. Gender, age at onset and duration of the disease can be neglected. Just by looking at the range of the effects, time from first observation seems to have the biggest influence, followed by EDSS at first observation, course$^{(2)}$ and course$^{(1)}$ in ascending order. But, with the excep-

tion of the techniqual evaluation of MCMC samples, the model performance hasn't yet been checked. To answer the question, how well the model fits to the data, residuals have been calculated. In the case of a mixed effects model, there are two levels of residuals: Population residuals, obtained by looking at fixed effects only and individual residuals, that take the subject-specific random effects into account. The plot of the population residuals (Figure 4.11) shows a skewed distribution with a lot of negative outliers. That is, the fixed part of the predictor highly underestimates the observed outcome variable for many patients. The introduction of random effects causes a shrinkage towards zero. The resulting distribution of individual residuals is symmetric, but has flattend tails compared to a normal distribution. This is also reflected in the corresponding normal-quantile plot.

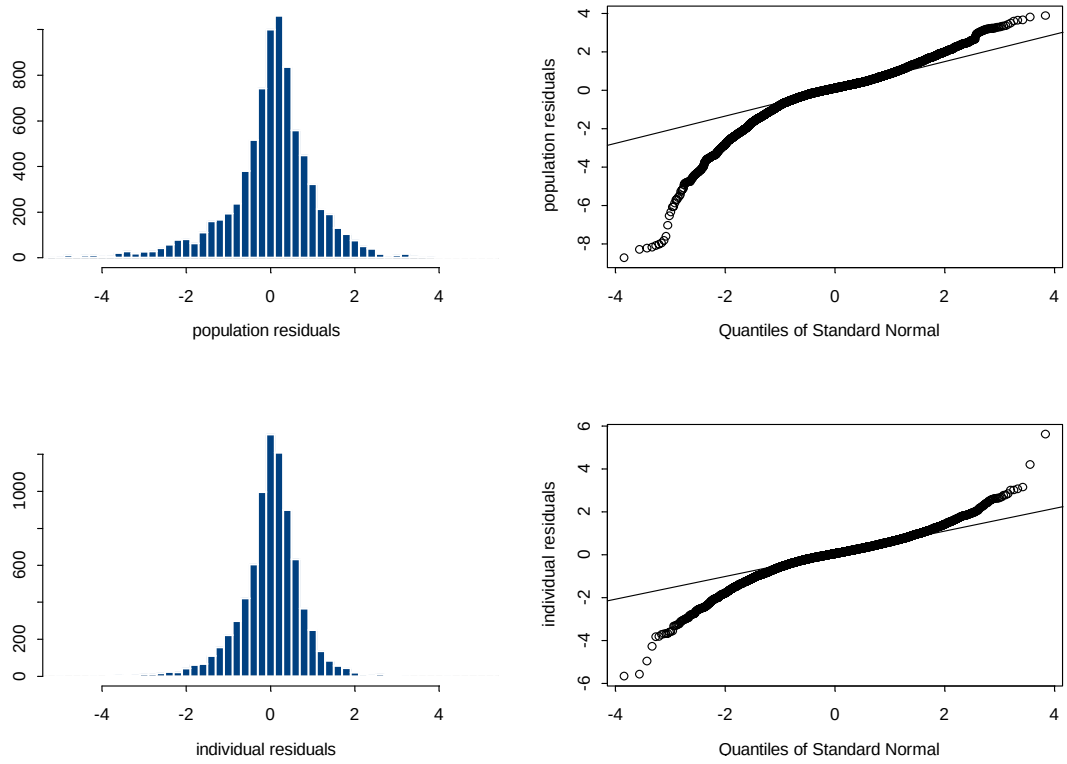

Figure 4.11: Histograms and normal-quantile plots for population residuals (top) and individual residuals (bottom)

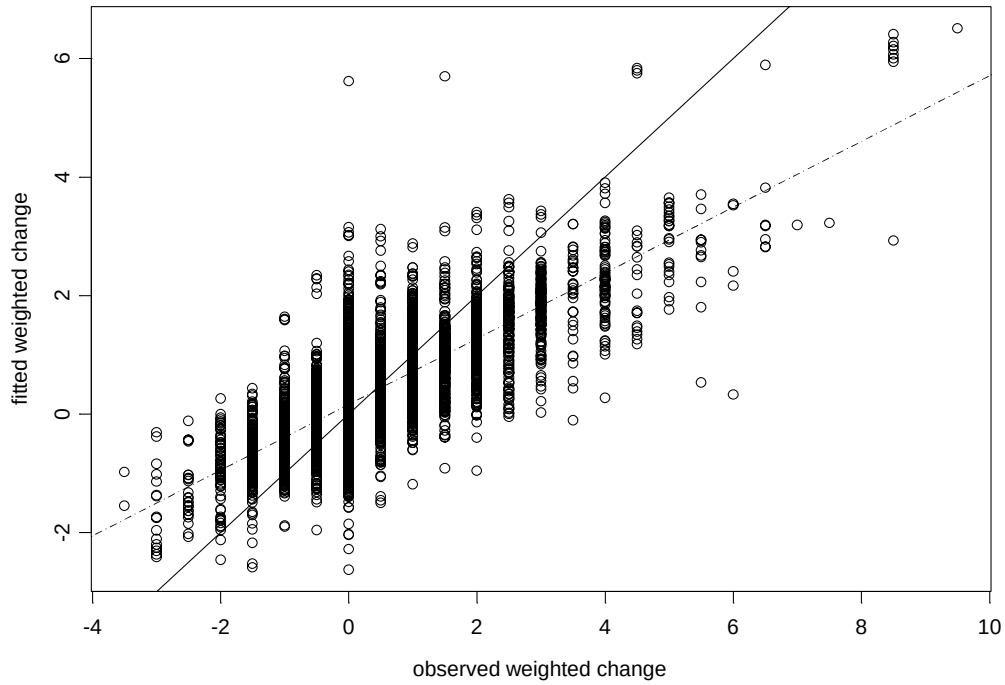

Figure 4.12: Plot of observed against fitted values (dashed line: linear regression line of the scatter plot; full line: the diagonal)

Plotting the observed weighted change against the fitted (Figure 4.12) shows a systematic trend: Negative values of the response are overestimated by the predictor, whereas positive values are underestimated. In general, the fitted values tend to be more conservative and estimations are shifted towards less change in EDSS.

It can be seen, that a high amount of variation is explained by the random effects. They are thought of accounting for variables, that haven't been observed or are not known. Including more explanatory variables in this model could reduce the big influence of the random effects and make it more stable and reliable for prediction.

weighted change	ordinal change	label	number of observations
≤ -2	big decrease	$<<$	126
$-1.5; -1$	small decrease	$<$	721
$-0.5; 0; 0.5$	stable	$=$	5438
$1; 1.5$	small increase	$>$	1440
≥ 2	big increase	$>>$	893

Table 4.5: Ordered change in EDSS

4.3 Analyzing an ordinal random intercept model

The change in EDSS, as well as the weighted change are ordinal variables. But so far, they have been assumed to be metric. Since it is not clear, if this assumption can be justified, the second part of the analysis took the ordinal structure of the data into account. Furthermore, it seems to be easier to interpret, especially, if predicting the outcome variable for new data is the goal. As it is numerically difficult, to handle an ordinal variable with 25 categories, some of them were combined to obtain a manageable number of categories. A new variable "changeord" with 5 ordered categories has been defined, ranging from a "big decrease" in disability over "stable" to a "big increase". The exact definition can be found in Table 4.5. Furthermore, the table gives the total number of observations in each of the categories.

This classification was motivated by results of Noseworthy et al [Noseworthy et al.(1990)]. The authors recommended, that at least a change of 1.0 on the EDSS-score is required to be confident of an important change in the degree of disability due to a big interrater variability. To ensure good interpretability, a symmetric classification was chosen. Other categorizations were considered, but didn't give better or significantly different results.

