Statistical analysis of TBE antibody decrease after vaccination

Essay in Statistics

10 Swedish credits

Level: 61-80 Swedish credits

Karin Kepplinger

karin_kepplinger@yahoo.de

May 2007

Handed in to:

Björn Holmquist, Department of Statistics

Lund University, Sweden

Supervisor:

Helga Wagner, Institute for Applied Statistics

Johannes Kepler University Linz, Austria

Contents

Chapter 1

Objective and outline

This study is concerned with the statistical analysis of tick-borne encephalitis (TBE) antibody decrease after vaccination. The objective is to derive conclusions about protection duration of vaccination. TBE is a serious infection affecting the central nervous system and Austria is a high risk region of TBE. Active immunization by vaccination prevents TBE and nowadays 88% of Austria's population is vaccinated.
The statistical analysis is based on observational data collected from persons who deliberately went for a test of their TBE antibody level at the *Institut für Hygiene, Mikrobiologie und Tropenmedizin des Allgemeinen öffentlichen Krankenhauses der Elisabethinen* in Linz . In addition to antibody concentration in blood serum (titer) obtained from an ELISA-test, the data set contains information on sex, number of booster vaccinations, age and time since last vaccination (covariates).
Various parametric regression type models are fit to investigate the decrease of titer after vaccination. Pure regression models turned out not to be adequate for explaining titer decrease due to considerably heterogeneity between individuals. Subject-specific random effects have to be taken into account.

The thesis is divided into four main parts:

1. introduction and overview of TBE epidemiology, pathology and vaccination in Austria
2. analysis of the first measurements of each person starting with simple explorative data analysis, followed by classical regression models and generalized linear models
3. analysis of repeated measurements starting with simple explorative data analysis, followed by models for longitudinal data in Gaussian and non-Gaussian case
4. comparison of all models and prediction of titer

In each chapter a short theoretical overview is provided before applying the presented models to the observed data using the software SAS ®9.1. For detailed explanations of procedures and syntax the SAS online documentation (SAS Institute Inc., 2003) is recommended.

Chapter 2

Tick-borne encephalitis

In this chapter background knowledge on the disease TBE itself and information on TBE vaccination in Austria are gathered and some interesting articles and studies considering TBE are pointed out.

2.1 Epidemiology

TBE is a serious viral infectious disease affecting the central nervous system (CNS) caused by the TBE virus (TBEV). TBEV is a Flavivirus and is distinguished by the following subtypes due to its endemic region (Rendi-Wagner (2004a, p.307); Zoehrer *et al.* (2003, p.1165)):
- European subtype: Central Europe, Eastern Europe, Southern Sweden, Southern Finland
- Far-Eastern subtype: Asian part of Russia, Northern China, Northern Japan
- Siberian subtype

The term *tick-borne encephalitis* already indicates that the TBEV is mainly transmitted by ticks. Rarely infection occurs after consumption of unpasteurized dairy products (Rendi-Wagner, 2004a, p.307). Humans become most likely infected with TBEV by the sting[1] of infected ticks, which might be wiped off from grass or bushes during outdoor activities. Ticks live in vegetation at ground level up to one meter. Children are often stung by ticks into the head, arms and the upper part of the body, whereas grown-ups are rather stung into the legs and the lower part of the body (Stanek & Hofmann, 1994).
In Central Europe ticks are active between April and November. Main activity takes place in May/June and September/October. Three to four weeks after those periods more occurrences of TBE cases can be observed. The first description of an unknown seasonal peak of serious meningitis cases in an eastern district of Austria was done by Schneider (1931). The incidence of ticks is highest in humid areas with annual average temperature above seven degrees Celsius (Kunz, 1992). Mild winters and humid summers support the reproduction of ticks and the activity of ticks increases. Ticks prefer meadows and woodlands where a lot of little mammals live which might serve as alternate hosts. Detailed explanations of development cycle of ticks and TBEV go beyond the scope of this introductive part and can be found in Dumpis *et al.* (1999), Charrel (2004) and Heinz (2006).
The prevalence of TBEV in ticks differs from year to year and varies between regions. An investigation on TBEV prevalence in ticks was for example done in Styria (Upper-Austria) 1992

[1] The term of sting instead of bite shall be used in this thesis on the account of Löser *et al.* (2002).

and the rate of infected ticks was 0.44% (Labuda, 1993). Areas in Austria with high risk to be infected by TBE are situated along the Danube, in Styria, Carinthia and in Burgenland, see figure 2.1.

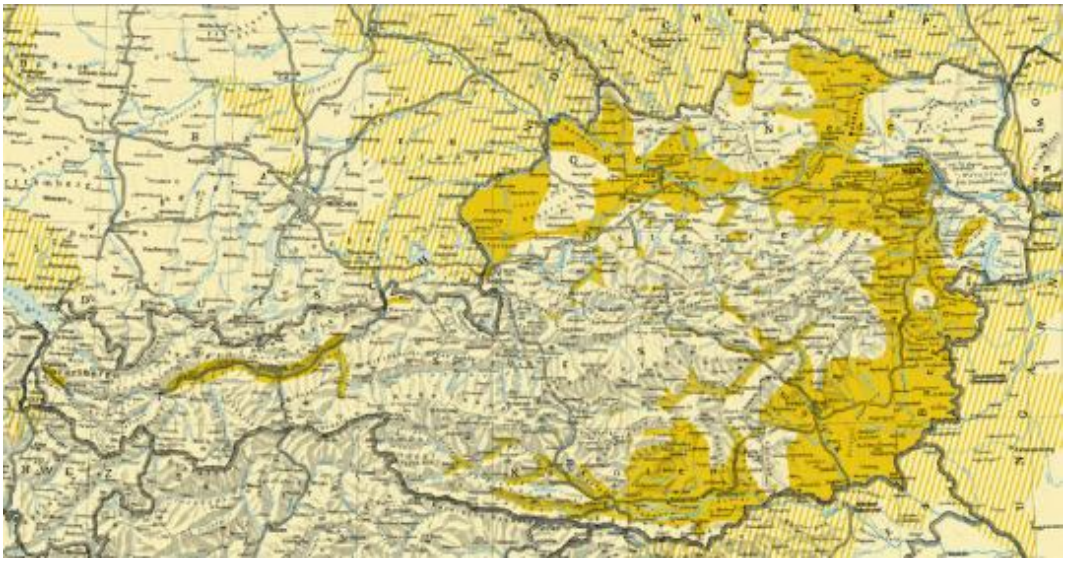

Figure 2.1: TBE risk areas in Austria, ARGE Gesundheitsvorsorge (Download: 2006-11-13)

2.2 Pathology

The incubation time for TBE normally lies between 4 and 14 days. The course of TBE can often be distinguished in two phases. In the first phase, which lasts three to five days, patients suffer from illness similar to influenza (Binder, 1996, p.52). After a symptom free interval of about one week, approximately two third of infected persons pass into the second phase of the disease with encephalitic symptoms involving the CNS (Kaiser, 2002). Depending on the affection of the CNS, time to recovery is extended to several months. The course of TBE disease is strongly related to the age of the victim. Especially persons above 40 years suffer severe discomfort. The risk of postencephalitic syndromes including spinal nerve paralysis, neuropsychiatric complaints, dysphasia, ataxia and paresis as well as the risk of permanent sequelae rises with increasing age (Charrel, 2004, p.1043). Mortality rate for TBE in Europe is less than 1%. More information on symptoms, different manifestations and prognosis of TBE is available in Kaiser (2006). Some examples for clinical studies investigating the severity of TBE diseases are Jezyna (1984) in Poland, Köck (1992) in Austria, Haglund (1996) in Sweden, Kaiser (1999) in Germany and Mickiene (2002) in Lithuania.

2.3 Vaccination

So far no efficient drug against TBEV has been found and treatment of TBE is only supportive (Dumpis *et al.* , 1999, p.887). Besides avoiding to get stung by an infected tick, accurately timed active immunization against TBEV appears the most effective way to prevent TBE. The first vaccine was developed in the early 1970s in Austria. The full vaccine dosage (0.5 ml) contains 2.0

to 2.75 μg TBEV-antigen. For children and adolescents up to 16 years only half of the dosage (0.25 ml) is used. (Gesellschaft für Virologie e.V., Download: 2007-03-18; P.N. *et al.* , 2003; Barrett, 2004)
The basic immunization protocol consists of three vaccinations. The first two vaccinations should be 2 - 12 weeks apart. The third vaccination follows 9 - 12 months later. The protection rate after three basic vaccinations is 96 - 98.7% (Kunz, 2003). Formerly it was recommended to do booster vaccinations every three years after the basic immunization. Several studies like Rendi-Wagner (2004b) and Kind (2004) evinced a longer protective immunity. The current proposal in Austria is to do the first booster vaccination after three years and the following vaccinations every five years. With increasing age the effect of vaccination and its protection time decrease (Hainz, 2002). The Austrian National Board for Immunization recommends people older than 60 years to do booster immunization every three years. In urgent cases an accelerated basic immunization schedule can be applied (0, 7, 21 days). This is especially interesting for travellers (ARGE Gesundheitsvorsorge, Download: 2007-03-31b). Problems with unprotected people travelling to TBE risk areas are discussed in Rendi-Wagner (2004a).

2.3.1 Mass vaccination campaign in Austria

In the 1970s Austria, former Czechoslovakia, Hungary, Slovenia and Croatia had the highest number of TBE incidents. In Austria a wide-spread TBE vaccination campaign was introduced in 1981. Nowadays around 88% of Austrians are vaccinated against TBE. Nevertheless there were 100 incidents of TBE in 2005 and 84 in 2006 (ARGE Gesundheitsvorsorge, Download: 2007-03-31a). Among the countries where TBE is endemic, Austria is the only one where TBE incidence has decreased since 1974. In all other countries the number of TBE cases has increased (Süss, 2006, p.19). TBE vaccination in Austria was especially promoted in schools and the proportion of TBE cases among older people increased. For this reason the age pattern of TBE cases in Austria is different than in other European countries. (Kunz, 2003, p.52)

2.4 Laboratory diagnosis

As soon the human immune system detects TBE agents in the blood circulation it reacts by producing specific defensive antibodies, so-called *immunoglobulin (Ig)*. Testing for immunity after a TBEV infection and immune response after TBE vaccination is based on ELISA (enzyme linked immuno sorbent assay) systems. These are methods to detect substances like hormones or proteins, e.g. Ig. The existence of those substances is proved if they react with so-called detection-antibodies, which are marked with an enzyme. This enzyme enables the conversion of a substrate and an antigen-antibody reaction can be proved.
TBE specific antibodies can be detected for several months after infection or basic immunization. In cases of former incidences with other Flaviviruses like vaccination against yellow fever or Japan encephalitis or a previous Dengue virus infection it is necessary to use the more costly but more specific neutralization test (Holzmann, 2003, p.39), because cross-reactions with other Flaviviruses are possible by using ELISA (Zoehrer *et al.* , 2003, p.1167). After a natural infection the victim is immune against TBEV and also against other subtypes for the whole life (Mickiene, 2006). Interested medical readers are recommended to study Holzmann (2003).

Chapter 3

Data

In the following chapter observed data on TBE vaccination are explored. First, the process of data collection and features of the data are described. Second, a population of interest is defined upon which statistical models for titer are applied in the subsequent chapters. Finally, simple frequency analysis on the constricted data is executed.
Similar studies on the protection of TBE vaccination with similar data situation have been done by Kind (2004), Kind (2005) in Switzerland and a clinical trial was done by Rendi-Wagner (2004b) in Austria.

3.1 Data collection and definition of the population of interest

From 1995 to 2004 data for 4031 measurements were recorded of persons who went to the microbiologic ambulance of the hospital Allgemein öffentliches Krankenhaus der Elisabethinen Linz for testing their TBE antibody level. The underlying questionnaire was designed by deputy Dr. Lothar Binder. The attending physician filled in the questionnaire in cooperation with the tested persons who are referred to as *proband or participants* in the following. Important items of the questionnaire were:
- How old is the proband at the moment?
- Which sex?
- Is the basic immunization protocol consisting of three vaccinations completed?
- When did those first three basic vaccinations take place?
- How many TBE booster vaccinations has the proband had so far?
- When did the last TBE-vaccination take place?
- Has the proband had any former incidents concerning yellow fever, Japan encephalitis or Dengue fever (vaccination or disease)?

If the proband had brought her/his personal vaccination card the physician could use this "exact" data. Otherwise the answers were referred to the more or less vague statements of the participants from their memory. A few persons went to the ambulance for testing their titer two or more times and several data records of them are available.
Unfortunately two different tests, Enzygnost^{TM} Anti-TBE Virus ELISA and VIE-ELISA, where used to determine the TBE antibody concentration. Both methods belong to ELISA systems, which have already been explained in the previous chapter, section 2.4. For simplification Enzygnost ELISA and VIE-ELISA are annotated as method A and method B, respectively.

The analysis of the measured titers is not so simple, because in 383 cases titers measured by method A were only noted down as "$< 7\ U/ml$". This values were coded with 6.99 to be able to use them as numerical values. Furthermore, titers measured by method B were often recorded as "<" or ">" some value, that makes an accurate analysis of titer difficult. However, these measures have been coded and are included in the analyzes. So far no functional relation between the results of these two ELISA methods have been found and there is no possibility to analyze the measured results of method A and method B together. Another option is to classify the results of both methods. According to the physicians at the Institut for Hygiene, Mikrobiologie und Tropenmedizin titers above 25 U/ml measured by method A and titers above 126 $VIEU$ measured by method B can be referred to as *seropositive* and a booster vaccination is not necessary at the moment. In table 3.1 the categories for the measured titers in both methods are summarized.

Table 3.1: Used categories for titers measured by method A and method B

category	method A in U/ml	method B in $VIEU$
negative	<7	<63
borderline	7-25	63-126
positive	>25	>126

As the data do not origin from a planned clinical trial, it is necessary to define properly a population on which the statistical analysis will be based. Background knowledge about TBE is taken into consideration. First all 35 cases, where former contact with other Flaviviruses was stated, are excluded to avoid biased results due to cross-reactivity of Flaviviruses in ELISA tests. The main interest in this study is the evolution of TBE titer after basic immunization. Therefore only those few data records are used, where the basic immunization is definitely completed. After this exclusions a data collection consisting of 430 data records is obtained where 250 records are measured by method A and 180 are measured by method B. This is going to be the sample of interest on which the following analyzes are based.

3.2 Restricting the analysis to the first measurements

In this section data is split in a part with only measurements by method A and another part with only measurements by method B. To guarantee that observations are independent from each other only the first measurement of each participant is used. This results in 213 first measurements by method A (sample A1) and 159 first measurements by method B (sample B1). In sample A1 66.20% and in B1 61.78% of the observations are from female participants. The slight difference between the groups is not significant (χ^2-, Fisher's exact test). In table 3.2 age at time of the titer test, time since last vaccination and number of booster vaccinations are compared for the samples A1 and B1. For none of these variables a significant difference between the samples A1 and B1 can be proved (Mann-Whitney-U-, χ^2- test).
Table 3.3 shows the frequency for the categorized variable. The relative frequency for seropositive titer is almost equal for both samples. A Fisher's Exact test did not detect a significant difference between sample A1 and sample B1. As there are only 9 cases in the negative category it is better for further analyzes to combine the classes borderline and negative in one category. Another

Table 3.2: Descriptive statistics of age (in years) at the time of titer test, time (in years) since last vaccination and number of booster vaccinations for the first measurements in the split data

age (in years)	min	$Q_{0.25}$	median	$Q_{0.75}$	max	mean	std	n	missing
A1	5	24	37	46	85	36.98	18.31	213 (100%)	0
B1	5	24	35	48	75	35.99	17.99	159 (100%)	0
time since last vacc. (in years)									
A1	0.01	3.00	4.75	6.00	11.00	4.83	2.34	194 (100%)	19
B1	0.01	3.00	4.00	5.00	15.05	4.46	2.30	152 (100%)	7
number of booster vacc.									
A1	0	0	2	4	6	2.14	1.93	142	71
B1	0	0	1	3	5	1.71	1.53	126	33

Table 3.3: Frequency analysis of categories for the first measurements in the split data

category	positive	borderline	negative	n
A1	193 (90.61%)	17 (7.98%)	3 (1.41%)	213 (100%)
B1	145 (91.19%)	8 (5.03 %)	6 (3.77%)	159 (100%)

positive effect of this procedure is that the relative frequencies in group A1 and B1 become almost equal.

3.2.1 Investigation of dependence of titer on other variables

The goal of this thesis is to describe the dependence of TBE antibody concentration on the measured covariates sex, number of booster vaccinations, age and time since last vaccination. Comparative boxplots were created in figure 3.1 and figure 3.2 to compare the distribution of measured titers by sex and number of booster vaccinations. According to Vittinghoff *et al.* (2005, p.12) a boxplot provides information on location, spread, range of observations, presence of outliers and some information about the shape of the distribution. The box sizes are proportional to the number of observations in the group.
In figure 3.3 and figure 3.4 scatterplots for logarithm of titers depending on age at the time of titer test and time since last vaccination are printed. The logarithmic scale was chosen to get better comparability of titers measured by method A and titers measured by method B. The two horizontal lines represent the borders for the categories seronegative and borderline, according to table 3.1. A non-parametric scatterplot smoother is added to get a better impression on the trend of the logarithm of titer depending on age at the time of titer test and time since last vaccination. For both methods logarithm of titer decreases with increasing age and time since last vaccination. After investigating the dependency of the outcome variable titer on each of the covariates sex, number of booster vaccinations, age and time since last vaccinations following tendencies can be summarized:
- Women have lower titers than men (at least for method A)
- Titers increase with increasing number of booster vaccinations
- Titers decrease with increasing age
- Titers decrease with increasing time since last vaccination

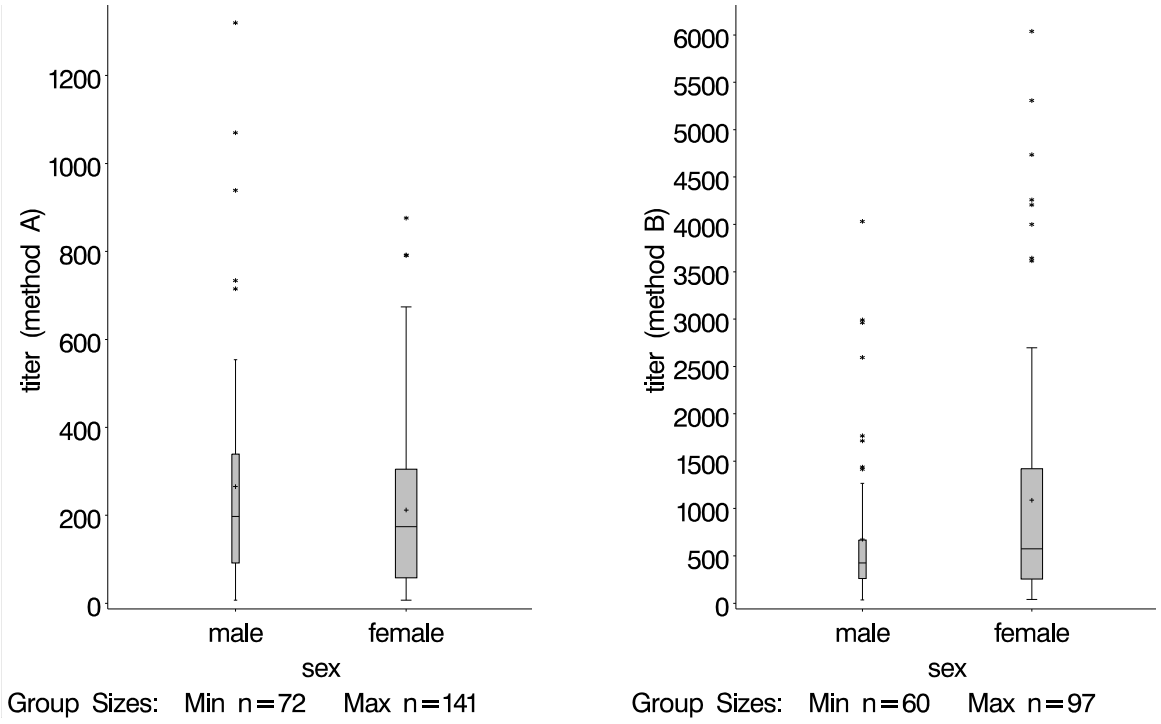

Figure 3.1: Boxplots of titers measured by method A (left) and method B (right) for female and male participants

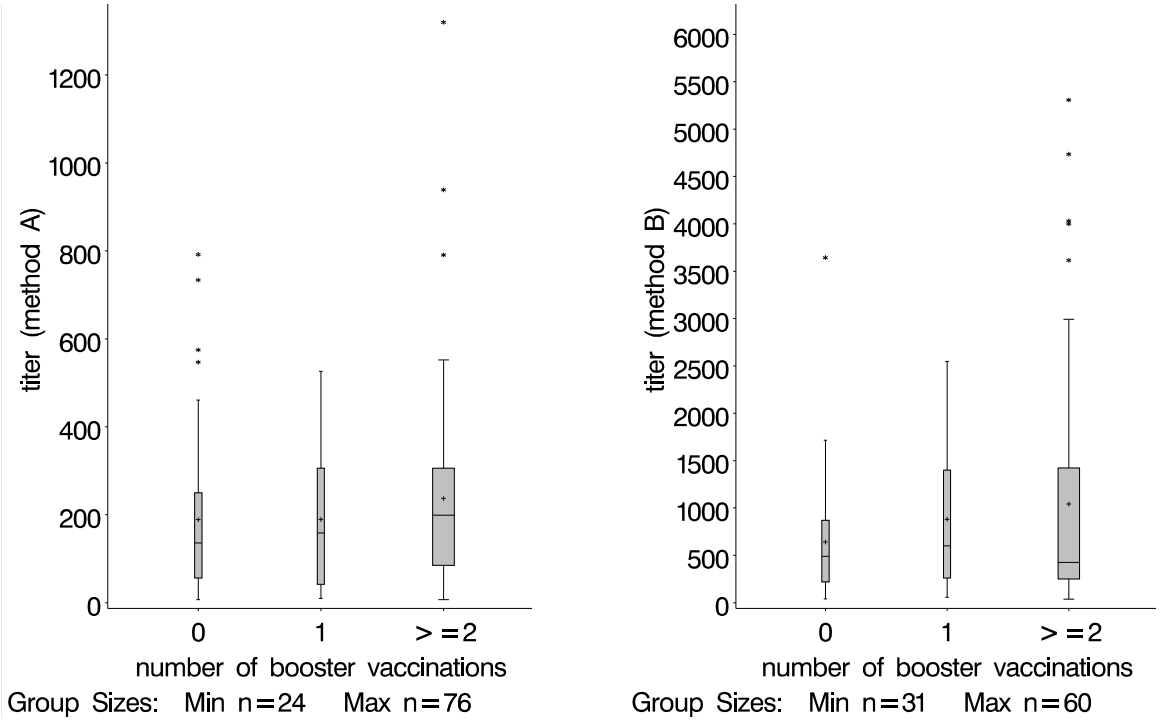

Figure 3.2: Boxplots of titers measured by method A (left) and method B (right) for different numbers of booster vaccinations

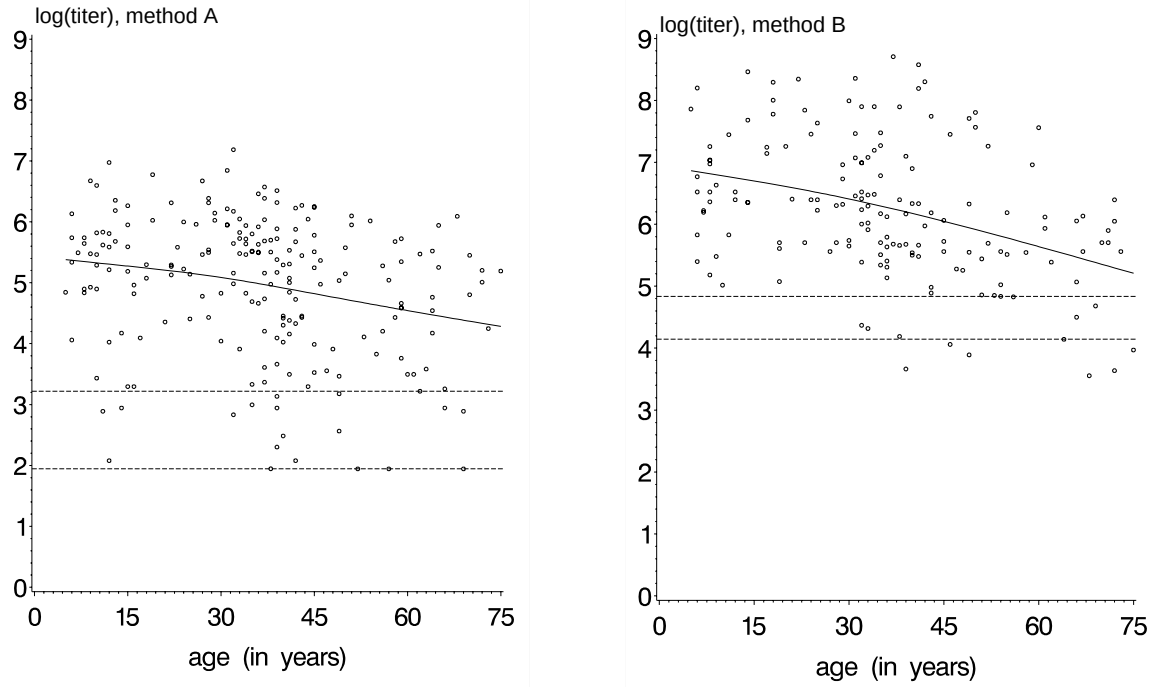

Figure 3.3: Scatterplots of logarithm of titers measured by method A (left) and method B (right) against age at the time of titer test

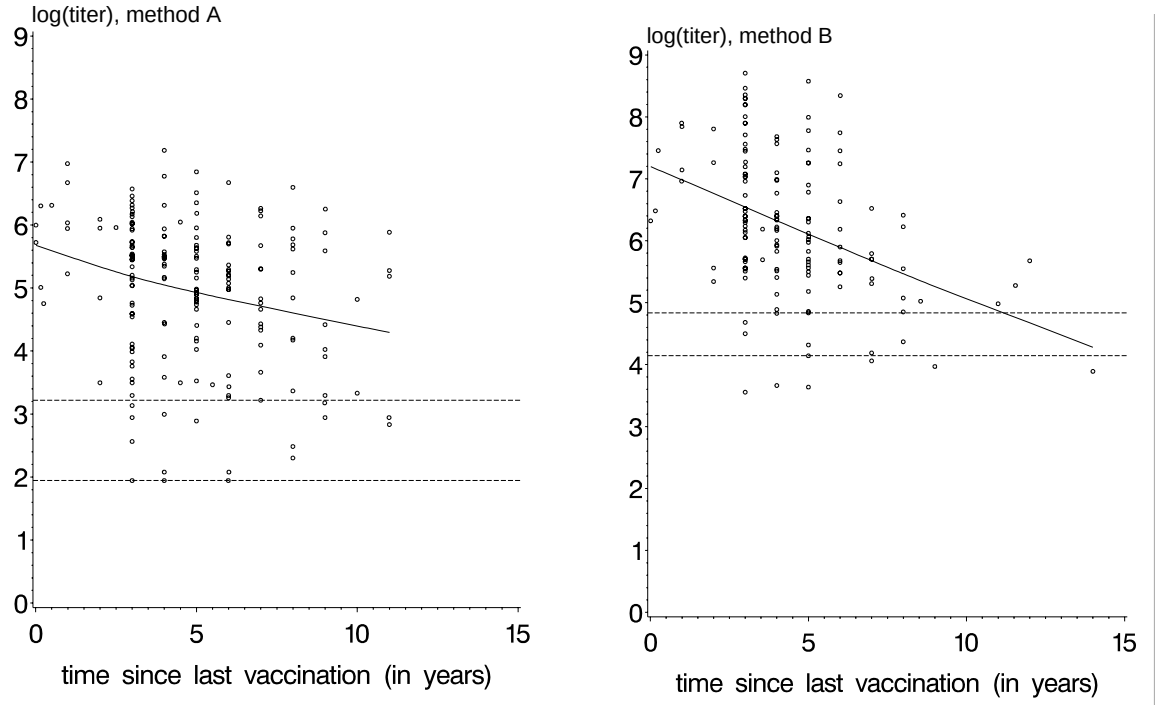

Figure 3.4: Scatterplots of logarithm of titers measured by method A (left) and method B (right) against time since last vaccination

Chapter 4

Linear regression

In the following chapter linear regression analysis is applied to investigate the relation between titer and the covariates sex, number of booster vaccinations, age and time since last vaccination in the sample A1.

4.1 The model

The outcome variable *titer* as well as the predictors age and time since last vaccination can be considered as continuous. From now on the variable age represents the age of the participant at her/his last vaccination. Sex is a nominal variable with two values (0 = male and 1 = female). The variable number of booster vaccinations is divided into three categories (0 = zero booster vaccinations, 1 = one booster vaccination, 2 = two or more booster vaccinations) and *dummy variables* have to be used. Figure 4.1 shows a scatterplot matrix of these variables.