Analyzing the dataset with the recategorized ordered response has turned out to be more complicated than with a metric response. Estimation of an ordered mixed effect model was not possible in the program BayesX, that had been used before, due to numerical difficulties. Thus, the model was first estimated without random effects in BayesX. This approach doesn't account for the repeated measures structure in the data. Thus, the results obtained in BayesX were used as a template for estimating the model in another pro-

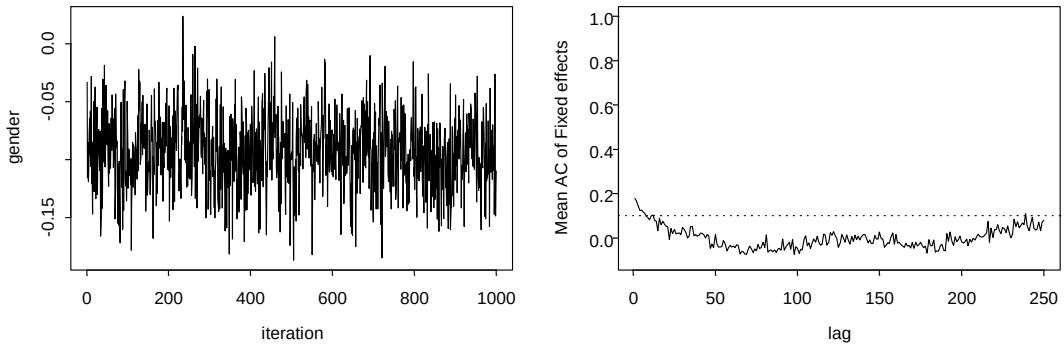

Figure 4.13: Sampling plot of gender (left) and autocorrelation plot of fixed effects (right)

gram. In this case, the program MIXOR was used. This program provides estimates for mixed-effects ordinal regression models. MIXOR uses marginal maximum likelihood estimation and produces empirical Bayes estimates for the random effects. Instead of P-Splines, polynomial splines were estimated. Since modelling the effect of a variable with splines is not implemented in MIXOR, choosing the correct model was very time-consuming. It was therefore necessary to reduce the parameters with the information obtained by estimating the ordinal model in BayesX first.

Note, that both approaches, the Bayesian in BayesX and the frequentistic in MIXOR, use the definition of the ordinal threshold model described in (3.28). Hence, higher values of the predictor or the effects in this predictor, correspond to a higher change in EDSS. That is, the response variable is shifted towards "increase". Whereas, negative parameter estimates stand for "decrease", i.e. an improving disease course.

4.3.1 The ordinal model without random effects

According to the model chosen for the metric situation, the ordinal threshold model without random estimates was estimated in BayesX and can be written as

$$\begin{aligned} changeord_{it} &= f_1(t_i) + f_2(age_i) + f_3(edss_i) + f_4(dur_i) + \beta_1 * course_i^{(1)} \\ &\quad + \beta_2 * course_i^{(2)} + \beta_3 * gender_i + \varepsilon_{it}, \end{aligned} \tag{4.2}$$

thresholds	changeord	
$\leq \theta_1$	big decrease	$(<<)$
$(\theta_1; \theta_2)$	small decrease	$(<)$
$(\theta_2; \theta_3)$	stable	$(=)$
$(\theta_3; \theta_4)$	small increase	$(>)$
$\geq \theta_4$	big increase	$(>>)$

Table 4.6: Boundaries of thresholds and their corresponding categories

As before, P-Splines of degree 3 with second order random walk penalty were chosen to model the metric variables. Instead of the scale parameter, a prior assumption for the ordinal variable had to be specified. According to other fixed effects, a diffuse and non-informative prior $p(\theta) \propto const.$ was chosen. All other prior distribution assumptions remained the same.

Like before, the convergence and mixing distribution is of interest. As the computation is more demanding, a much larger number of iterations is required to obtain a behavior comparable to the previous model. A burn-in period of 500000 and a step-width of 1000 was chosen. This choice guaranteed an almost ideal behavior for the samples and autocorrelation plots of all P-Spline parameters. Those plots haven't been displayed. The diagnosis plots of the constant effects, shown in Figure 4.13, are also satisfying. However, the trace plots of the threshold parameter samples (Figure 4.14) illustrate a bad mixing behavior. This oscillating behavior is also reflected in the autocorrelation plot (Figure 4.15). Positive and negative correlations seem to alternate. Hence, the estimation of threshold parameters is not very stable.

4.3.2 Results of the ordinal model

Threshold parameters

The estimated threshold parameters are listed in Table 4.7. It is assumed, that there is an underlying latent variable, with the thresholds defining the boundaries of this variable. Table 4.6 illustrates the connection between thresholds and the corresponding predicted outcome level of the ordered variable. That is, for a predicted value less than θ_1, a big decrease is assumed, between θ_1 and θ_2 a small decrease and so on. For example, if the effect of a predictor for one patient is 0.22, then it lies between θ_2 and θ_3 and the category "no change" is predicted as an outcome.

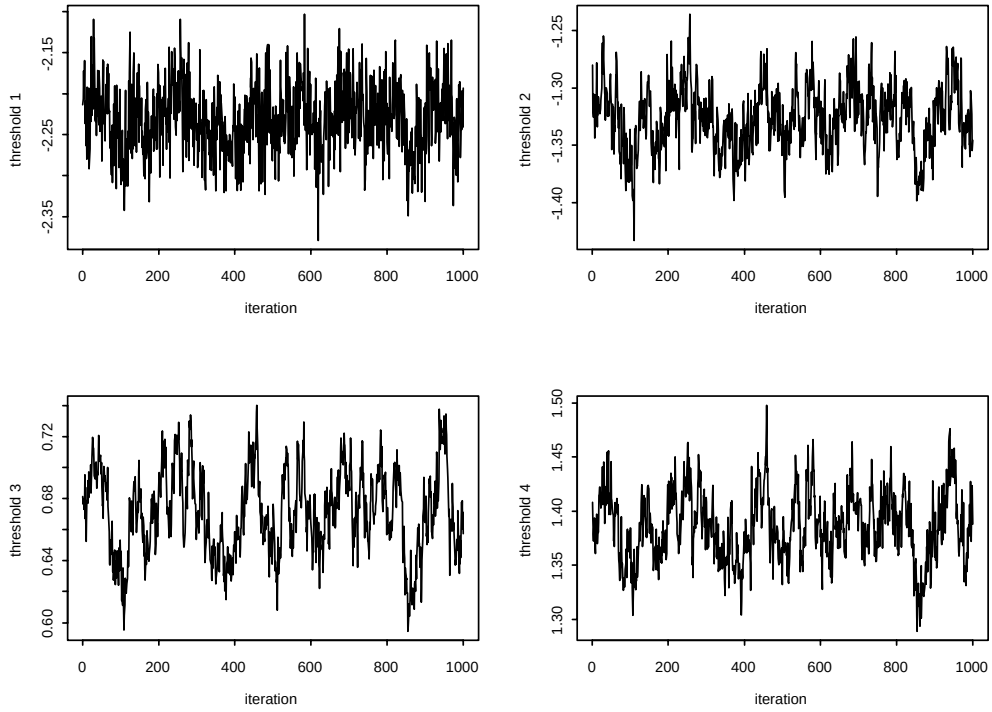

Figure 4.14: Sampling plots of threshold parameters

threshold	Mean	Std.Dev.	10% Qu.	50% Qu.	90% Qu.
θ_1	-2.23052	0.027015	-2.28281	-2.23142	-2.17701
θ_2	-1.32659	0.014089	-1.36323	-1.3259	-1.28996
θ_3	0.671491	0.023296	0.635749	0.672089	0.707502
θ_4	1.38643	0.034724	1.34592	1.38571	1.42779

Table 4.7: Estimates of threshold parameters

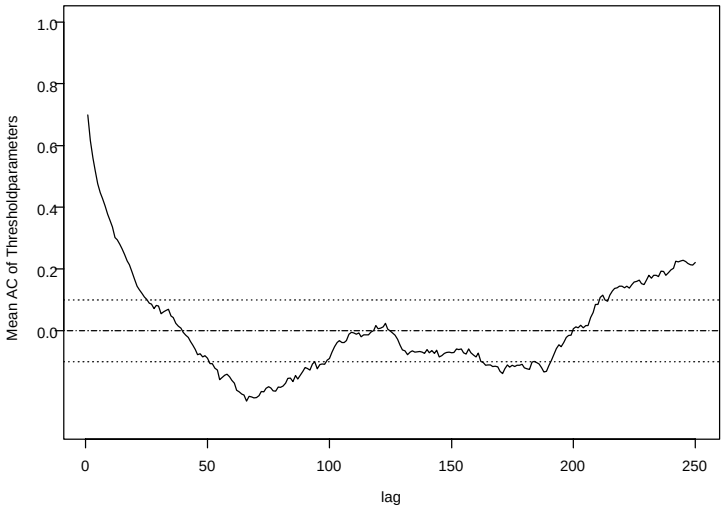

Figure 4.15: Mean autocorrelation of threshold parameters

Constant effects

Not surprisingly, the trends of the estimated constant effects, shown in Table 4.8 didn't change. The effects of the progressive courses (0.441641 and 0.45082) are significantly different to the reference category (relapsing-remitting). That is, a more severe course can be expected for patients categorized into sp, pp or pr. The influence of gender on the weighted change in EDSS slightly increased compared to the Gaussian model. Although the magnitude is very small, the posterior mean, 10% and 90% quantiles are all negative, meaning that women increase more in their level of disability than men. This result is in opposite to previous findings in MS clinical trials.