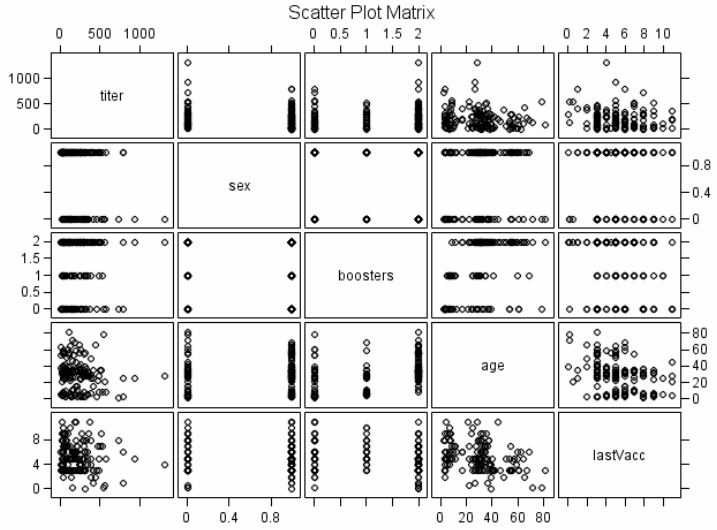

Figure 4.1: Scatterplot matrix of titer, sex, number of booster vaccinations, age and time since last vaccination

Supposing that the Gauss-Markov conditions hold, the following linear regression model is as-

sumed

$$\begin{aligned} titer_i &= \beta_0 + \beta_1 sex_i + \beta_2 boosters1_i + \beta_3 boosters2_i + \beta_4 age_i + \beta_5 lastVacc_i + \varepsilon_i \quad (4.1) \\ \varepsilon_i &\sim \boldsymbol{N}(0, \sigma^2) \quad i.i.d. \end{aligned}$$

4.2 Estimation of parameters

The estimated parameters for model (4.1) are given in table 4.1.

Table 4.1: Estimated parameters for regression model (4.1)

n=140, missing=73

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	406.97	63.55	6.40	<.0001
sex 1	1	-68.50	35.19	-1.95	0.0537
boosters 1	1	3.50	49.88	0.07	0.9442
boosters 2	1	81.52	42.04	1.94	0.0546
age	1	-3.06	1.01	-3.04	0.0029
lastVacc	1	-18.50	7.34	-2.52	0.0129

It is estimated that female participants have lower titer than men. The expected titer increases with the number of booster vaccinations. These results are not significant (significance-level = 0.05) but both parameters for the continuous predictors age and time since last vaccination are significant. The estimated error variance is very high, $\hat{\sigma}^2 = 36906$.

4.3 Model validation

The coefficient of determination $R^2 = 0.1210$ is very low and indicates that the variability of titer is poorly explained by the variables sex, number of boosters, age and time since last vaccinations. Although accounting for the observed covariates a high degree of heterogeneity in the data , so-called *unobserved heterogeneity*, is left. Several interactions were tested if their inclusion would improve the model, but t-tests did not show any significance. Hence, the model with only main effects of the predictors is analyzed in the next steps. In the next step the model is checked for validity by analyzing the residuals.

Normality assumption

Histogram and normal quantile plot in figure 4.2 show a strong departure from normality and transformations are applied to the outcome variable titer. Among the transformations $\log(titer)$, $\sqrt{titer}$, $titer^{1/3}$ the cubic root is most satisfying and the model

$$\begin{aligned} titer_i^{1/3} &= \beta_0 + \beta_1 sex_i + \beta_2 boosters1_i + \beta_3 boosters2_i + \beta_4 age_i + \beta_5 lastVacc_i + \varepsilon_i \quad (4.2) \\ \varepsilon_i &\sim \boldsymbol{N}(0, \sigma^2) \quad i.i.d. \end{aligned}$$

is estimated. Histogram and normal quantile plot of this model do not have any severe departures from the normality assumption anymore, see figure 4.3.

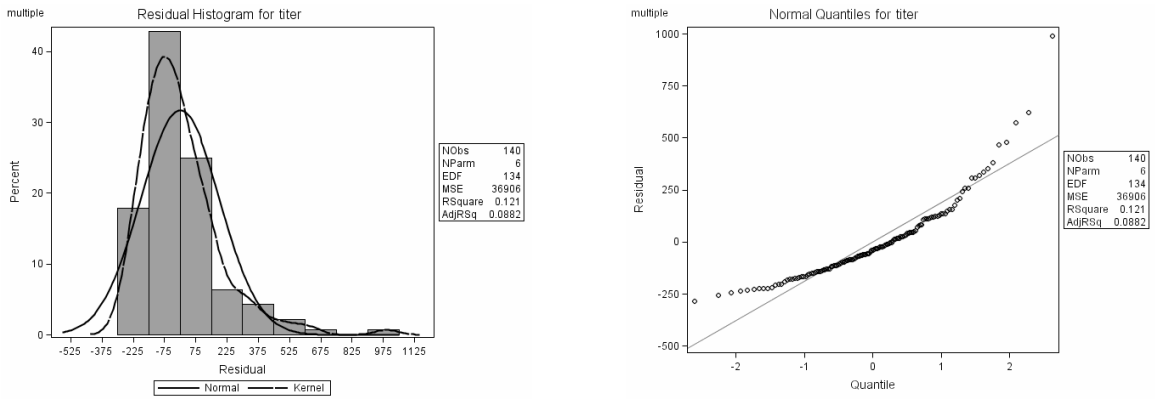

Figure 4.2: Check of normality assumptions for regression analysis of titer by histogram (left) and normal quantile plot (right) of the residuals

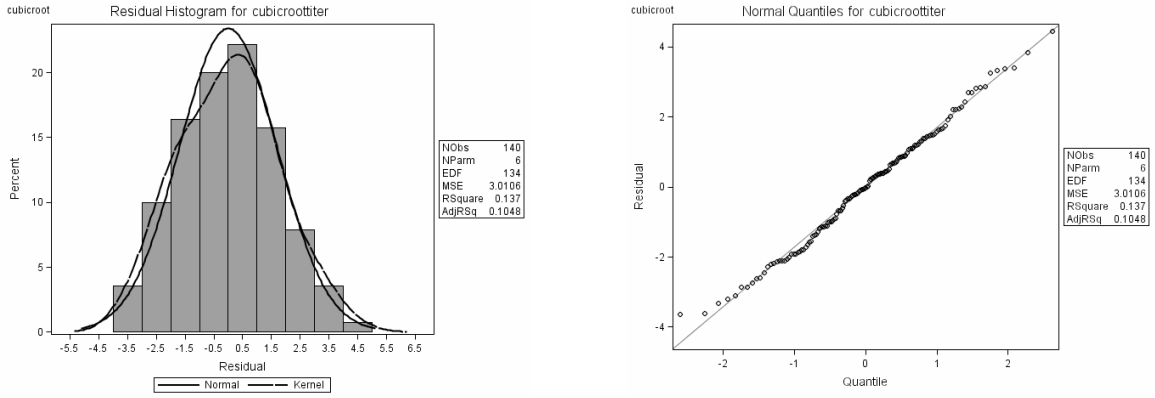

Figure 4.3: Check of normality assumptions for regression analysis of cubic root titer by histogram (left) and normal quantile plot (right) of the residuals

Independence assumption

Plots of residuals against observation number for the regression model on cubic root titer (4.2) did not show a visible structure and the assumption of independence among observations seems appropriate.

Linearity assumption

For checking the linearity assumption in model (4.2) residuals have been plotted against each continuous predictor. The linearity assumption for age seems adequate. The plot of residuals against time since last vaccination shows a slightly quadratic trend, that might be due to only a few observations with less than 3 years since last vaccination. A regression model including the quadratic term $lastVacc^2$ was fitted, but t-test for this new parameter was not significant.

Assumption of constant variance

The important assumption of homoscedasticity is checked by plotting the residuals against the fitted values for both models (4.1) and (4.2), see figure 4.5. The residuals in the left panel clearly have the form of a funnel whereas the residuals in the right panel seem quite unstructured. The

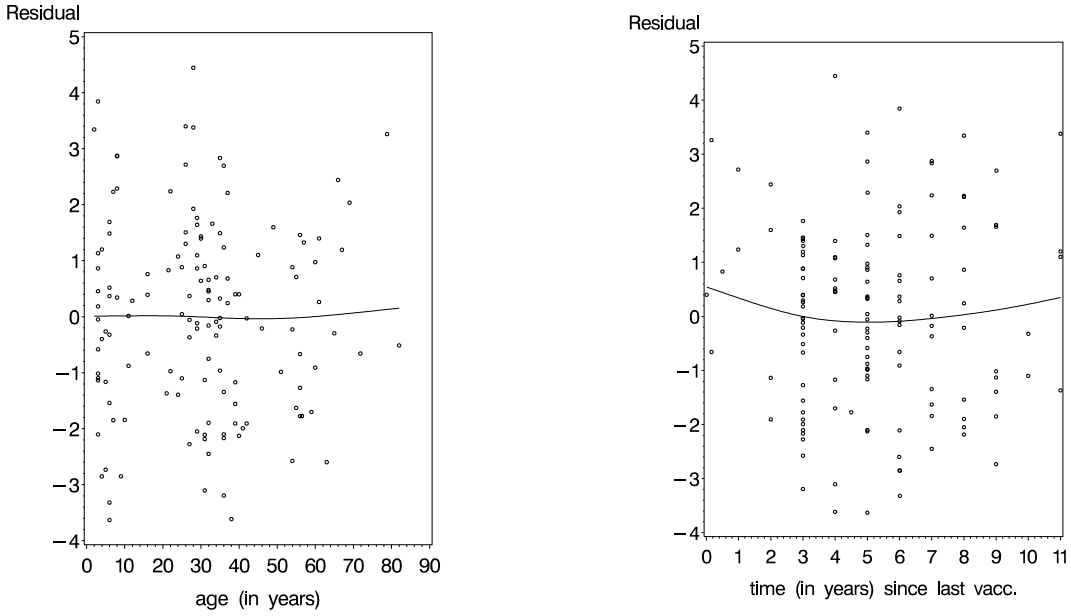

Figure 4.4: Plot of residuals against continuous predictors age (left) and time since last vaccination (right) for regression analysis of cubic root titer

cubic root transformation stabilizes the variance.

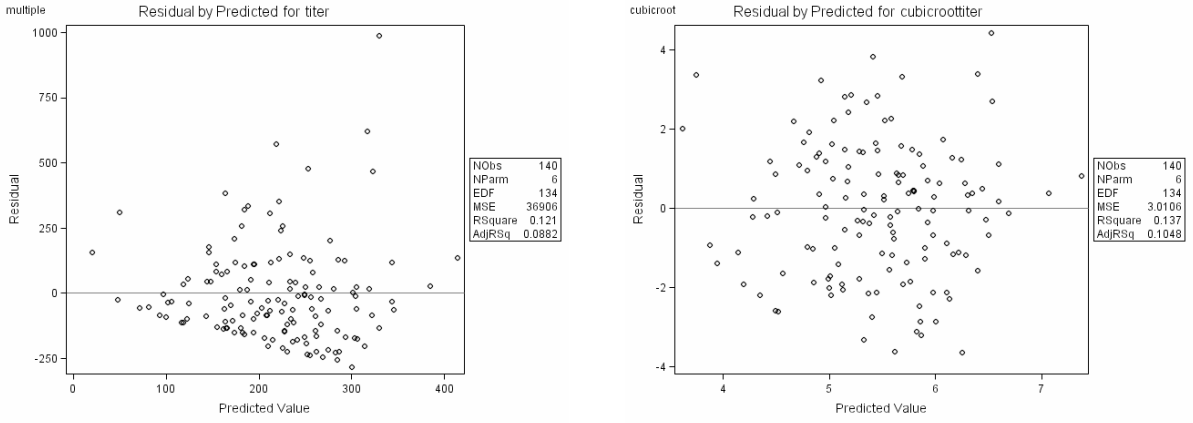

Figure 4.5: Plot of residuals against fitted values for regression analysis of titer (left) and cubic root titer (right)

Outliers, influential points

At last influential points are investigated. In table 4.2 leverage, residuals, studentized residuals and DFFIT-statistics are printed for outstanding observations in both models (regression on *titer* and $titer^{1/3}$). Leverages are equal because the design matrix X is the same in both models. Subjects ($ID = 769, 917, 1574, 3151$) have been identified as influential in both models. Residuals and influential statistics are not displayed for observations which are only influential in the other model. Subjects $ID = 769, 917, 3151$ can be referred to as x-outliers. Measurement #42 ($ID = 769$) refers to a 79-year-old man with a titer of 574 U/ml after 0 booster vaccinations and the last vaccination took place 1.5 months before the test. Observation #50 ($ID = 917$)

Table 4.2: Influence diagnostics for regression analysis of *titer* and $titer^{1/3}$

obs	ID	model	leverage	residual	stud. residual	DFFITS
7	69	*titer*	0.045	468.913	2.548	0.551
33	703	$titer^{1/3}$	0.064	-3.630	-2.193	-0.573
42	769	*titer*	0.157	384.217	2.209	0.952
42	769	$titer^{1/3}$	0.157	3.261	2.071	0.893
50	917	*titer*	0.078	310.706	1.696	0.493
50	917	$titer^{1/3}$	0.078	3.379	2.052	0.597
71	1326	*titer*	0.035	573.710	3.140	0.601
90	1461	*titer*	0.032	622.565	3.423	0.624
105	1574	*titer*	0.033	991.185	5.864	1.084
105	1574	$titer^{1/3}$	0.033	4.447	2.665	0.493
195	3151	*titer*	0.051	481.148	2.627	0.608
195	3151	$titer^{1/3}$	0.051	3.344	2.000	0.463

was measured on a woman with no booster vaccination who had her last vaccination 11 years ago at the age of 28. Observation #195 belongs to a boy with zero booster vaccinations who had his last vaccination at the age of 2, 8 years ago. Finally measurement #105 ($ID = 1574$) was detected as influential because it has the largest residual. This observation was obtained on a man with two or more booster vaccinations who had his last vaccination 4 years ago at the age of 28.

4.4 Conclusion on regression models

It seems that the model for cubic root transformed titers measured by method A fulfills the requirements for a linear regression model best. But for this model it is theoretically possible that a negative value for $titer^{1/3}$ is predicted. This is not reasonable because titer can only be positive and therefore the cubic root of titer has also to be positive. For a woman with no booster vaccination who had her last vaccination 30 years ago at the age of 50 the expected value for cubic root of titer is

$$E(titer^{1/3}) = 7.24 - 0.62 * 1 + 0.12 * 0 + 0.85 * 0 - 0.03 * 50 - 0.19 * 30 = -0.49$$

Although this case is very unlikely to observe it would be better to create a model which takes into account that the outcome variable can only be positive. Furthermore cubic root of titer is hard to interpret and linear regression was also applied to the logarithm of titer. Following parameters have been estimated:

$$E(\log(titer)_i) = 5.96 - 0.38sex_i + 0.10boosters1_i + 0.54boosters2_i - 0.02age_i - 0.11lastVacc_i$$

However, the coefficient of determination R^2 is only 0.1370 in the first and 0.1339 in the second case and probably a more complicated model is required to explain more of the variability of titers. As already mentioned in earlier sections a very flexible kind of modelling is provided by general linear models (GLMs). This model family is presented in the next chapter.

Chapter 5

Generalized linear models

Obviously the untransformed outcome variable titer in U/ml do not fulfill the requirements for a linear regression. Titer can only have a positive value and the distribution is strongly right-skewed. Cubic root of titer appeared to be an appropriate transformation, but is hard to interpret. An analysis for the outcome variable on the original measured scale is desirable and is provided by the family of generalized linear models (GLM), which covers linear and nonlinear models for normal or non-normal, discrete or continuous outcome variables (Montgomery, 2001, p.594). A further relaxing feature of GLMs in comparison with linear regression is, that the variance of the outcome variable does not have to be constant. Instead, the variance can be modelled as a function of the mean.
In this chapter essential ideas of GLMs for continuous and discrete outcomes are presented. For further reading McCullagh & Nelder (1989), one of the standard books on GLMs, is recommended. An introduction to theory and concepts of GLMs is presented by Dobson (2002) and an applied approach is given in Olsson (2002).

5.1 The model

In linear regression models the expected value of an individual observation $E(y_i)$ is modelled as a linear combination of the predictors. In GLMs any monotonic and differentiable function $g(\cdot)$ of the expected value can be modelled as a linear combination of the predictor variables. This function is called *link-function* because it "links" the expected value for the outcome variable to the data.

$$g\left(E(y_i)\right) = g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} \tag{5.1}$$

The distribution of the outcome variable Y in GLMs is assumed to belong to the exponential family which includes e.g. normal, Poisson, binomial, exponential and gamma distribution. The general form for the density of a distribution belonging to the exponential family is:

$$f(y) \equiv f(y|\theta,\phi) = c\left(y,\phi\right) \cdot \exp\left(\frac{\theta y - b(\theta)}{\phi}\right), \tag{5.2}$$

where θ and ϕ are unknown parameters which are often called *natural parameter* and *scale parameter*, respectively. The functions $b(\cdot)$ and $c(\cdot,\cdot)$ are known and determine the form of the

distribution. First and second moment for this distribution are given by

$$E(Y) = b'(\theta) \tag{5.3}$$
$$V(Y) = \phi \cdot b''(\theta) \tag{5.4}$$

and combination of both leads to following relationship between mean and variance:

$$V(Y) = \phi b''(\theta) = \phi b'' \left(b'^{-1}(E(Y))\right) = \phi v(E(Y)), \tag{5.5}$$

where $v(\cdot)$ is called *variance function*, see for example Molenberghs & Verbeke (2005, p.27f).
It is assumed that each outcome Y_i is from the same distribution with density $f(y_i|\theta_i, \phi)$. Each observation can have a different natural parameter θ_i but the scale parameter ϕ has to be the same for all observations. The natural parameter θ_i can be inserted in expression (5.1).

$$g(E(y_i)) = g(b'(\theta_i)) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$$

In most applications $g^{-1}(\cdot) = b'(\cdot)$ is used and $g(\cdot)$ is called the *natural link-function*, because then the natural parameter follows a linear regression model:

$$\theta_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}.$$

Distribution, link-function and predictors have to be specified in GLMs. For choosing a reasonable distribution for the outcome variable, the histogram of titer is compared with several probability distribution functions in figure 5.1. The normal distribution would clearly not be an appropriate assumption for the distribution of titer. The density of the logarithmic normal distribution has a very high peak for low titers. Exponential, Weibull and gamma distribution seem to be more appropriate. The Weibull distribution can not be brought into the canonical form of the exponential family and exponential distribution is just a special case of gamma distribution. It is reasonable to choose the gamma distribution which is appropriate for non-negative variables with long right-tailed distributions. The density for a gamma distribution $\mathbf{\Gamma}(\lambda, \mu)$ with a shape parameter λ and a scale parameter μ is defined as:

$$f(y) = \frac{y^{\lambda-1}}{\Gamma(\lambda)} \frac{1}{\mu^\lambda} \exp\left(-\frac{y}{\mu}\right), \tag{5.6}$$

where $\Gamma(\cdot)$ denotes the gamma function. To get the form of an exponential family the parametrization $\mathbf{\Gamma}(\lambda, \frac{\nu}{\lambda})$ is used. A lot of different parameterizations exist for the gamma distribution and cause a lot of confusion. However, the density can be rewritten as:

$$f(y) = \frac{y^{\lambda-1}}{\Gamma(\lambda)} \left(\frac{\lambda}{\nu}\right)^\lambda \exp\left(-\frac{\lambda y}{\nu}\right) = \frac{y^{\lambda-1}\lambda^\lambda}{\Gamma(\lambda)} \exp\left(\frac{-\frac{1}{\nu}y - \log\nu}{\frac{1}{\lambda}}\right) \tag{5.7}$$

The natural parameter θ and the scale parameter ϕ equals $-\frac{1}{\nu}$ and $\frac{1}{\lambda}$, respectively. The function $b(\theta)$ has the form $-\log(-\theta)$. According to (5.3), (5.4) and (5.5) expectation, variance and

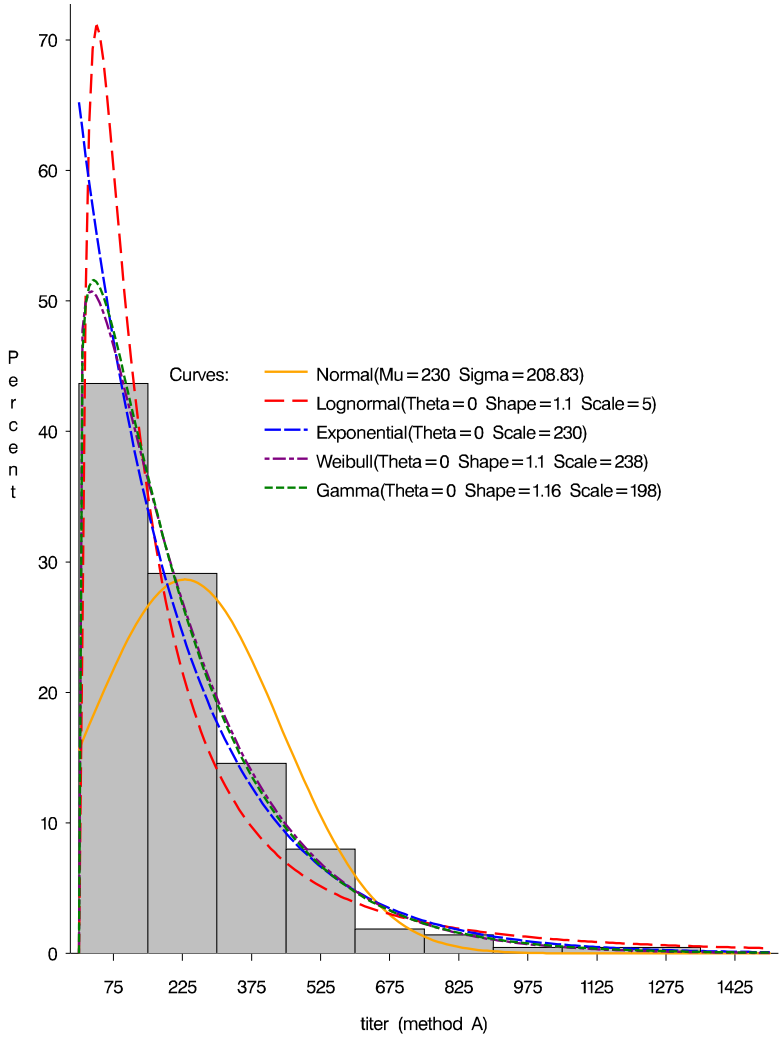

Figure 5.1: Histogram of titer in U/ml compared with several distributions

variance function have the following form:

$$\begin{aligned} E(Y) &= b'(\theta) = -\frac{1}{\theta} = \nu = \lambda\mu \\ V(Y) &= \phi b''(\theta) = \frac{1}{\lambda\theta^2} = \frac{\nu^2}{\lambda} = \lambda\mu^2 \\ v(E(Y)) &= E(Y)^2 = \lambda^2\mu^2 \end{aligned} \tag{5.8}$$

In conclusion the standard deviation depends proportionally on the mean:

$$std(Y) = \sqrt{V(Y)} = \sqrt{\frac{1}{v(E(Y))}} = \frac{1}{\sqrt{\lambda}} \cdot E(Y)$$

The variance in (5.8) is μ-times the expectation. If a residual plot for a positive outcome variable shows that the standard deviation is increasing proportionally to the mean, it is reasonable to

choose the gamma distribution. (Vittinghoff *et al.* , 2005, p.297)
The expected value of a gamma-distributed variable has to be positive. To ensure that the predicted mean as a function of the predictors is positive, the log-link-function for the expected value of titer depending on the predictors sex, number of boosters, age and time since last vaccination is used.

$$\log(\mu_i) = \log(E(titer_i)) = \beta_0 + \beta_1 sex_i + \beta_2 boosters1_i + \beta_3 boosters2_i + \beta_4 age_i + \beta_5 lastVacc_i \quad (5.9)$$

5.2 Estimation of parameters

Usually the parameters $\boldsymbol{\beta}$ in GLMs are estimated by maximum likelihood (ML) method. ML estimation is a widely used estimation method where the likelihood function of the observed data is maximized. The likelihood function reflects how probable it is to observe the actual data under the assumed model. Therefore the roles of parameters and random variables in the density function are simply switched. The random variable is held fix (using the observed data) and the parameters are varied.
For estimating GLMs the log-likelihood function

$$l(\boldsymbol{\beta}, \phi) = \frac{1}{\phi} \sum_{i=1}^{n} (\theta_i y_i - b(\theta_i)) + \sum_{i=1}^{n} \ln(y_i, \phi) \quad (5.10)$$

is maximized by setting the first order derivative with respect to the parameter-vector $\boldsymbol{\beta}$ to zero. As θ_i are functions of $\boldsymbol{\beta}$ and $\mu_i = \frac{\partial b(\theta_i)}{\partial \theta_i}$ we get

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{\partial \theta_i}{\partial \boldsymbol{\beta}} \cdot (y_i - \mu_i) = \sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \cdot v_i^{-1} \cdot (y_i - \mu_i) = 0 \quad (5.11)$$

The last equation in (5.11) is obtained by considering $v_i = v(\mu_i) = b''(\theta_i) = \frac{\partial^2 b(\theta_i)}{\theta_i^2}$ and $\frac{\partial \mu_i}{\partial \boldsymbol{\beta}} = v_i \cdot \frac{\partial \theta_i}{\partial \boldsymbol{\beta}}$. The obtained equations in (5.11) are called *score equations* and can be solved iteratively by e.g. Newton-Raphson algorithm or Fisher scoring.

Table 5.1: Estimated parameters for GLM model (5.9) with assumed gamma distribution and log-link-function

n=140, missing=73

			Standard	Wald 95%			
Parameter	DF	Estimate	Error	Conf. Limits		χ^2	Pr > χ^2
Intercept	1	6.234	0.300	5.647	6.821	433.15	<.0001
sex 0	0	0	0	0	0	.	.
sex 1	1	-0.341	0.161	-0.658	-0.025	4.48	0.0343
boosters 0	0	0	0	0	0	.	.
boosters 1	1	-0.004	0.228	-0.451	0.443	0.00	0.9854
boosters 2	1	0.319	0.184	-0.040	0.679	3.03	0.0819
age	1	-0.014	0.004	-0.023	-0.006	10.65	0.0011
lastVacc	1	-0.081	0.032	-0.144	-0.018	6.41	0.0113
Scale	1	1.290	0.139	1.045	1.592		
	NOTE: The scale parameter was estimated by maximum likelihood.						

Estimated parameters for model (5.9) are given in table 5.1 and the expected value of titer is given by:

$$\begin{aligned} E(titer_i) &= \exp(6.23 - 0.34sex_i - 0.004boosters1_i + 0.32boosters2_i - 0.01age_i - 0.08lastVacc_i) \\ &= e^{6.23}e^{-0.34sex_i}e^{-0.004boosters1_i}e^{+0.32boosters2_i}e^{-0.01age_i}e^{-0.08lastVacc_i)} \end{aligned} \quad (5.12)$$

Equation (5.12) clearly shows that the model is multiplicative. In contrary to an additive model, where the absolute change of the outcome variable is measured, multiplicative models measure the relative change. For example it is estimated that a woman in average has only 71% of a man's titer given the other predictor variables are held constant. If age is increased by one year titer reduces by 1%. All parameters except the number of booster vaccinations are significant (significance-level = 0.05). The estimated confidence interval for the scale parameter in table 5.1 does not cover 1. The hypothesis that an exponential distribution would be adequate is rejected.

5.3 Model validation

(Scaled) deviance and Pearson-χ^2 can be used for assessing goodness of fit in GLMs. The deviance (=121.89) is roughly equal to the degrees of freedom (=134) and indicates a quite fair model fit. However, for validation of the model a residual analysis has to be carried out.
The plot of standardized deviance residuals against time shows again a slightly departure from the linearity assumption on time since last vaccination. All other residual plots do not show worrying departures from the model assumptions and the specified model seems to be adequate. Methods for checking the choice of link and variance function are explained in McCullagh & Nelder (1989).

5.4 Conclusion on GLMs

Fitting a GLM based on gamma distribution with log-link-function seems to be quite a good model for the data. This coincides with the linear regression model on $titer^{1/3}$ in section (4.1), because it is known that an accurate normalizing transformation for a gamma distributed variable Y is

$$3\left(\left(\frac{Y}{\nu}\right)^{1/3} - 1\right),$$

where ν is the parameter used in (5.7), see McCullagh & Nelder (1989, p.289). Also discrete outcome variables can be modelled with GLMs and were applied to the binary variables indicating seropositive titer for both, titers measured by method A and titers measured by method B. Dichotomizing the variable titer induces a loss of information. The idea behind modelling the less informative binary variable was to find a possibility for modelling the outcome variables of method A and method B together, but the results of GLMs on the binary variable for method B were quite different from GLMs on the binary variable for method A and a common analysis would not be reasonable. Method A is more commonly used in the laboratories and further analyzes are concentrated on modelling titers measured by method A.

Chapter 6

Repeated measures analysis

In the previous chapters models for independent observations were presented and only the first measurement of each person was used in analyzes to guarantee independence between observations. Now, all available measurements should be included and models are required which take a possible dependence between repeated measurements on the same observational unit into consideration. The umbrella term for modelling repeatedly measured data is *repeated measures analysis*.
Analysis of repeated measures in general is described in Crowder & Hand (1990). More specific books on analysis of longitudinal data especially in health sciences are provided by Diggle (2002) and Fitzmaurice *et al.* (2004). Very extensive books on linear mixed models and models for discrete longitudinal data are given by Verbeke & Molenberghs (2000) and Molenberghs & Verbeke (2005), respectively. In the last three books SAS program code is printed as well as in Brown & Prescott (1999). A lot of books on longitudinal analysis exist and only some of them have been picked out.