Variable	Mean	Std.Dev.	10% Qu.	50% Qu.	90% Qu.
gender	-0.093089	0.032357	-0.137522	-0.091968	-0.052025
course$^{(1)}$	0.441641	0.058101	0.367282	0.444438	0.515193
course$^{(2)}$	0.450820	0.056695	0.378324	0.451493	0.524575

Table 4.8: Estimates of constant effects for the ordinal model

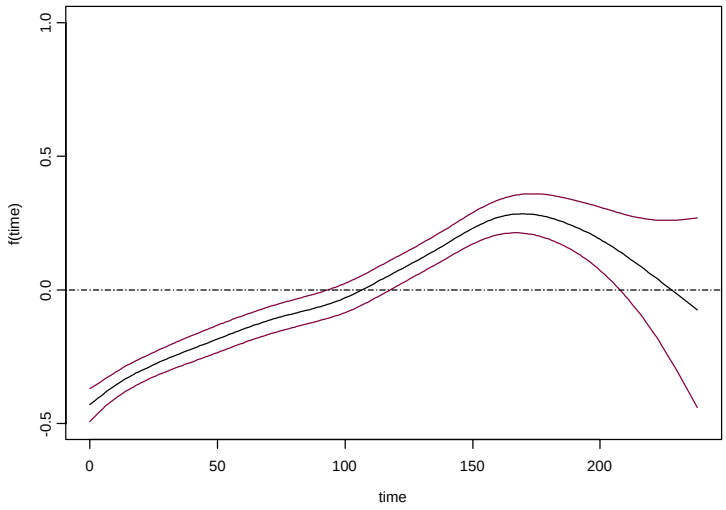

Figure 4.16: P-Spline for time since study entry

Effect of time

The P-Spline for the time effect (illustrated in Figure 4.16) obtained in the ordinal model is in accordance with the P-Spline in the metric model. It is just shifted to more negative values. But since the predictor denotes the weighted change itself in the one model, and the latent variable in the other, the magnitudes of the effects are not comparable. The shape of the curves are important. Like before, the trend increases linearly, but decreases beyond weeks 160-170.

Effect of age

Figure 4.17 shows the estimated P-Spline for the effect of age on the ordinal response variable. It is remarkable, that effects, that show only very slightly in the previous model, seem to be intensified in the ordinal model, like the peak at age 30 and the dip at age 45. As before, a confident estimation can only be done between the ages 18 and 45, as sufficient information is available in this range.

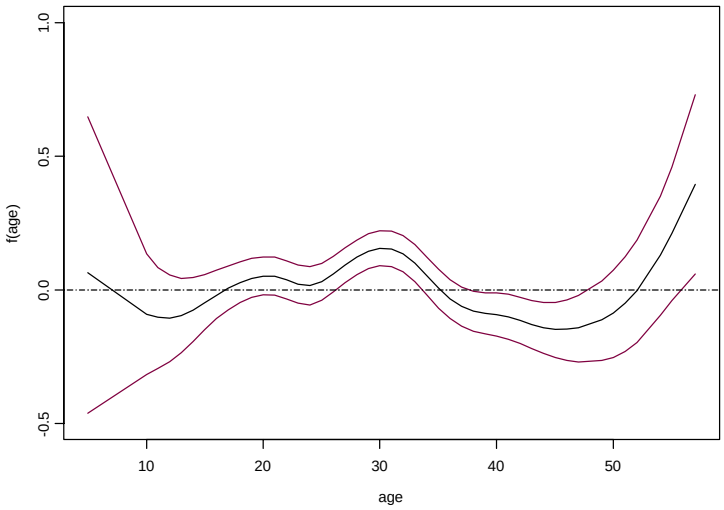

Figure 4.17: P-Spline for age at onset

Effect of baseline EDSS

In contrast to the very smooth function obtained in the Gaussian model, the P-Spline for entry EDSS, that has been estimated in the ordinal model, is very rough. The interpretation of the curve (Figure 4.18) is not very different, though. The level of disability increases for very low, as well as very high values of EDSS. Significant negative effects, indicating a decreasing disease course, can be found for EDSS values equal to 2.5, 4, 4.5 and 5.5.

Effect of duration

The effect of duration, illustrated in Figure 4.19 doesn't follow a clear trend. The effects are positive for a duration of 100-200 months and once again for approximately 350 months. The duration times in between, before 100 and above 400 months speak for a decreasing disease course. However, the magnitude of the effects is very small and the zero effect line almost always lies in between the 10 and 90% quantiles, indicating non-significant effects.

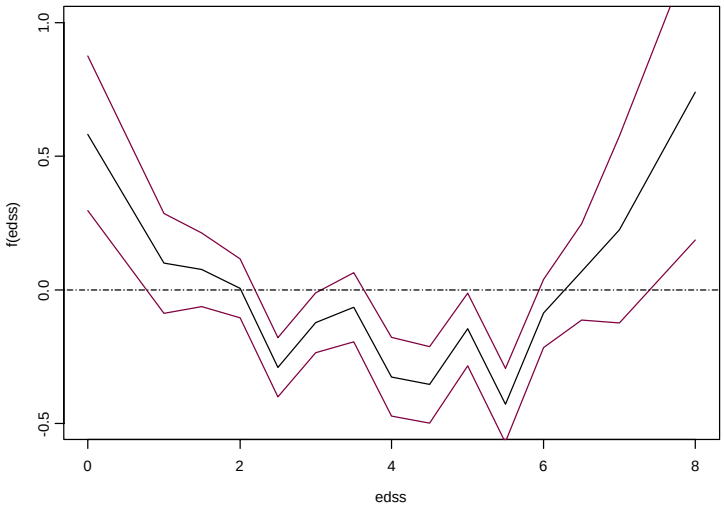

Figure 4.18: P-Spline for baseline EDSS

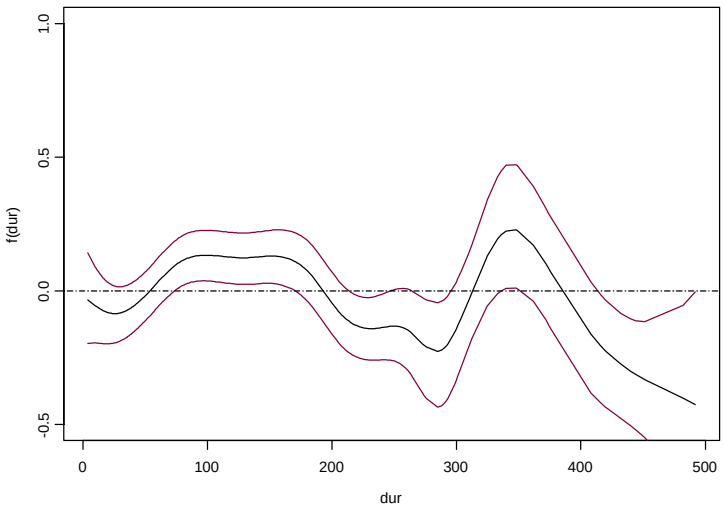

Figure 4.19: P-Spline for duration from onset

4.3.3 The ordinal model with random effects

As already stated, information obtained by estimating the fixed effects model in BayesX is used to construct an appropriate ordinal regression model for MIXOR. In principle, the previous model was taken unchanged, except the inclusion of a random-intercept effect, so that the following mixed-effects model will be estimated:

$$\begin{aligned} changeord_{it} &= f_1(t_i) + f_2(age_i) + f_3(edss_i) + f_4(dur_i) + \beta_1 * course_i^{(1)} \\ &\quad + \beta_2 * course_i^{(2)} + \beta_3 * gender_i + b_i + \varepsilon_{it}. \end{aligned} \tag{4.3}$$