6.1 Features of longitudinal data

Longitudinal data are obtained by taking several measurements on a subject at different points of time. Observations are no longer independent. In Fitzmaurice *et al.* (2004, p.36ff.) three sources of correlation in longitudinal data are mentioned:

Between-individual heterogeneity: Especially in health sciences there is a lot of natural heterogeneity among individuals due to genetic, environmental, social and behavioral factors. It is expected that repeated measurements on the same subject are more similar than observations across different individuals.

Within-individual biological variation: Most health-related variables vary considerably over time. Circadian rhythms, temperature, light, season, diet, infections etc. might have an influence. It is assumed that random departures from the underlying biological process are more similar for measurements close together in time. Serial correlation is assumed to decrease over time between repeated measures on an individual.

Measurement error: Random measurement errors occur in almost all studies. The precision of measurements can be expressed by terms of the variance of the measurement errors. If

this variance is large the correlation among repeated measurements appears lower than it would be in a study where more reliable measurements were obtained.

Ignoring the correlation in longitudinal data results in misleading inference on regression coefficients, less efficient estimates and less protection against bias caused by missing data. (Diggle, 2002, p.19). If the number of repeated observations is the same for all individuals taken at a common set of occasions, longitudinal data are called *balanced* (Fitzmaurice *et al.* , 2004, p.23). In retrospective studies unbalanced data are common. Incomplete data require some care and a lot of literature exists on the topic of missing data. In health sciences longitudinal data are rarely complete and balanced over time and methods for longitudinal analysis should be able to handle those specifics. Once more the goal of longitudinal analysis should be emphasized:

> "In summary, the fundamental objective of a longitudinal analysis is the assessment of within-individual changes in the response and the explanation of systematic differences among individuals in their changes." (Fitzmaurice *et al.* , 2004, p.21)

6.2 Explorative analysis of repeatedly measured data

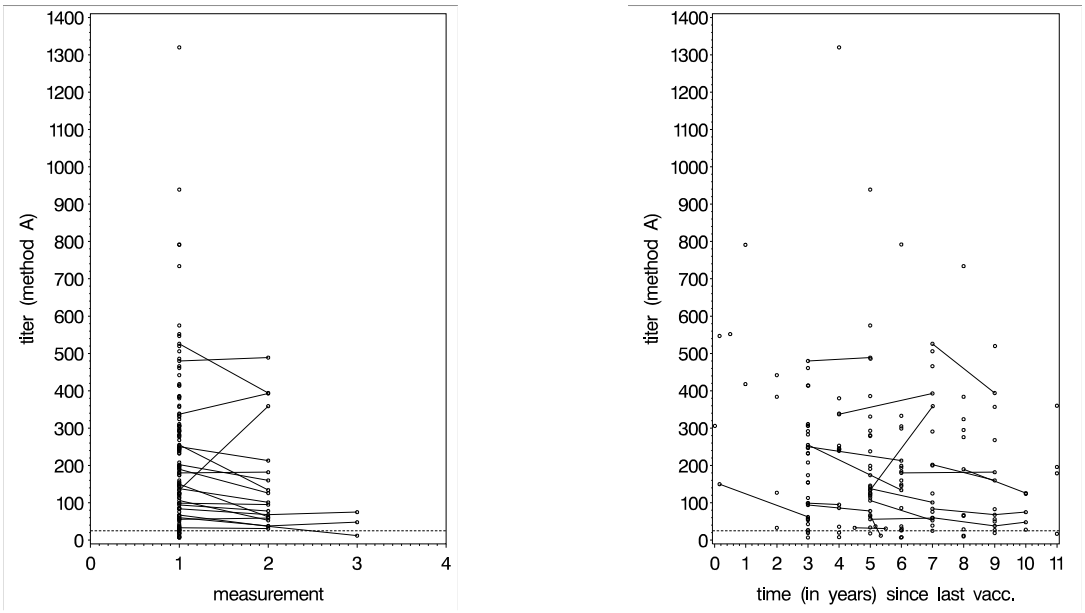

Figure 6.1: Distribution of all 162 available titers at each occasion (left) and against time since last vaccination (right). Repeated measurements on the same person are connected.

The repeated measures in the TBE data at hand belong to the class of longitudinal data, because several persons went to the hospital two or more times for testing their titer. Unfortunately, different tests (method A or method B) were sometimes used on the same person. It has already been mentioned that no formula is known to recalculate results of one test into a measurement of the other test. One idea to deal with this problem was using only categories seropositive and seronegative for titer measurements. But the application of GLMs for discrete variables gave very different results for observations measured by method A and observations measured by method B. The analysis of a pooled sample would not at all be reliable. Hence, only all available measurements by method A are analyzed. There are 3 persons with 3 measurements by method

Table 6.1: Descriptive statistics of titer (in U/ml), age (in years) at the time of titer test and time (in years) since last vaccination at first, second, third measurement

titer (in U/ml)	min	$Q_{0.25}$	median	$Q_{0.75}$	max	mean	std	n
M1	6.99	64.50	176.50	302.00	1320.00	217.44	201.18	140 (100%)
M2	31.00	59.00	101.00	213.00	489.00	161.68	141.96	19 (100%)
M3	12.00	12.00	48.00	75.00	75.00	45.00	31.61	3 (100%)
Total	6.99	60.00	154.50	295.00	1320.00	207.70	195.03	162 (100%)
age (in years)								
M1	5	17.5	36	43.5	85	35.18	18.53	140 (100%)
M2	11	19	45	66	86	45.58	23.97	19 (100%)
M3	20	20	46	57	57	41.00	19.00	3 (100%)
Total	5	18	37	45	86	36.51	19.41	162 (100%)
time since last vacc. (in years)								
M1	0.01	3.00	5.00	7.00	11.00	5.17	2.38	140 (100%)
M2	3.00	5.17	7.00	9.00	10.00	6.83	1.96	19 (100%)
M3	5.34	5.34	10.00	10.00	10.00	8.45	2.69	3 (100%)
Total	0.01	3.00	5.00	7.00	11.00	5.42	2.42	162 (100%)

Table 6.2: Frequency analysis of sex at first, second, third measurement

sex	male	female	n
M1	44 (31.43%)	96 (68.57%)	140 (100%)
M2	8 (42.11%)	11 (57.89%)	19 (100%)
M3	2 (66.67%)	1 (33.33%)	3 (100%)
Total	54 (33.33%)	108 (66.66%)	162 (100%)

A where the covariates *sex*, *boosters*, *age* and *lastVacc* are completely observed. 16 persons have 2 available measurements by method A and 121 persons have only 1 titer observation. All 162 available observations are illustrated in figure 6.1. The 140 first measurements by method A which were analyzed in the preceding chapters are a subset of these repeated measures. Data are highly unbalanced because the number of observations per measurement is quite different. Furthermore the distances between successive measurements vary individually. Although none of the participants got a booster vaccination after their titer tests, titer increases in 7 cases of successive measurements. The outstanding increase by 225 U/ml within 2 years was measured for a boy ($ID = 3022$) last vaccination five years ago at the age of 5. In only one case of successive measurements on a participant, a positive titer becomes negative.

Although measurements are not equidistant and not balanced descriptive summaries are provided for first, second and third measurements to get some hints on the evolution of all variables over time. Table 6.1 summarizes descriptive statistics of the continuous variables titer, age at the time of titer test and time since last vaccination. Average titer and median titer reduces from first to the third measurement. Time since last vaccination increases in average because none of the persons with more than one observation got a vaccination in-between. Average age at the time of titer test would also be expected to increase. The decrease of average age from second to third measurement is due to the little number of observations. At least the median of age increases.

Table 6.3: Frequency analysis of number of booster vaccinations at first, second, third measurement

number of boosters	0	1	≥2	n
M1	40 (28.57%)	24 (17.14%)	76 (54.29%)	140 (100%)
M2	3 (15.79%)	7 (36.84%)	9 (47.37%)	19 (100%)
M3	1 (33.33%)	2 (66.67%)	0 (0.00%)	3 (100%)
Total	44 (27.16%)	33 (20.37%)	85 (52.47%)	162 (100%)

Tables 6.2 and 6.3 present the proportions of sex and number of booster vaccinations at first, second and third measurement. These proportions vary highly due to the small numbers of observation at second and third measurement. Only a third of all 162 measurements origin from men. Half of all observations are made on participants with two or more booster vaccinations.

For exploring the degree of association between successive measurements a scatterplot matrix is printed in figure 6.2 and correlations are calculated. For computation of correlations all available pairs of repeatedly measured titers are used. The correlation between first and second measurement 0.86 is calculated on 19 pairs and the other two correlations are based on 3 pairs. It seems that the correlation between successive measurements is equal and decreases with increasing measurements in-between.

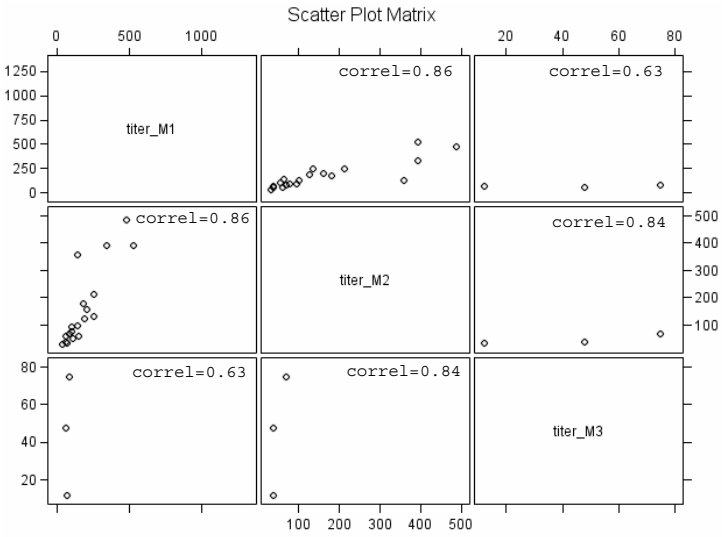

Figure 6.2: Correlation matrix of titers at at first, second, third measurement

For the TBE data a selection effect on repeated measurements can be observed. Persons who had a very low titer at their first measurement go rather for a booster vaccination than to do another blood test. Repeated measurements are only available of people who expect having a high titer. One reason for this phenomenon is that the costs for vaccination and blood test are almost the same. For instance the vaccine in Austria costs 22.50 € for adults (ARGE Gesundheitsvorsorge, Download: 2007-03-31b) and the titer test at the microbiological ambulance of the hospital *Allgemeines öffentliches Krankenhauses der Elisabethinen* in Linz costs 25 € .

Chapter 7

Models for longitudinal data in the Gaussian case

Only a few repeated measures are available and the following chapters are rather an outlook over possible models for longitudinal data and a guideline how analysis on repeated data could be performed. Reliable statistical inference for the TBE data is not really possible due to the scarcity of repeated measurements. For further reading on analysis of Gaussian longitudinal data Diggle (2002) and Verbeke & Molenberghs (2000) are recommended. Applied approaches are given in Brown & Prescott (1999) andFitzmaurice *et al.* (2004).
Following notation for longitudinal data is used: On $i = 1, \ldots, N$ subjects an outcome variable Y is measured n_i times per subject i. y_{ij} denotes the j^{th} observation of subject i and $\boldsymbol{x}_{ij} = (x_{ij,1}, \ldots, x_{ij,p})'$ is the corresponding p-dimensional vector of covariate values. The n_i-dimensional vector $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{in_i})'$ represents the vector of successive measurements for subject i and X_i is the corresponding $(n_i \times p)$-matrix of predictor values. Finally, the N-dimensional vector $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N)'$ denotes the vector of all measurements on all subjects and $\boldsymbol{X}$ is the corresponding $(N \times p)$-matrix of covariate values.
In this chapter regression models for Gaussian data discussed in chapter 4 are extended by including repeated measures. Titers have turned out not to be normally distributed, but the cubic root appeared to be a normalizing transformation and models for Gaussian longitudinal data are applied to the outcome variable $titer^{1/3}$.

7.1 Marginal models for longitudinal data in the Gaussian case

Marginal or population-averaged models do not account explicitly for between-individual heterogeneity and are mainly used in population studies like epidemiology. The interest is on differences between groups in the population with different risk factors (Zeger *et al.* , 1988, p.1051).

7.1.1 The model

Most parametric models for continuous longitudinal data are based on normality assumptions. The univariate linear regression model is simply extended to longitudinal data by assuming that each response vector $\boldsymbol{y}_i$ of subject i can be modelled as the multivariate version of a normal

linear regression model.

$$\begin{aligned} \boldsymbol{y}_i &= X_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i \\ \boldsymbol{Y}_i &\sim \boldsymbol{N}(X_i\boldsymbol{\beta}, V_i) \end{aligned}$$

V_i is the covariance matrix for measurements of subject i. The diagonal-elements represent the variances and the off-diagonal-elements denote the covariances for two repeated measurements on subject i. Conclusively, a model for all measurements is given by:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{7.1}$$

$$\boldsymbol{Y} \sim \boldsymbol{N}(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{V}), \tag{7.2}$$

where $\boldsymbol{V}$ is a $N \times N$-block diagonal matrix with the individual covariance matrices V_i on the diagonal and zero elsewhere. This structure assumes that outcomes for different subjects are independent. If all subjects have an equal number of measurements ($n_i = n$ for all $i = 1, \ldots, N$) taken at fixed points of time then the choice of the same general $n \times n$ positive definite covariance matrix for all subjects ($V_i = V$) would be appropriate. In this case the covariance structure is called homogenous. But also heterogeneous version for covariance structures are possible. (Molenberghs & Verbeke, 2005, p.36)
For the TBE data the outcome vector $\boldsymbol{y}_i$ with components $y_{ij} = titer_{ij}^{1/3}$ for each subject i is modelled as a regression model depending on the covariates sex, number of booster vaccinations, age at last booster vaccination and time since last vaccination. Sex is of course constant over time and $lastVacc$ represents the time variable. Number of boosters and age at last booster vaccination do not change over time because none of the participants with repeated observations had a booster vaccination between the measurements. Different covariance pattern models (unstructured, Toeplitz, AR(1), compound symmetry) for V_i were applied and compared by likelihood ratio tests and Akaike information criterion (AIC). It turned out that an auto-regressive covariance structure (AR(1)) is most appropriate.

$$Cov(\boldsymbol{Y}_i) = V_i = \sigma^2 \begin{pmatrix} 1 & \alpha^1 & \alpha^2 & \ldots & \alpha^{n-1} \\ \alpha^1 & 1 & \alpha^1 & \ldots & \alpha^{n-2} \\ \alpha^2 & \alpha^1 & 1 & & \alpha^{n-3} \\ \vdots & \vdots & & \ddots & \vdots \\ \alpha^{n-1} & \alpha^{n-2} & \ldots & & 1 \end{pmatrix} \tag{7.3}$$

7.1.2 Estimation of parameters

Estimation of parameters in marginal models for Gaussian longitudinal data is based on ML-principles. Depending on the choices for the block matrices V_i there might be some additionally unknown parameters $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots)'$, so-called *variance components* which describe aspects of the covariance or correlation. Usually variance components are not known and must be estimated from the data at hand. Primary interest is on the estimation of the regression parameters $\boldsymbol{\beta}$ and variance components are referred to as *nuisance parameters*. Their estimation is of secondary

interest for the research question but they have to be taken into account to assure an appropriate method of analysis (Fitzmaurice *et al.* , 2004, p.72).
If the covariance matrix V_i has been wrongly specified the point estimators for the parameters $\boldsymbol{\beta}$ are still unbiased but the standard errors become incorrect. Valid standard errors for $\hat{\boldsymbol{\beta}}$ can be obtained by the so-called *information sandwich estimator* or *Huber estimator* (Dobson, 2002, p.200), which is robust to specification of the structure of the covariance matrices and is a consistent estimator for the variance of the estimated parameter vector $V(\hat{\boldsymbol{\beta}})$ as long as the mean is correctly specified (Molenberghs & Verbeke, 2005, p.61). In finite samples ML-estimates of V_i have well-known bias because the ML-estimates do not take into consideration that the parameter vector $\boldsymbol{\beta}$ was also estimated from the data. Therefore the method of restricted maximum likelihood (REML) estimation was introduced by Patterson & Thompson (1971) for estimating variance components.
REML-estimated parameters are given in table 7.1 for the marginal model of $titer^{1/3}$ assuming an AR(1) covariance structure. For the standard errors the robust Huber estimator was used.

Table 7.1: Estimated parameters for the marginal model of cubic root of titer with assumed AR(1) covariance structure

number of subjects=140, number of observations=162

Parameter	Estimate	Standard Error	DF	t value	$Pr > \|t\|$
Intercept	7.112	0.517	135	13.76	<.0001
sex 0	0	.	.	.	.
sex 1	-0.643	0.323	135	-1.99	0.0484
boosters 0	0	.	.	.	.
boosters 1	0.177	0.458	135	0.39	0.6996
boosters 2	0.846	0.402	135	2.11	0.0369
age	-0.029	0.009	135	-3.10	0.0024
lastVacc	-0.162	0.049	21	-3.33	0.0032
α	0.92				
σ^2	3.03				

The parameter estimates and the estimated error variance ($\sigma^2 = 3.03$) are quite similar to those of the linear regression model on cubic root of titer in (4.2) in chapter 4. Residual checks do not show any severe departures from the model assumptions.

Marginal models do not take the different sources of variability into account. Furthermore most covariance pattern models are not appropriate for unbalanced data. In the next section subject-specific models are applied for explicitly considering natural differences between individuals and avoiding problems with unbalanced data.

7.2 Linear mixed models for longitudinal data

A different approach to model longitudinal data is provided by random-effects models. These subject-specific models are used when the individual response-evolution of subjects is of interest.

7.2.1 The model

Each subject is supposed to have its own trajectory over time which can be modelled by a linear regression model on the subject-specific regression coefficients $\tilde{\boldsymbol{\beta}}_i = (\tilde{\beta}_{i0}, \ldots, \tilde{\beta}_{ip})$. The j^{th} observation of subject i is assumed to follow the model

$$y_{ij} = \tilde{\beta}_{i0} + \tilde{\beta}_{i1} x_{ij,1} + \ldots + \tilde{\beta}_{ip} x_{ij,p} + \varepsilon_{ij}.$$

The error component ε_{ij} is assumed to be normal distributed with mean zero. As subjects are a random sample from a population of subjects, the subject-specific regression coefficients $\tilde{\boldsymbol{\beta}}_i$ are considered as a sample from a population of regression coefficients. Denoting the departure of the subject-specific regression parameters from the population-averaged parameters by the so-called *random effects* $b_{ij} = \tilde{\beta}_{ij} - \beta_j$ the model can be re-written:

$$y_{ij} = (\beta_0 + b_{i0}) + (\beta_1 + b_{i1}) x_{ij,1} + \ldots + (\beta_p + b_{ip}) x_{ij,p} + \varepsilon_{ij}.$$

Usually random effects are assumed to be normally distributed with mean zero and covariance matrix D. In the above case all covariates were supposed to have a random effect and the model for subject i could be written in matrix form as

$$\boldsymbol{y}_i = X_i \boldsymbol{\beta} + X_i \boldsymbol{b}_i + \boldsymbol{\varepsilon}_i, \text{ where } \boldsymbol{\varepsilon}_i \sim \boldsymbol{N}(\boldsymbol{0}, \Sigma_i) \quad i = 1, \ldots, N$$

The elements in the parameter vector $\boldsymbol{\beta}$ are called *fixed effects* and the elements in the vector of subject-specific regression coefficients $\boldsymbol{b}_i$ are called *random effects*. Models including fixed and random effects are called *mixed models*. In practice some predictors are assumed to be fixed across all individuals and only a subset of covariates is assumed to vary randomly across subjects. The general linear mixed model including a vector of p unknown regression coefficients $\boldsymbol{\beta}$ and a vector of q subject-specific regression coefficients $\boldsymbol{b_i}$ is formulated as

$$\begin{aligned} \boldsymbol{y}_i &= X_i \boldsymbol{\beta} + Z_i \boldsymbol{b}_i + \boldsymbol{\varepsilon}_i, \quad i = 1, \ldots, N \\ &\quad \boldsymbol{b}_i \sim \boldsymbol{N}(\boldsymbol{0}, D) \text{ and } \boldsymbol{\varepsilon}_i \sim \boldsymbol{N}(\boldsymbol{0}, \Sigma_i) \text{ independent} \end{aligned} \tag{7.4}$$

X_i and Z_i are $(n_i \times p)$ and $(n_i \times q)$ design matrices and $q \leq p$. The errors $\boldsymbol{\varepsilon}_i$ are assumed to be independent from the subject-specific effects $\boldsymbol{b}_i$. The error components $\boldsymbol{\varepsilon}_i$ may be regarded as sampling or measurement errors which are uncorrelated and have equal variance across time. In this case Σ_i is assumed to be a diagonal matrix $\sigma^2 I_{n_i}$. (Fitzmaurice *et al.* , 2004, p.195)

Given the subject-specific parameters $\boldsymbol{b}_i$ the conditional model is formulated as:

$$\boldsymbol{Y}_i | \boldsymbol{b}_i \sim \boldsymbol{N}(X_i \boldsymbol{\beta} + Z_i \boldsymbol{b}_i, \Sigma_i) \tag{7.5}$$

$$\boldsymbol{b}_i \sim \boldsymbol{N}(\boldsymbol{0}, D) \tag{7.6}$$

From the conditional model in linear mixed models a marginal model formulation can be easily obtained using the properties of normal distribution. Under the assumption $E(\boldsymbol{b}_i) = \boldsymbol{0}$ and

independence assumption for $\boldsymbol{b}_i$ and $\boldsymbol{\varepsilon}_i$, marginal mean and variance can be obtained by

$$\begin{aligned} E(\boldsymbol{Y}_i) &= E(E(\boldsymbol{Y}_i|\boldsymbol{b}_i)) = E(X_i\boldsymbol{\beta} + Z_i\boldsymbol{b}_i) = X_i\boldsymbol{\beta} + Z_iE(\boldsymbol{b}_i) = X_i\boldsymbol{\beta} \\ V_i = V(\boldsymbol{Y}_i) &= V(X_i\boldsymbol{\beta} + Z_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i) = Z_iV(\boldsymbol{b}_i)Z_i' + V(\boldsymbol{\varepsilon}_i) = Z_iDZ_i' + \Sigma_i. \end{aligned}$$

The marginal model of linear mixed model is given by

$$\boldsymbol{Y}_i \sim \boldsymbol{N}(X_i\boldsymbol{\beta}, Z_iDZ_i' + \Sigma_i) \tag{7.7}$$

This close relationship between conditional and marginal model in linear mixed models results from the specific feature of multivariate normal distributions having normally distributed marginal distributions, see e.g. Hafner (1989, p.119).

Now, a linear mixed model is applied for the transformed variable $titer^{1/3}$. If sufficient repeated measurements on each person would be available individual random intercept and random slope for each participant would be included in the model.

$$titer_{ij}^{1/3} = (\beta_0 + b_{i0}) + \beta_1 sex_i + \beta_2 boosters1_i + \beta_3 boosters2_i + \beta_4 ageM1_i + (\beta_5 + b_{i1}) lastVacc_{ij} + \varepsilon_{ij}$$

Linear mixed models including random intercepts and random slopes are very common in studies of evolution of titers or antibody concentrations. A random intercept takes into account that different individuals reach naturally different antibody concentrations after basic immunization. The random slope takes into account that some people's titer decrease faster than the average and other peoples's titer declines slower over time.
However, for TBE data too few repeated measurements are available for estimating a random slope and only a random intercept is assumed. For the fixed effects the predictors sex, number of boosters, age and time since last vaccinations are used and the model equation for the cubic root transformed titer at occasion j of participant i is:

$$titer_{ij}^{1/3} = \beta_0 + b_{i0} + \beta_1 sex_i + \beta_2 boosters1_i + \beta_3 boosters2_i + \beta_4 ageM1_i + \beta_5 lastVacc_{ij} + \varepsilon_{ij} \tag{7.8}$$

The random intercept is assumed to be normally distributed with mean zero and variance τ^2, $b_{i0} \sim \boldsymbol{N}(0, \tau^2)$. The random errors are also assumed to be normally distributed with mean zero and covariance matrix $\sigma^2 I_{n_i}$, $\boldsymbol{\varepsilon}_i \sim \boldsymbol{N}(\boldsymbol{0}, \sigma^2 I_{n_i})$.

7.2.2 Estimation of parameters

Estimation and inference in linear mixed models is based on the marginal model formulation in (7.7), except the models are fitted in a Bayesian context (Verbeke & Molenberghs, 2000, p.41). The interest is usually directed towards estimating the parameters in $\boldsymbol{\beta}$ in the marginal model. ML-, REML-estimation methods and robust inference as briefly explained in section 7.1 are applied by using $V_i(\boldsymbol{\alpha}) = Z_iDZ_i' + \Sigma_i$. If the research interest is prediction of subject-specific evolutions, estimates for the random effects $\boldsymbol{b}_i$ are needed and the conditional model specification in (7.5) must be considered. As the subject-specific parameters $\boldsymbol{b}_i$ are random they are predicted with Bayesian methods.

Table 7.2: Estimated parameters for the random intercept model of cubic root of titer (7.8)

number of subjects=140, number of observations=162

Parameter	Estimate	Standard Error	DF	t-value	$Pr > \|t\|$
Intercept	7.106	0.511	135	13.91	<.0001
sex 0	0	.	.	.	.
sex 1	-0.640	0.323	21	-1.98	0.0608
boosters 0	0	.	.	.	.
boosters 1	0.175	0.459	21	0.38	0.7063
boosters 2	0.846	0.401	21	2.11	0.0472
age	-0.029	0.009	21	-3.10	0.0054
lastVacc	-0.162	0.047	21	-3.44	0.0024
Variance components	**Estimate**	**Standard Error**		**Z-value**	$Pr > Z$
τ	2.740	0.377		7.27	<.0001
σ^2	0.286	0.088		3.24	0.0006

Estimated fixed effects and variance components for the random intercept model (7.8) are given in table 7.2.
The parameter estimates for the fixed effects in table 7.2 are almost the same as the parameter estimates in the marginal model, see table 7.1. Although a marginal model formulation can be obtained from a random-effects model they are not the same. Different random-effects models can produce the same marginal model and there are also some marginal models which are not implied by a mixed model. In contrast to the covariance patterns mentioned in section 7.1 random effects covariance structures do not require balanced data and the structure can be described with few parameters. Linear mixed effects models are particularly useful for the analysis of unbalanced longitudinal data. (Fitzmaurice *et al.* , 2004, p.198f.)

The assumed covariance structure could be checked by so-called *empirical semi-variograms*, but for the data at hand too few repeated measurements are available. Residual checks for the linear mixed model of cubic root titer do not display any worrying departures from the model assumptions.

7.3 Conclusion on models for Gaussian longitudinal data

In figure 7.1 estimated titers are compared to the actually observed titer for the participant with $ID = 766$. This man had two or more booster vaccinations and had had his last vaccination at the age of 36. Titers measured by method A are available 7, 9 and 10 years after the last vaccination. The increase of titer after the second measurement might be due to within-individual biological variation or measurement error. However, this participant was chosen because he represents the most ordinary case among the three persons with three measurements ($ID = 530, 763, 766$). Measurements for the woman with $ID = 530$ were taken within 4 months and the participant with $ID = 763$ had diabetes and had had his last vaccination at the age of 10.
Estimates for titer were obtained by applying models for Gaussian longitudinal data to the transformed variable cubic root of titer. The estimates for titer on the original scale were simply

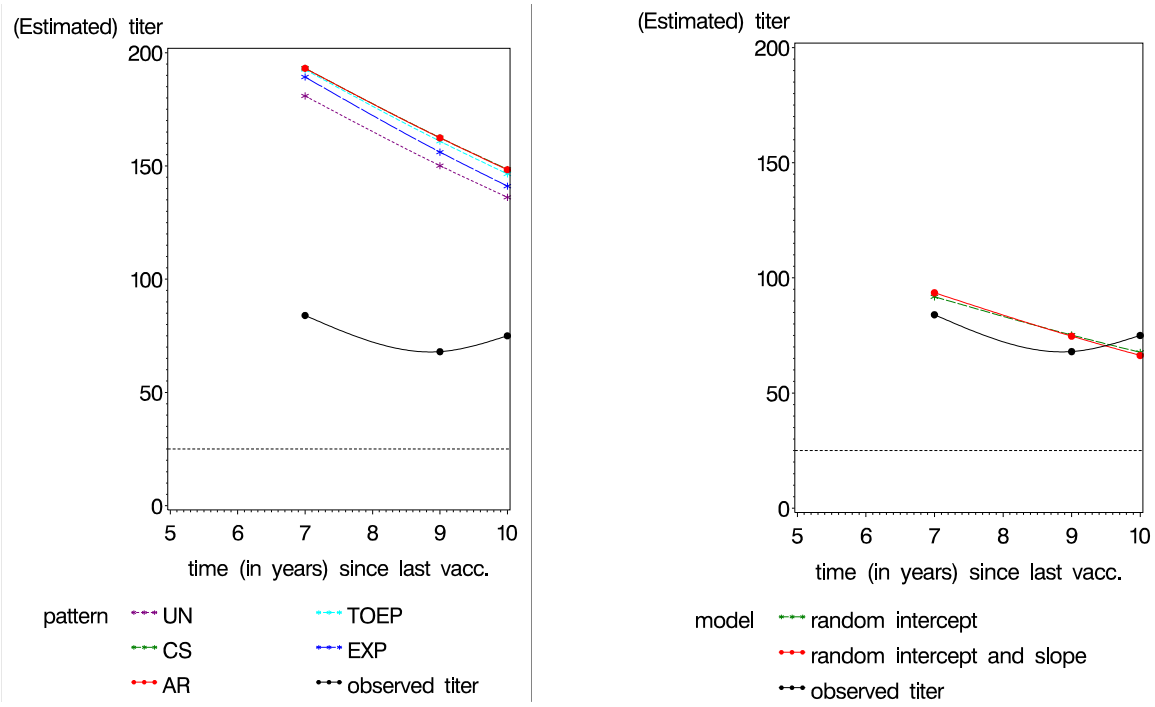

Figure 7.1: Comparison of estimated titers in marginal models (left) and subject-specific models (right) for cubic root of titer for the participant with $ID = 766$

obtained by taking the fitted values $\hat{y}_{ij} = \widehat{titer^{1/3}}_{ij}$ to the power of 3.