Another difference to the previous model lies in the estimation method. Here, marginal maximum likelihood estimation, utilizing a Fisher-scoring algorithm, is used. Moreover, the smooth functions for the metric variables are not estimated with P-Splines, but with polynomial regression splines. Available information has been used to reduce the number of parameters to estimate:

- As the time effect is linear up to a cut-point of approximately week 160 to 170, a piecewise linear function was used to estimate the effect of time. The choice of the cut-point is crucial to the interpretation. Thus, a grid search between week 155 and 180 was performed and the cut-point maximizing the log-likelihood was chosen. With the obtained cut-point of 170, the function for the time effect can be written as

$$f_1(t_i) = \alpha_1 t_i + \alpha_2 (t_i - 170) I_{t_i > 170}, \tag{4.4}$$

  where

$$I_{t_i>170} = \begin{cases} 1, & \text{if } t_i > 170 \\ 0, & \text{if } t_i \leq 170 \end{cases}.$$

- In the previous model, the P-Spline for entry EDSS also shows a piecewise character: the effect is linearly decreasing from 0 to 2.5, fluctuating between 2.5 and 5.5 and positive linear beyond 5.5. Hence, a very sparse parametrization with only three knots was chosen, i.e. at EDSS equal to 2.5, 4, and 5.5. The regression spline of degree 3 is defined by

$$\begin{aligned} f_3(edss_i) &= \sum_{j=1}^{3} \alpha_j edss_i^j + \sum_{l=1}^{3} \alpha_j (edss_j - knot_l)_+^3, \qquad (4.5) \\ \text{where } knot_l &= \{2.5, 4, 5.5\}_{l=1}^3. \end{aligned}$$

- The P-Splines for age and duration didn't show such a clear pattern. The resulting choice of knots represents a trade-off between model fit, indicated by the log-likelihood, smoothness and parameter sparseness. Fortunately, a high number of equidistant knots was not required, so that the splines of 3rd order are given by:

$$\begin{aligned} f_2(age_i) &= \sum_{j=1}^{3} \alpha_j age_i^j + \sum_{l=1}^{8} \alpha_j (age_j - knot_l)_+^3, \qquad (4.6) \\ \text{where } knot_l &= \{15, 20, 25, 30, 35, 40, 45, 50\}_{l=1}^{8} \end{aligned}$$

and

$$\begin{aligned} f_4(dur_i) &= \sum_{j=1}^{3} \alpha_j dur_i^j + \sum_{l=1}^{8} \alpha_j (dur_j - knot_l)_+^3, \qquad (4.7) \\ \text{where } knot_l &= \{50, 100, 150, 200, 250, 300, 350, 400\}_{l=1}^{8}. \end{aligned}$$

Note, that the starting values of polynomial regression splines are set to 0. As before, a positive coefficient indicates a positive association between regressor and ordinal outcome. Confidence intervals are not provided in the following plots of the regression splines. A rugplot at the bottom line of each spline plot illustrates the distribution of the corresponding covariate, i.e. each tickmark stands for one observation.

4.3.4 Results of the ordinal mixed model

Threshold parameters

For identification, the first threshold was set to 0. Thus, the results from the previous model cannot be compared directly. The shape of the spline curves is important for interpretation. The estimated threshold parameters are given in Table 4.9 and the predicted outcome is interpreted as in Table 4.6. Since the intervals between the thresholds are bigger than in the previous model, it can be expected, that the magnitude of the estimates will be bigger.

threshold	Estimator	std.error	z-value	p-value
θ_1	0	–	–	–
θ_2	1.46894	5.66944	44.35629	<0.0001
θ_3	4.59204	0.03922	117.09799	<0.0001
θ_4	5.66944	0.04102	138.22608	<0.0001

Table 4.9: Estimates of threshold parameters

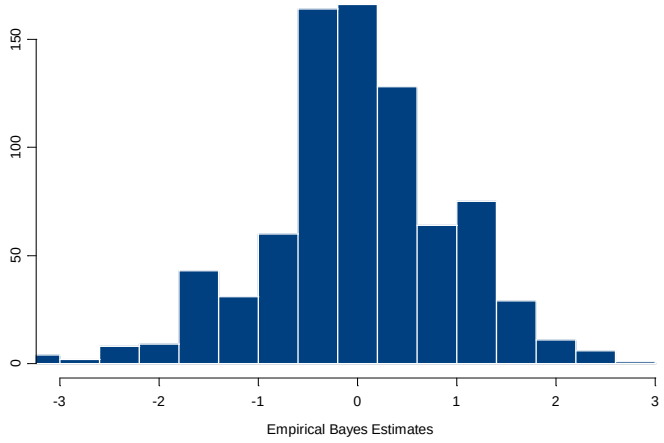

Figure 4.20: Histogram of Empirical Bayes Estimates

Variance components and random effect

In MIXOR, a residual variance of 1 is assumed. Relative to that, the estimated "between-patients" variance is 1.237. Hence, the variance component obtained by introducing the random effect even exceeds the residual variance, indicating, that the random effect is responsible for explaining a substantial part of variance in this model.

The histogram of empirical Bayes estimates (Figure 4.20) illustrates an almost symmetric distribution that departs from a normal distribution by a higher curtosis and several outliers.

Constant effects

Table 4.10 shows the estimates for all constant effects. The estimates are highly significant for both progressive courses, compared to relapsing-remitting.

	Estimate	Std.Error	z-value	p-value
gender	-0.08919	0.08839	-1.00902	0.31297
course$^{(1)}$	0.54872	0.17777	3.08677	0.00202
course$^{(2)}$	0.62386	0.12768	4.88632	0.00000

Table 4.10: Estimates of constant effects

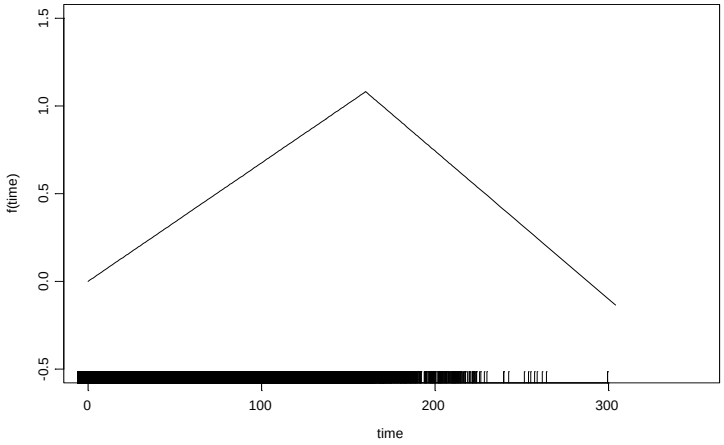

Figure 4.21: Regression spline for time

The estimated effect of gender is still negative, but not significant. Thus, it doesn't go against common knowledge.

Effect of time

The effect for time since first observation is illustrated in Figure 4.21. The cut-point clearly seperates the two linear pieces: The curve increases from 0 to 1 in the first part and decreases with almost the same slope from then on.

Effect of age

The effect of age, shown in Figure 4.22 , seems to follow a sinusoidal curve with several peaks. Compared to all other effects, the influence of this oscillating function is not big. Considering, that the interval between θ_2 and θ_3, for example, is almost 3 points, it is almost neglectable. The high negative

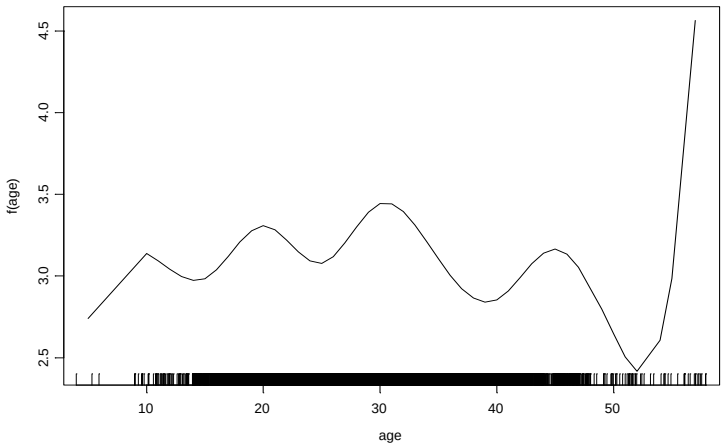

Figure 4.22: Regression spline for age at onset

effect at age 57, followed by a steep increase is remarkable. As already stated, sufficient information is only available for ages 18 to 45.