In marginal models estimation is done for the average of subgroups of a population. In the left panel of figure 7.1 estimates are quite far apart from the observed titer because the estimation is not done on the individual level but for an average person in the group of men with two or more booster vaccinations who had their last vaccination at the age of 36. By assuming the unstructured covariance pattern the estimated titers are lower than for other choices of covariance structure. The subject-specific estimated titers in the right panel of figure 7.1 are much closer to the observed titers because individual random effects were included.

In the case of normally distributed outcome variables, marginal models and subject-specific models have a close relationship. For other distributions there is no such close connection (Molenberghs & Verbeke, 2005, p.47). In the following chapter models are presented which extend on the one hand marginal models and on the other hand subject-specific models to distributions other than normal distribution.

Chapter 8

Models for longitudinal data in the non-Gaussian case

So far only models for longitudinal data on the cubic root transformation of titer have been applied. This transformation "normalizes" the strongly right-skewed outcome variable titer, but there is no illustrative interpretation of cubic root of titer and might be difficult to communicate. Application of GLMs to the 140 first measurements has shown that a gamma distribution might be a good choice for the distribution of titer measured by method A and because of that models are fitted by using *generalized estimating equations* (GEE) and *generalized linear mixed models*. Due to the elegant properties of multivariate normal distribution the connection between marginal and subject-specific model formulation is straightforward. Such close relations are not provided for other distributions. Depending on the research question either a marginal model formulation is used to model population-averaged means or a subject-specific model is formulated for inference on an individual level.
This chapter provides only a short outlook on models for non-Gaussian longitudinal data because no profound analysis is possible pertinent to the scarcity of repeated measurements. For further reading Diggle (2002) and Fitzmaurice *et al.* (2004) are recommended as well as the elaborated book on models for discrete longitudinal data by Molenberghs & Verbeke (2005).

8.1 Marginal models for longitudinal data in the non-Gaussian case

If researchers are primarily interested in population means, marginal model formulations are used by extending the ideas of GLMs, see chapter 5. The same notation for longitudinal data as in chapter 7 and same terminology as for GLMs in chapter 5 are used in the following.

8.1.1 The model

Expectations of the outcome variable Y measured on subject i are linked by a known link-function to a linear combination of covariate values annotated in the design matrix X_i:

$$g(\boldsymbol{\mu}_i) = X_i\boldsymbol{\beta},$$

where $\boldsymbol{\mu}_i = E(\boldsymbol{y}_i)$ denotes the n_i-dimensional vector of expected values for the outcome variable. The variance of each response Y_{ij} is depending on the mean by a scale parameter ϕ and the variance function $\nu(\cdot)$.

$$V(Y_{ij}) = \phi\nu(\mu_{ij})$$

In contrast to GLMs for independent observations, additionally a covariance structure $Cov(\boldsymbol{Y}_i) = V_i$ for repeated measurements on subject i has to be specified with variances $V(Y_{ij}) = \phi\nu(\mu_{ij})$ along the diagonal. An additional set $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots)$ of within-subject association parameters is needed and the covariance matrix will be denoted by $V_i(\boldsymbol{\alpha})$. This matrix can be decomposed by

$$V_i(\boldsymbol{\alpha}) = A_i^{\frac{1}{2}} R_i(\boldsymbol{\alpha}) A_i^{\frac{1}{2}}, \tag{8.1}$$

where A_i is a diagonal matrix with the variances $V(Y_{ij})$ on the main diagonal and $R_i(\boldsymbol{\alpha})$ is the correlation matrix for $\boldsymbol{Y}_i$ depending on the within-subject association parameters $\boldsymbol{\alpha}$. If the model for the marginal mean $g(\boldsymbol{\mu}_i) = X_i\boldsymbol{\beta}$ is correctly specified the parameter vector $\boldsymbol{\beta}$ can be estimated even if the correlation matrix $R_i(\boldsymbol{\alpha})$ is not correct. For a start only a so-called *working correlation* structure has to be assumed.
In order to fit a marginal model to the original outcome variable *titer*, gamma distribution is assumed and log-link-function and the same predictor variables are used as for the fitted GLM in chapter 5.

$$\log(E(titer_{ij})) = \beta_0 + \beta_1 sex_i + \beta_2 boosters1_i + \beta_3 boosters2_i + \beta_4 age_i + \beta_5 lastVacc_{ij} \tag{8.2}$$

For the marginal model in the Gaussian case the AR(1) covariance pattern model seemed to be satisfying. On that account an AR(1) working correlation structure is chosen for $R_i(\boldsymbol{\alpha})$.

8.1.2 Parameter estimation

For analyzing correlated data that can be discrete or continuous an alternative method to ML-estimation is provided by generalized estimating equations (GEE). This method is based on extension of the usual score equations for GLMs (5.11) by including the covariance matrix $V_i(\boldsymbol{\alpha})$.

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{N} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}'} V_i^{-1}(\boldsymbol{\alpha})(\boldsymbol{y}_i - \boldsymbol{\mu}_i) = \sum_{i=1}^{N} D_i' V_i^{-1}(\boldsymbol{\alpha})(\boldsymbol{y}_i - \boldsymbol{\mu}_i) = \mathbf{0}, \tag{8.3}$$

where D_i simply denotes the matrix of derivatives $D_i = \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}$. The equations in (8.3) are called *generalized estimating equations* or *quasi-score equations* (Dobson, 2002, p.203). The correlation parameters $\boldsymbol{\alpha}$ are usually estimated by standardized residuals

$$r_{ij} = \frac{y_{ij} - \mu_{ij}}{\sqrt{v(\mu_{ij})}}. \tag{8.4}$$

These residuals depend on the expectation $\mu_{ij} = \boldsymbol{x}_{ij}'\boldsymbol{\beta}$ and conclusively on the parameter vector $\boldsymbol{\beta}$. On the other hand the solution for $\boldsymbol{\beta}$ in equation (8.3) depends on the correlation parameters $\boldsymbol{\alpha}$. Because of this interdependence between the parameters an iterative procedure is used to estimate alternately $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$:

1. Choice of an initial correlation matrix R_i. Usually the identity matrix is used, assuming

independence between all measurements and a univariate GLM is fit to the data.

2. Calculation of the initial parameter estimate $\boldsymbol{\beta}^{(0)}$ by solving the score equations (8.3).
3. Computation of the estimated mean $\hat{\boldsymbol{\mu}}_i = g^{-1}(X_i\hat{\boldsymbol{\beta}})$ and the standardized residuals
4. Residuals are used to estimate the parameters of A_i and $R_i(\boldsymbol{\alpha})$
5. Computation of $R_i(\boldsymbol{\alpha})$ and $V_i(\boldsymbol{\alpha})$
6. The current estimate for $\boldsymbol{\beta}$ is updated by

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \left[\sum_{i=1}^{N} D_i' V_i^{-1} D_i\right]^{-1} \times \left[\sum_{i=1}^{N} D_i' V_i^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i)\right] \tag{8.5}$$

The last four steps are iterated until convergence is reached. This standard GEE method is the most common way to estimate parameters for correlated non-Gaussian data. An advantage of GEE is that the joint distribution of the outcomes needs not to be fully specified (Molenberghs & Verbeke, 2005, p.158). Even if the working correlation matrix is not correct an appropriate variance estimator for $\hat{\boldsymbol{\beta}}$ can be obtained by using the robust *sandwich-estimator* or *empircal corrected variance*.
Table 8.1 displays GEE parameter estimates and empirical standard error estimates for the marginal model on *titer* specified in (8.2). The parameter estimates in table 8.1 are quite similar to the parameter estimates for the GLM (5.12) in chapter 5.

Table 8.1: Estimated parameters obtained by GEE for marginal model of titer (8.2) with assumed gamma distribution, log-link-function and AR(1)-working correlation structure

number of subjects=140, number of observations=162

		Standard	95% Conf.			
Parameter	Estimate	Error	Limits		Z	$Pr > \|Z\|$
Intercept	6.178	0.263	5.663	6.693	23.51	<.0001
sex 0	0	0	0	0	.	.
sex 1	-0.352	0.164	-0.673	-0.031	-2.15	0.0314
boosters 0	0	0	0	0	.	.
boosters 1	0.021	0.229	-0.429	0.470	0.09	0.9288
boosters 2	0.320	0.220	-0.110	0.751	1.46	0.1445
age	-0.014	0.005	-0.025	-0.003	-2.54	0.0112
lastVacc	-0.071	0.026	-0.123	-0.020	-2.71	0.0067
Scale	0.846	.	.	.	.	.
	NOTE: The scale parameter for GEE estimation was computed as the square root of the normalized Pearson's χ^2					

In the next section random effects are included into the mean model of GLMs to explicitly account for between-subject heterogeneity.

8.2 Generalized linear mixed models for longitudinal data

Generalized linear mixed models are an extension of linear mixed models to the non-Gaussian case by including random effects in the GLM framework. Especially for discrete repeated outcomes

this kind of modelling becomes more and more common for studies in social and health sciences.

8.2.1 The model

Corresponding to the linear mixed model applied to the TBE data in section 7.2, a subject-specific model is now fitted to the outcome variable *titer* by including a random intercept in the mean model of the GLM (5.9). The expected value of titer given the random intercepts is modelled as

$$\begin{aligned} E(titer_{ij}|b_{0i}) &= \exp\left(\beta_0 + b_{0i} + \beta_1 sex_i + \beta_2 boosters1_i + \beta_3 boosters2_i \right. \\ &\quad \left. + \beta_4 age_i + \beta_5 lastVacc_{ij}\right). \end{aligned} \tag{8.6}$$

The random intercepts are supposed to be normally distributed with mean zero and variance τ^2. In contrast to GLMs the distribution assumption is not made on the outcome variable *titer* itself but on the conditional outcome $titer_{ij}|b_{0i}$. It is supposed that titer given the random effects is gamma distributed.

$$titer_{ij}|b_{0i} \sim \boldsymbol{\Gamma}(shape, scale) \text{ and } b_{i0} \sim \boldsymbol{N}(0, \tau^2)$$

The covariance matrix of the vector of titers for subject i conditional on the random intercept is

$$CoV(\boldsymbol{titer}_i|b_{0i}) = A_i^{\frac{1}{2}} R_i A_i^{\frac{1}{2}}.$$

A_i is a diagonal matrix and contains the variance function of the specified distribution. R_i is assumed to be ϕI_{n_i}. Hence, the conditional variance for one observation is $V(titer_{ij}|b_{0i}) = \mu_{ij}^2 \phi$.

8.2.2 Estimation of parameters

Parameters in generalized linear mixed models are based on maximizing the likelihood for the fixed parameters $\boldsymbol{\beta}$, the covariance matrix D for the random effects and the scale parameter ϕ:

$$L(\boldsymbol{\beta}, D, \phi) = \prod_{i=1}^{N} f_i(\boldsymbol{y}_i|\boldsymbol{\beta}, D, \phi) = \prod_{i=1}^{N} \left[\int \left(\prod_{j=1}^{n_i} f_{ij}(y_{ij}|\boldsymbol{b}_i, \boldsymbol{\beta}, \phi) q(\boldsymbol{b}_i|D) \right) \mathrm{d}\boldsymbol{b}_i \right]$$

The problem in maximizing this likelihood are the N integrals over the q-dimensional vector $\boldsymbol{b}_i$. Only in some special cases, like for the normal assumption, an analytical solution exists. According to Molenberghs & Verbeke (2005, p.268ff.) there are three possibilities for numerical approximations of the likelihood: approximation of the integrand, the data or the integral. Detailed explanations of these numerical methods go beyond the scope of this thesis. The interested reader is recommended to study Molenberghs & Verbeke (2005). For the following it is assumed that procedures exist which provide estimates for the parameters based on ML-principles. Usually estimation of the parameters $\boldsymbol{\beta}, D, \phi$ is of primary interest. In cases where the prediction of subject-specific evolution is required, estimates for the random effects are needed. Like in linear mixed models those are obtained by Bayes methods based on their posterior distribution.

Table 8.2: Estimated parameters for the random intercept model of titer (8.6) with assumed gamma distribution and log-link-function

number of subjects=140, number of observations=162

Parameter	Estimate	Standard Error	DF	t-value	$Pr > \|t\|$
Intercept	5.885	0.305	135	19.28	<.0001
sex 0	0	.	.	.	.
sex 1	-0.399	0.189	21	-2.11	0.0470
boosters 0	0	.	.	.	.
boosters 1	0.133	0.300	21	0.44	0.6622
boosters 2	0.533	0.256	21	2.08	0.0501
age	-0.017	0.006	21	-2.89	0.0088
lastVacc	-0.098	0.028	21	-3.52	0.0020
τ^2	1.046	0.143			
ϕ	0.103	0.032			

In most statistical software packages tools for fitting generalized linear mixed models are not standard. For computing estimates for the specified generalized linear mixed model (8.6) the new SAS procedure `PROC GLIMMIX` is used. This procedure needs to be downloaded from SAS software site `http://www.sas.com/apps/demosdownloads/setupcat.jsp?cat=SAS\%2FSTAT+Software`. Very detailed explanations on estimation methods and model theory can be found in the SAS documentation (SAS Institute Inc., 2006) and an introduction for the usage of this procedure is given by Schabenberger (2005). In SAS so-called *residual pseudo likelihood estimation* is used by `PROC GLIMMIX` and computed estimates for fixed parameters and variance components are displayed in table 8.2. These estimates are quite different from those obtained from GEE for the marginal model, see table 8.1. Once more it should be emphasized that there is no close relation between marginal models and subject-specific models in non-Gaussian case.
Plots of Pearson residuals were checked for departures from the model assumptions. The plot of the residuals against linear predictor displayed a slightly linear trend. Histogram, normal quantile plot and box plot of Pearson residuals did not show any departures from the model assumptions.

8.3 Conclusion on models for non-Gaussian longitudinal data

In this chapter models for non-Gaussian longitudinal data were briefly presented. Depending on the research question the appropriate type of modelling - marginal or subject-specific - has to be chosen. For the analysis of repeatedly measured titers, random effects models are more appropriate, because they explicitly take different sources of correlation in longitudinal data into consideration and there are no problems with unbalanced data.

The next chapter completes this thesis with comparison of estimates for all investigated models and forecasts of titers.