Effect of baseline EDSS

Due to the fact, that the starting value of each polynomial spline in this model is set to 0, all EDSS-values from 0 to 6 have negative effects and thus are associated with an improving clinical course. For EDSS values above 5.5, the spline increases again. Nevertheless, the interpretation of the curve in Figure 4.23 doesn't change compared to the previous Gaussian or Bayesian ordinal model: Patients starting with EDSS values between 2 and 6 can be expected to develop better than patients with lower or higher EDSS-values. That is, the disability of patients with these baseline EDSS values decreased during the clinical trial.

Effect of duration

Figure 4.24 shows the regression spline for the effect of duration on the ordinal outcome variable. As in the Bayesian ordinal model, the effect increases in the beginning and stays positive up to approximately month 200. Then it decreases, until it reaches its minimum at about 300 months. At 350 months, another peak can be found, that decreases again in the upper tail of

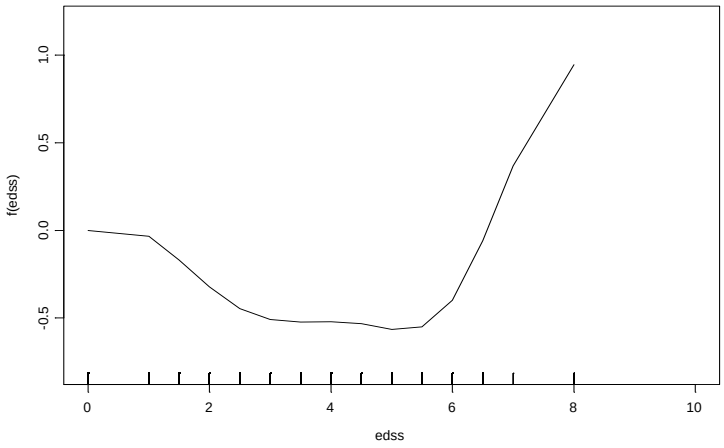

Figure 4.23: Regression spline for baseline EDSS

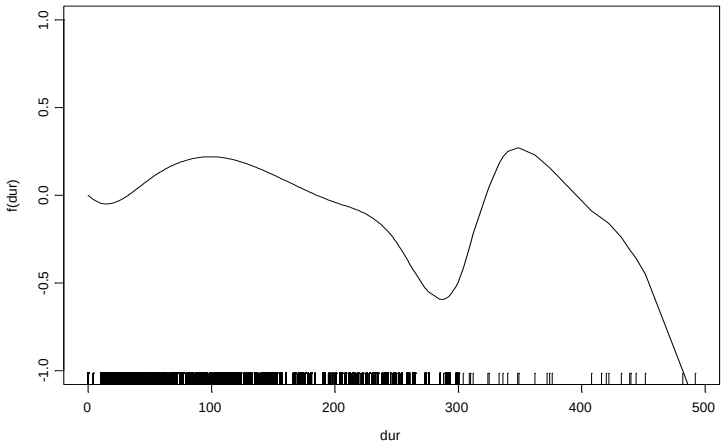

Figure 4.24: Regression spline for duration from onset

observed category	fitted category: <<	<	=	>	>>	Total
<<	**10**	61	47	0	0	118
<	16	**120**	556	4	0	696
=	85	128	**4684**	278	6	5181
>	2	15	704	**661**	17	1399
>>	0	6	154	453	**261**	874
Total	113	330	6145	1396	284	8268

Table 4.11: Crosstab of observed and fitted response

the curve. It has to be stated once again, that the vast majority of patients had a duration less than 200 months (see Figure 2.3). In this range, the magnitude of the curve is very small. For bigger values, the variance of the estimates is increasing, so that a reliable estimation is not possible.

Conclusion

Categorizing the weighted change in EDSS into 5 ordered categories didn't result in deviating interpretation of the estimated parameters. Results, that appear very different in the first sight, like the regression splines for duration and age at onset, don't show such a discrepancy, when looked at closer. Due to different outcome variables, the estimates cannot be compared directly. Yet the hypothesis that effects appear to be of bigger influence in the ordinal model, has been confirmed. Furthermore, the variance of the estimations has to be taken into account. Unfortunately, confidence bands couldn't be added in the plots of regression splines obtained in MIXOR. But the P-Splines of the ordinal model in BayesX as well as the rugplots can serve as indicators, how accurate the estimation is.

As in the Gaussian model, the model fit will be analyzed by comparing the fitted values against the observed values. The crosstab in Table 4.11 also indicates a systematic error. Only 69.4% of all observations are classified correctly. A good fit is only achieved in the category, that defines a "stable" disease progression. All other fitted values are shifted towards this same category. That is, observations on both extreme ends of disease progression cannot be explained very well by the ordinal model as well as the Gaussian

model. Moreover it has to be noted, that many computational problems occured during the estimation of the ordinal model. The autocorrelation and trace plots of the threshold samples (Figures 4.14 and 4.15) showed a bad mixing and convergence behavior, although random effects havn't been included in this stage of modelling. Analyzing the mixed-effects ordinal regression model in MIXOR also led to numerical difficulties. Adjustments had to be made to improve the chances of convergence. Thus, a Gaussian model should be preferred. Using the Gaussian model also seems to be justified, as the results of both approaches don't differ substantially. Nevertheless, the fit of the model to the data, as shown in Table 4.11 and Figure 4.12 is not satisfying. Since only one random variable, namely a random intercept, was included so far, adding other random variables should be considered.

4.4 Analyzing a random slopes model

4.4.1 Formulation of the random slopes model

The random intercept models proposed before are not appropriate for fitting the repeated measures for weighted change in EDSS. They underestimate the change for patients, whose disability greatly decreased or increased within the time frame of a clinical study. This could be due to the fact, that progression of disability not only differs in magnitude within the group of patients, but also in speed. The disability of one patient may rise fast in the beginning and then stabilize, wherease it rises steadily, but slow for another patient. Of course, the introduction of a random slope alone is not sufficient, as it assumes a constant slope over the whole range of the time frame. By adding a quadratic random effect, the curvature in the progression of disability can be reflected. The splines for the time effect in the previous model (Figures 4.7 and 4.16) also showed a quadratic trend. Hence, a quadratic random slopes model seems to be appropriate to account for the nonparametric effect of time detected before, as well as the heterogeneity in disability increase. To ensure interpretability, fixed effects for the intercept, linear slope and quadratic slope were also included. Random effects can then be interpreted as deviations from the population mean. Thus, the model can be formulated in the following way:

source of variation	Mean	Std.Dev.	10% Qu.	50% Qu.	90% Qu.
scale	0.360485	0.006067	0.352562	0.360414	0.368427
intercept	0.024117	0.009096	0.012335	0.024135	0.036359
linear slope	0.000449	$3.66*10^{-5}$	0.000404	0.000448	0.000495
quadr. slope	$1.77*10^{-5}$	$9.30*10^{-7}$	$1.65*10^{-5}$	$1.76*10^{-5}$	$1.89*10^{-5}$

Table 4.12: Estimates of variance components

$$\begin{aligned} changew_{it} &= f_1(age_i) + f_2(edss_i) + f_3(dur_i) + \beta_1 * course_i^{(1)} + \\ &\quad \beta_2 * course_i^{(2)} + \beta_3 * gender_i + (b_{i0} + \beta_0) + \\ &\quad (b_{i1} + \beta_4) * t_i + (b_{i2} + \beta_5) * t_i^2 + \varepsilon_{it}, \end{aligned} \tag{4.8}$$

where β_{i0} is the random intercept, β_{i1} the random slope and β_{i2} the quadratic random slope parameter. The fixed intercept is denoted by β_0, the fixed time by β_4 and the fixed quadratic time effect by β_5. As in the Gaussian model, prior distributions of all fixed effects are diffuse, whereas all random effects are assumed to be normally distributed.

To ensure comparability of the random intercept and random slopes model, calculation in BayesX was performed with the same number of iterations, i.e. a burn-in of 20000 and a step-width of 500. The convergence and mixing behavior was comparable to the ones obtained in the random intercepts model (Figure 4.5). Hence, non of the Gaussian models is numerically superior.

4.4.2 Results of the random slopes model

Variance components and random effects

The scale parameter and the variance components of the random effects, shown in Table 4.12, are reduced by a significant amount in comparison to the random intercept model. Above all, it is noticeable, that the variance components belonging to the random effects are much smaller than the scale parameter. This is due to the higher number of parameters in the model, especially the random slope and quadratic random slope effects.

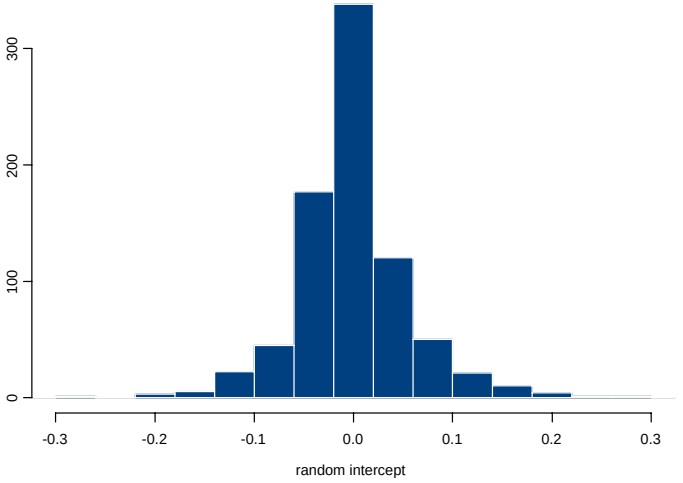

Figure 4.25: Histogram of random intercept estimates

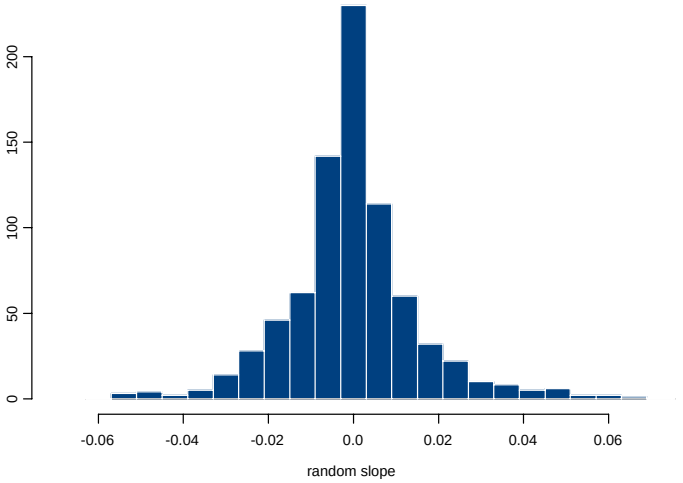

Figure 4.26: Histogram of random slope estimates

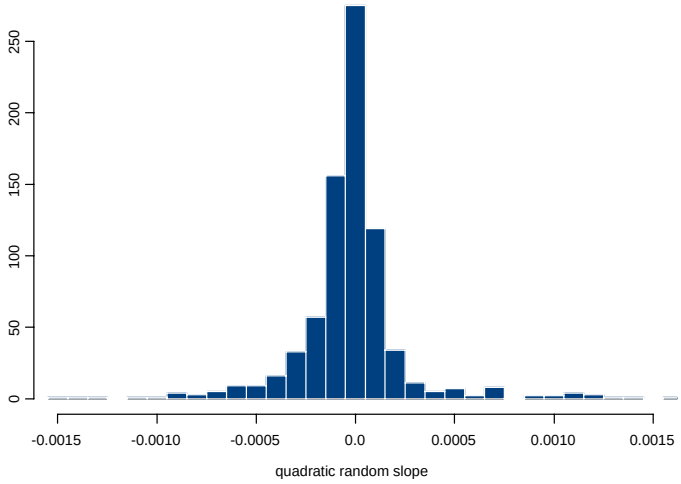

Figure 4.27: Histogram of quadratic random slope estimates

Histograms of the random effects (Figures 4.25, 4.26 and 4.27) illustrate the corresponding posterior distributions. The curtosis is higher than a normal distribution and a lot of outliers can be detected. Although the distributions are almost symmetric and bellshaped, a deviation from a normal distribution is likely.

Constant effects

Table 4.13 shows the fixed effects of the intercept, time and quadratic time trend, both progressive courses and gender. The positive posterior mean of

Variable	Mean	Std.Var.	10% Qu.	50% Qu.	90% Qu.
intercept	0.017732	0.006067	0.352562	0.360414	0.368427
time	0.004538	0.001049	0.003242	0.004499	0.005897
$time^2$	$1.19*10^{-5}$	0.000156	0.000191	0.000209	0.000218
$course^{(1)}$	0.129323	0.049309	0.070234	0.128142	0.192123
$course^{(2)}$	0.089652	0.042889	0.032993	0.089353	0.144362
gender	-0.070236	0.030946	-0.11044	-0.06945	-0.110436

Table 4.13: Estimates of constant effects

course		mean of random linear slope	mean of random quadratic slope
pp or pr	(course$^{(1)}$)	-0.000024	0.000133
sp	(course$^{(2)}$)	0.000927	0.000060
rr	(reference category)	-0.000561	-0.000154

Table 4.14: Mean of random slope estimates, stratisfied for courses

both fixed time and quadratic time parameters indicate, that the disability averaged over the whole population is increasing over time. The increase even seems to speed up. This stands in contrast to the P-Spline curves of time in the previous models. They indicate, that disability increases faster in the beginning and then flattens during the course of a clinical study. This behavior might be captured by the random effects for the quadratic time effect. The histogram in Figure 4.27 illustrates, that there are a lot of negative outliers. The mean and median are also slightly negative. Furthermore, patients, that are observed over a longer time generally deviate from the population effect in the negative direction. Thus, the flattend and almost negative trend at the upper tail of the time distribution, that was detected before, is reflected in the random estimates.

The positive estimates of the progressive courses, although they are still existent, are noticeably smaller than before. It seems, that some amount of information, that was given by the course of disease in the previous model, is captured by the individual linear or quadratic slopes. To confirm this assumption, a closer look has been taken into the distribution of the random slope estimates within each group. Table 4.14 shows the mean values for the linear and quadratic random slope parameters. Slopes for progressive patients (sp, pp or pr) are generally higher than for relapsing-remitting patients (rr). Hence, the categorization into disease courses reflects the kind of progression over time, so that a time-dependent effect rather than a constant effect per group could be an alternative.

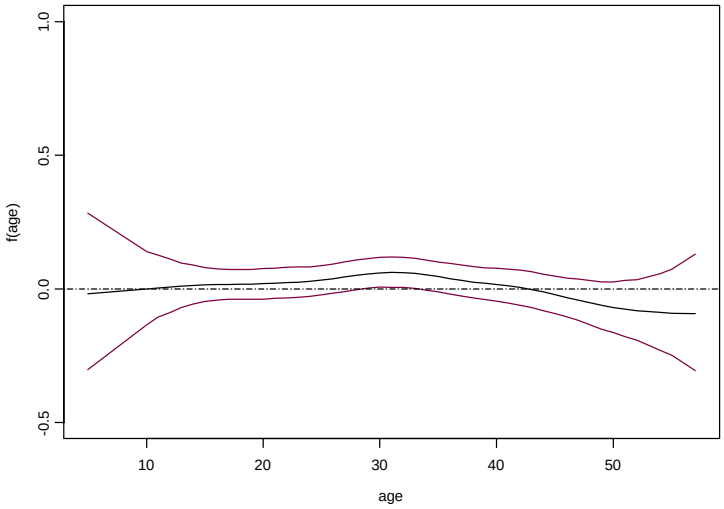

Figure 4.28: P-Spline for age at onset

Effects of metric covariates

The random slopes model doesn't change the information obtained by the P-Spline curves substantially. The splines for age at onset (Figure 4.28) and duration (Figure 4.29) are horizontal and the corresponding 80% credible intervals include the zero-effect line. The negative slope of duration, that was detected before, disappears. But since the credible interval was wider in the random-intercept model and also overlapped zero, the interpretation didn't change significantly.

The influence of the EDSS value at first observation (see Figure 4.30) remains the same. The increase in disability for patients with an entry EDSS of 2 to 6 is lower than for patients with a lower or higher entry EDSS. Note, that the introduction of random slope effects reduced the variance of the P-Spline estimates. The credible intervals are noticeably narrower than before.

Conclusion

The histograms and normal-quantile plots of residuals (Figure 4.31) show almost symmetric distributions, but with more outliers than a normal distribution and a higher curtosis. Compared to the random slopes model, the

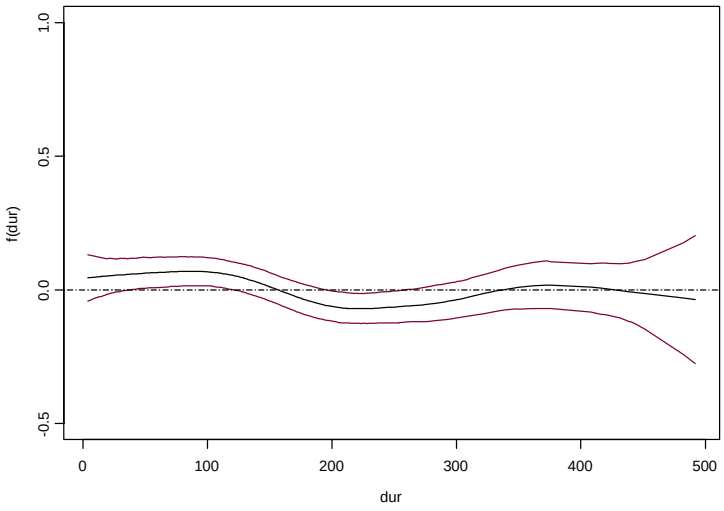

Figure 4.29: P-Spline for duration from onset

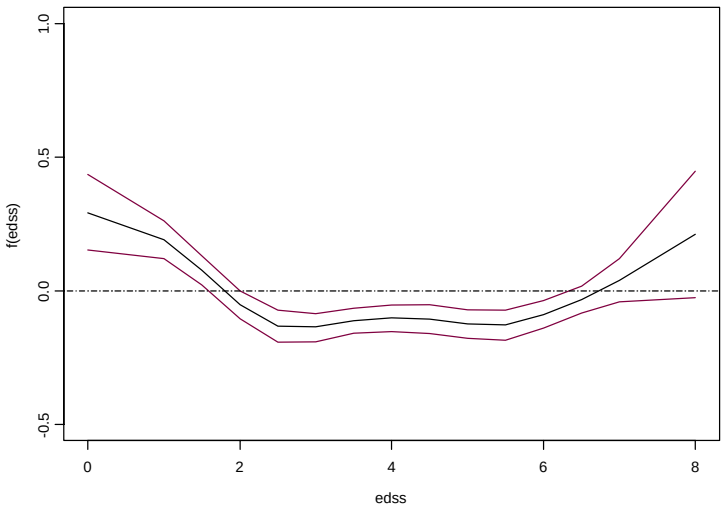

Figure 4.30: P-Spline for baseline EDSS

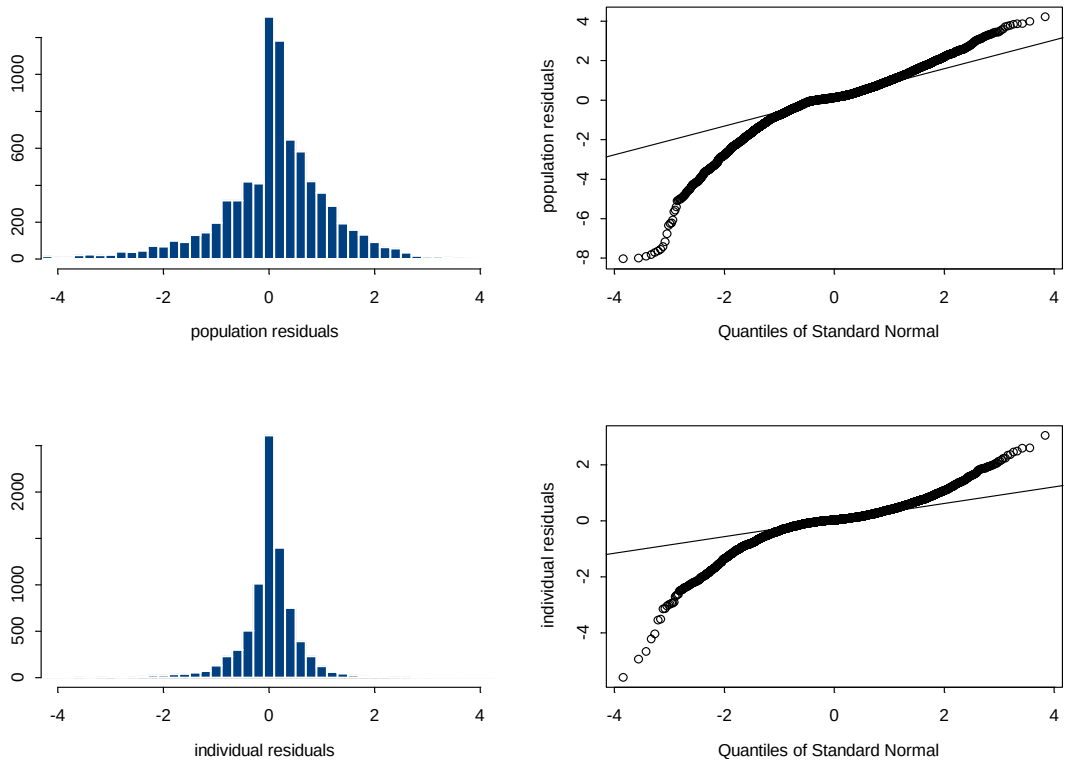

Figure 4.31: Histograms and normal-quantile plots for population (top) and individual residuals (bottom)

magnitude of residuals are smaller. As before, the introduction of random variables causes a shift of the residuals towards zero. Moreover, random slopes helped to reduce the bias, that was detected in the previous models. The plot of fitted against observed values (Figure 4.32) still shows a systematic trend. But as the dashed line, indicating a linear regression fit, lies closer to the diagonal, model fit is improved significantly. However, outliers can still be detected, especially on both extreme ends of the weighted EDSS change.

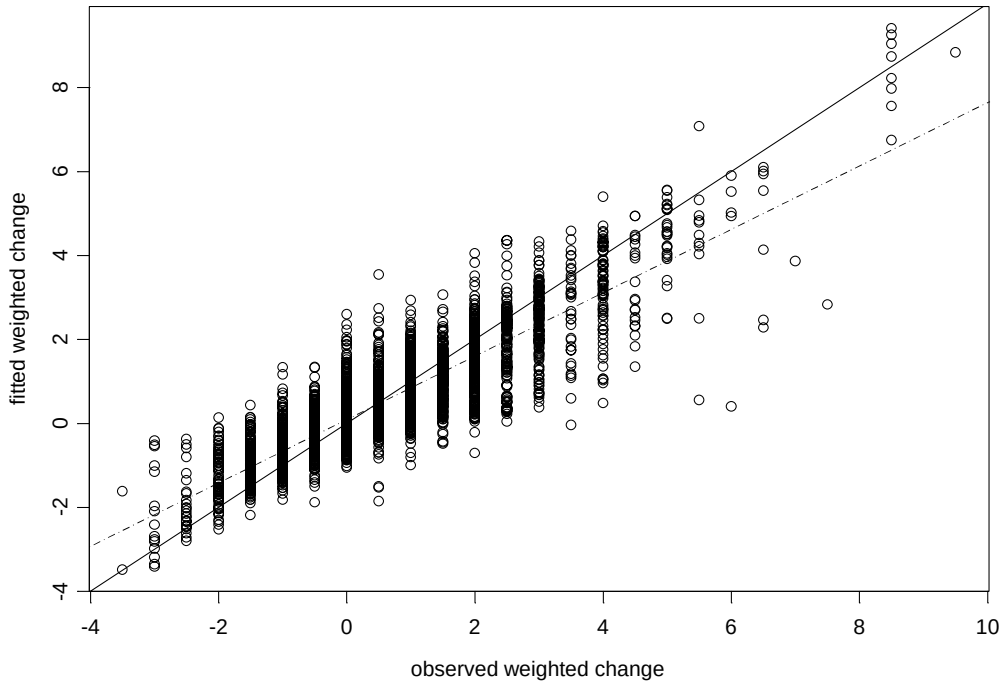

Figure 4.32: Plot of observed against fitted values

Chapter 5

Conclusion

Different statistical methods have been presented in this work. Overall it can be noted that mixed models are a powerful tool for analyzing longitudinal data. In the case of MS, random effects are necessary to account for unobserved heterogeneity and to obtain unbiased estimates of the remaining effects. Introducing random effects reduced the error variance significantly. Thus, they are explaining a substantial part of the variance in the data. It was shown, that a random intercept is not sufficient. Random linear and quadratic slopes are required to assess the disease progression for each patient. Especially for patients, who experienced a high increase or decrease in disability, disease progression cannot be fitted by a mixed model with random intercept alone. Both random slope terms help to reduce the systematic trend, although a too conservative behavior of the model can still be detected.

Another question was whether it is justified to use the change in EDSS as a metric outcome variable. The assumption, that statistics can only be as good as the data it is based on, also holds in this case. To account for the well-known unlinearity of the EDSS, the changes have been weighted according to a widespread acceptance. But the criticism, that a 1 point change, although weighted, doesn't mean the same over the range of the EDSS, remains. Analyzing an ordinal model should clarify this question. Combining levels of the outcome variable to 5 ordered categories not only accounts for the ordinal structure in the response. But it is presumed that this oversimplifying ensures comparability of the responses. Overall it was shown, that the computationally demanding and time-consuming estimation of an ordinal mixed effects model is not necessary. The interpretation of

results didn't change dramatically. Furthermore, numerical problems occured in both fixed and mixed ordinal models, so that the reliability of results is not given. Reducing the number of ordered categories to three could be an alternative. In this case, the levels of the response are reparametrized to obtain stable estimates and then, analysis can be carried out in BayesX. In this thesis, however, Gaussian models with the weighted change as a metric variable was followed up.

Bayesian methods, based on MCMC algorithms, emerged as extremely flexible. Models, that are too complex for classical maximum likelihood estimation, can be estimated in a Bayesian context. The big advantage of the program BayesX is the incorporation of smooth P-Splines for metric covariates. Nonlinear effects, like the influence of the EDSS at first observation couldn't have been detected using simple linear models or with other strict assumptions on the functional form.

Overall, the entry EDSS appears to be of big influence on the weighted change in EDSS. Patients enrolled in a study with an EDSS lower than 2 or bigger than 5.5 can be expected to experience a higher increase in disability as patients in between. The effects of duration of disease, age at onset and gender are marginal and even neglectable. However, there is a "positive" time trend during a clinical study. That is, the more time elapses, the higher is the change in EDSS. The course of the disease also emerged as a predictive factor. This variable can be seen as a summary of the past disease progression. As soon as there are more variables available, that hopefully explain the previous disease course of a patient, a shrinkage of this effect can be expected. In general, the disability of patients, that are categorized in one of the progressive courses, increases more. It is remarkable, that the influence of the disease course is much smaller in the random slopes model. It was shown, that this is due to deviating random effects between the 3 groups of patients. Thus, it is advisable to think about time-dependent effects, i.e. to estimate one slope for each group of patients.

Future analyses will be concentrating more and more on prediction of EDSS progression within a time frame of a trial to approach the goal of a "virtual" patient. The random slopes model proposed in section 4.4 will be followed up on new releases of the SLCMSR database. For prediction, it is necessary, to check model assumptions more thoroughly. It has already been detected, that normal distribution assumptions are violated to some extent. Furthermore, validation procedures have to be developed. Another approach, that is worth following is the already mentioned ordinal model with 3 ordered

categories. MCMC methods will be used for modelling mixed effect models. In any case, it proved useful to leave common paths of statistical analysis in MS research.

Finding questions to data and answers to questions requires communication between physicians and statisticians. In this thesis, medical advice has been followed to find an appropriate modelling method for disability progression of MS patients. It is certain that hints and advices from a medical viewpoint can still improve the statistical analyses that will evolve out of this work in the future.

Bibliography

[Brezger(2000)] A. Brezger. Bayesianische p-splines. Master's thesis, Department of Statistics, University of Munich, 2000.

[Brown and Prescott(1999)] H. Brown and R. Prescott. *Applied Mixed Models in Medicine*. John Wiley and Sons, 1999.

[Compston et al.(1998)] A. Compston et al. *McAlpines's Multiple Sclerosis*. Churchill Livingstone, 1998.

[Cutter(1999)] G. Cutter. Development of a multiple sclerosis functional composite as a clinical trial outcome measure. *Brain*, 122:871–882, 1999.

[Eilers and Marx(1996)] P. Eilers and B. Marx. Flexible smoothing with b-splines and penalties. *Statistical Science*, 11(2):89–121, 1996.

[EMEA(2001)] EMEA. Note for guidance on clinical investigation of medicinal products for the treatment of multiple sclerosis, 2001. URL `http://www.emea.eu.int/pdfs/human/ewp/056198en.pdf`.

[Fahrmeir and Lang(2001)] L. Fahrmeir and S. Lang. Bayesian semiparametric regression analysis of multicategorical time-space data. *Annals of the Institute of Statistical Mathematics*, 53:11–30, 2001.

[Fahrmeir and Tutz(2001)] L. Fahrmeir and G. Tutz. *Multivariate statistical modelling based on generalized linear models*. Springer Verlag, New York, 2001.

[Gilks et al.(1996)] W. Gilks et al. *Markov Chain Monte Carlo in Practice*. Chapman & Hall, 1996.

[Hammerlin and Hoffman(1994)] G. Hammerlin and K. Hoffman. *Numerische Mathematik*. Springer Lehrbuch, 1994.

[Hastie and Tibshirani(1990)] T. Hastie and R. Tibshirani. *Generalized Additive models*. Chapman & Hall, 1990.

[Hedeker and Gibbons(1994)] D. Hedeker and R. Gibbons. A random-effects ordinal regression model for multilevel analysis. *Biometrics*, 50:933–944, 1994.

[Kesselring(1996)] J. Kesselring. *Multiple Sclerosis*. Cambridge University Press, Cambridge, 1996.

[Kurtzke(1983)] J. Kurtzke. Rating neurologic impairment in multiple sclerosis: An expanded disability status scale (edss). *Neurology*, 33:1444–1452, 1983.

[Lang and Brezger(2001)] S. Lang and A. Brezger. Bayesian p-splines. Technical report, Department of Statistics, University of Munich, 2001.

[Lublin and Reingold(2001)] F. Lublin and S. Reingold. Placebo-controlled clinical trials in multiple sclerosis: Ethical considerations. *Annals of Neuology*, 49:677–681, 2001.

[McCullough and Nelder(1989)] P. McCullough and J. Nelder. *Generalized Linear Models*. Chapman & Hall, 1989.

[McDonald(2000)] I. McDonald. International database centre aimed at speeding up ms research to be established in munich. Final Press Release MSIF, 2000.

[Milliken and Johnson(1992)] G. Milliken and D. Johnson. *Analysis of Messy Data*, volume 1. Chapman & Hall, 1992.

[Noseworthy et al.(1990)] J. Noseworthy et al. Interrater variability with the expanded disability status scale (EDSS) and functional systems (FS) in a multiple sclerosis clinical trial. *Neurology*, 40:971–974, 1990.

[Polman et al.(2001)] C. Polman et al. *Multiple Sclerosis: the Guide to Treatment and Management*. Demos Medical Publishing, New York, 2001.

[Tutz(2000)] G. Tutz. *Die Analyse kategorialer Daten*. Oldenbourg Wissenschaftsverlag, Munich, 2000.

[Verbeke and Molenberghs(2000)] G. Verbeke and Molenberghs. *Linear Mixed Models for Longitudinal Data*. Series in Statistics. Springer, New York, 2000.

[World Medical Association(2000)] World Medical Association. Declaration of helsinki, 2000. URL `http://wma.net`.

Software

The following software has been used in this thesis:

- **Scientific Word 4.10**, MacKichan Software, Inc, for typesetting

  More information at http://www.mackichan.com

- **S-Plus 6, Insightful Corp.,** for descriptive statistics and graphics

  More information at http://www.insightful.com

- **BayesX** - Software for Bayesian Inference based on Markov Chain Monte Carlo simulation techniques

  Software and User Manual available at
  http://www.stat.uni-muenchen.de/∼lang/bayesx/bayesx.html

- **MIXOR**: a computer program for mixed-effects ordinal regression analysis

  Software and User Manual available at
  http://tigger.uic.edu/∼hedeker/mix.html

Hiermit versichere ich, dass ich diese Diplomarbeit selbständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe.

München, den 19. September 2002

(Claudia Lamina)