Chapter 9

Model comparison, prediction and review

In this last chapter all models are compared and titers are predicted. A short review on models, results and predictions closes this thesis.

9.1 Model comparison

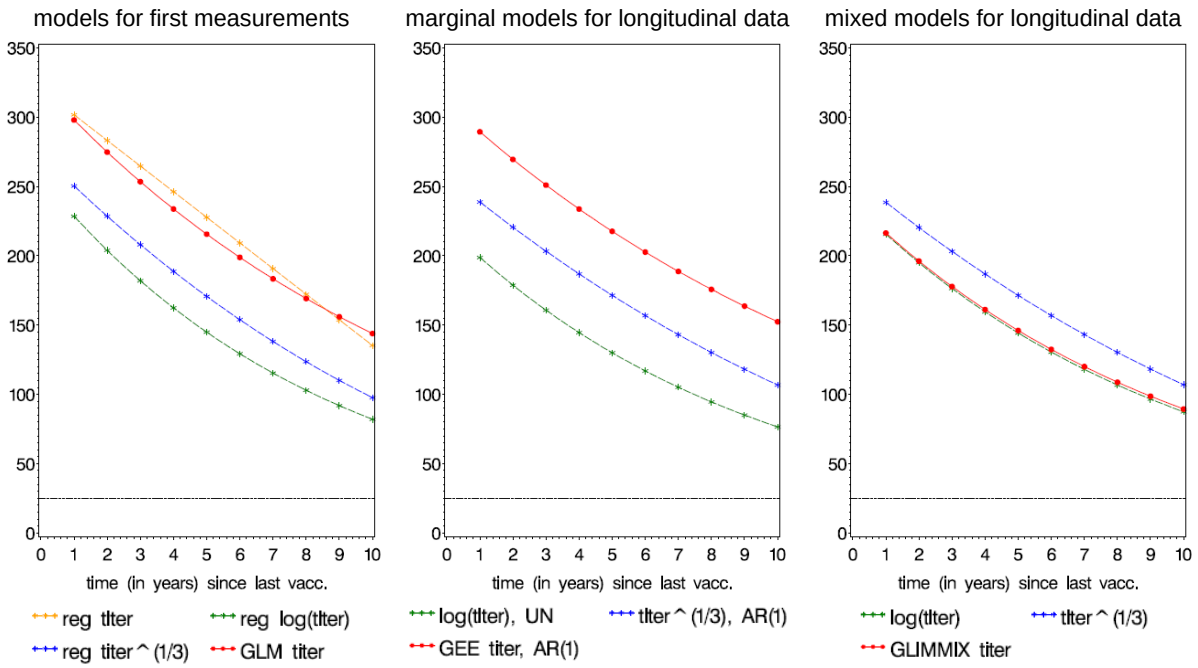

Figure 9.1: Comparison of estimated titers in different models by means of a fictitious man with two or more booster vaccinations who had his last vaccination at the age of 55

In figure 9.1 predictions are compared for models for the independent first measurements and for marginal and mixed models for longitudinal data. The predicted titers are illustrated by means of a fictitious man with two or more booster vaccinations who had his last vaccination at the age of 55. All models were fit by using the same predictors sex, number of booster vaccinations in categories 0, 1 or ≥ 2, age at last booster vaccination and time since last vaccination.

For the first observations measured by method A, classical regression analysis was performed on the original outcome variable *titer*. Because *titer* did not fulfill the required normality assumption, transformations $titer^{1/3}$ and $\log titer$ were used instead. The cubic root transformed variable coped best with the underlying model assumptions in linear regression but is hard to interpret. Therefore the more common logarithmic transformation was also applied. The distribution of the residuals in linear regression analysis of $\log(titer)$ showed some left-skewness, but other model assumptions were adequate. Predictions for $titer^{1/3}$ and $\log(titer)$ were back-transformed to the original scale of *titer* and following estimates are compared for linear regression models:

- $\hat{E}(titer)$
- $\exp\left(\hat{E}(\log titer)\right)$
- $\left(\hat{E}(titer^{1/3})\right)^3$

Because the original outcome variable *titer* seemed to fit a gamma distribution quite well, a generalized linear model was applied using the log-link-function. The difference between performing a regression analysis on a logarithmic transformed variable $E(\log y_i) = \boldsymbol{x}_i'\boldsymbol{\beta}$ and using a logarithmic link-function in a GLM $\log\left(E(y_i)\right) = \boldsymbol{x}_i'\boldsymbol{\beta}$ is emphasized. The resulting estimates of the specified GLM

- $\hat{E}(titer)$

are compared to the predictions of linear regression models. The estimated titers obtained from regression analysis of $titer^{1/3}$ are higher than those obtained from analysis of $\log(titer)$. The predicted values of the GLM are higher than the predictions of regression models on the transformed variables. The curvature is similar to the curvature of predicted values for the model on $\log(titer)$.
Only around 13% of variability in the first measured titers is explained by the covariates sex, age, number of booster vaccinations and time since last vaccination by fitting linear regression models. For accounting on this unobserved heterogeneity, models for longitudinal data are applied including the few available repeatedly measured titers. There are two types of models for longitudinal data: marginal or so-called population-averaged models and subject-specific models.

In marginal models the correlations between repeated measurements are considered by specifying a covariance structure. As long as the mean is correctly specified valid point estimates for the regression parameters are obtained even if these covariance pattern is not correct. For computing valid standard errors robust estimation techniques are used.
Marginal models were applied to the transformed variables $titer^{1/3}$ and $\log(titer)$ supposing a normal distribution. For $titer^{1/3}$ the assumption of a first-order autoregressive and for $\log(titer)$ the assumption of an unstructured covariance pattern seemed most satisfying. Predicted values obtained from these models were back-transformed as explained above. A marginal model under the assumption of a gamma distribution for the variable *titer* and using the log-link-function was estimated by applying the method of generalized estimating equations . The predicted values for marginal models have almost the same range as estimates for models on the first measurements, see figure 9.1. The curves are a bit flatter than in the first panel.

In vaccination studies large between-individual heterogeneity is expected because of highly varying response of humans to vaccinations. Subject-specific models take between-individual heterogeneity into account by including individual random effects.
Linear mixed models are widely used in health sciences for modelling normally distributed outcome variables and were applied to the transformed variables $titer^{1/3}$ and $\log(titer)$.
Generalized linear mixed models are applied to non-Gaussian distributed response variables by including random effects in GLMs. A generalized linear mixed model was applied to the untransformed outcome *titer*, using log-link-function and assuming a gamma distribution. In the case of TBE data only a randomly varying intercept is included in the mixed models because estimation of a random slope for these few repeated measurements would be overstated. The back-transformed predictions of linear mixed models for $titer^{1/3}$ and $\log(titer)$ are almost the same as for models on the first measurements. The curve of estimated titers of the generalized linear mixed model is much lower than for predictions of GLM in the first panel and predictions of GEE in the second panel.

The generalized linear mixed model is now used to forecast titers for different fictitious persons.

9.2 Prediction of titer

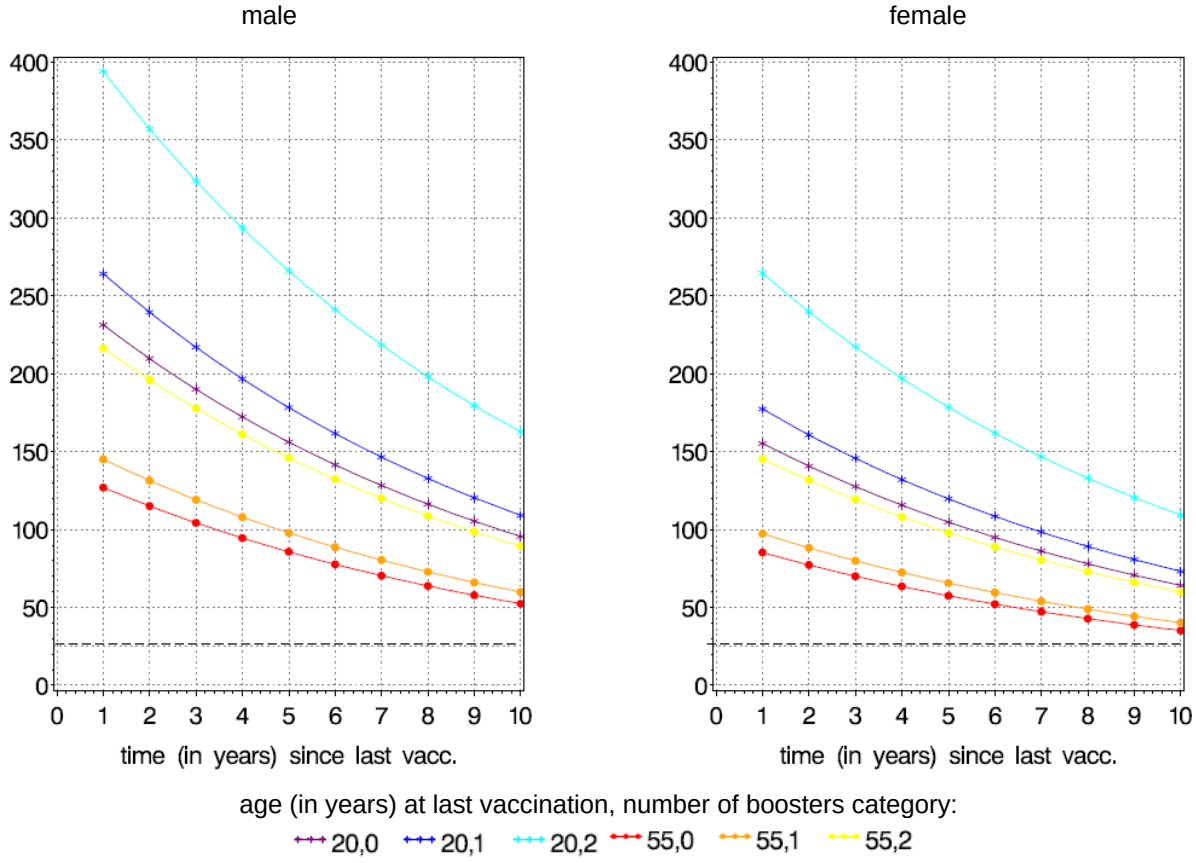

Figure 9.2: Prediction of titer

In figure 9.2 forecasts of titers for several fictitious persons are displayed from 1 to 10 years after

their last vaccination. Titers are predicted for men (left panel) and women (right panel) who had their last vaccination at the age of 20 or 55 and had 0, 1 or ≥ 2 booster vaccinations before. Estimated titers for women are lower than for men. Titer decreases with increasing age and the more booster vaccinations the higher the estimated titer. For instance the titer of a 56-year-old woman who had her last vaccination 1 year ago is estimated to be as high as the titer of a 27-year-old woman who had her last vaccination 7 years ago. In all cases the predicted titers are seropositive ($> 25\ U/ml$).

9.3 Review and outlook

The goal of this thesis was to investigate the TBE antibody decrease after completion of the basic immunization protocol. Data were presented on which the further analyzes would be based after knowledge on TBE, vaccination and methods for measuring antibody concentration had been acquired. Because lots of data records were not complete and different test methods for assessing antibody concentrations were used, only a small amount of data was available for further statistical analyzes. On these data several statistical models were performed like linear regression, GLM, linear mixed models and generalized linear mixed models. Titers were estimated for all these models and showed that women have lower titer than men and the more booster vaccinations the higher the antibody concentrations. Titer is decreasing with increasing age and time since last vaccination.

In Austria it is currently recommended to do a booster vaccination every 5 years, but in all applied models estimated titers were seropositive longer than five years after the last vaccination. To prove this hypothesis a planned clinical trial would be necessary where at least 3 measurements per person would be available. Because immune response is very heterogeneous across human beings, subject-specific models including a random intercept and random slope would be appropriate for modelling titer. To estimate a random slope at least 3 measurements per person would be necessary. If even 5 measurements per person would be available it could be controlled if the slope remains constant over time. These successive measurements should be taken on men and women of different age with different number of precedent booster vaccinations. On this data set it would be possible to make solid estimates on the duration of protection provided by TBE vaccination.

Bibliography

ARGE Gesundheitsvorsorge. Download: 2006-11-13. *FSME. Verbreitungsgebiete. Verbreitungsgebiete Österreich.* Internet: `http://www.zecken.at/Zecken.aspx_param_target_is_49690.v.aspx`.

ARGE Gesundheitsvorsorge. Download: 2007-03-31a. *FSME. Statistiken.* Internet: `http://www.zecken.at/Zecken.aspx_param_target_is_49690.v.aspx`.

ARGE Gesundheitsvorsorge. Download: 2007-03-31b. *Schutzimpfung. Impfplan.* Internet: `http://www.zecken.at/Zecken.aspx_param_target_is_49690.v.aspx`.

Barrett, P.N. et al. 2004. *Vaccines.* 4 edn. Elsevier Inc (USA). Chap. Tick-borne encephalitis virus vaccine.

Binder, Lothar. 1996. Frühsommer-Meningoenzephalitis (*FSME*) in Österreich. Übersicht über *E*rkrankung und Impfung mit Fallbericht. *Pages p. 50–65 of:* Mittermayer, H., & Pietsch, M. (eds), *Impfmanagement.*

Brown, Helen, & Prescott, Robin. 1999. *Applied Mixed Models in Medicine.* John Wiley & Sons Ltd.

Charrel, R.N. et al. 2004. Tick-borne virus diseases of human interest in Europe. *Clinical Microbiology and Infection*, **10**(12), 1040–1055.

Crowder, M.J., & Hand, D.J. 1990. *Analysis of Repeated Measures.* Monographs on Statistics and Applied Probability, vol. 41. Chapman & Hall.

Diggle, Peter J. et al. 2002. *Analysis of Longitudinal Data.* Second edition edn. Oxford University Press.

Dobson, Annette J. 2002. *An Introduction to Generalized Linear Models. Second Edition.* Chapman & Hall/CRC texts in statistical science series.

Dumpis, Uga, Crook, Derrick, & Oksi, Jarmo. 1999. Tick-Borne encephalitis. *Clinical Infectious Diseases*, April, 882–890.

Fitzmaurice, Garrett M., Laird, Nan M., & Ware, James H. 2004. *Applied Longitudinal Analysis.* John Wiley & Sons, Inc.

Gesellschaft für Virologie e.V. Download: 2007-03-18. *Steckbrief Frühsommer-Meningoenzephalitis (FSME) Impfstoff.* Internet: `http://www.g-f-v.org/inhalt_de.php?lmnop=1&modul=TEXTE&aktion=DETAILS&id=197`.

Hafner, Robert. 1989. *Wahrscheinlichkeitsrechnung und Statistik*. Springer Wien, New York.

Haglund, M. et al. 1996. A 10-year follow-up study of tick-borne encephalitis in the Stockholm area and a review of the literature: need for a vaccination strategy. *Scandinavian journal of infectious diseases*, **28**(3), 217–224.

Hainz, U. et al. 2002. Vaccine protection in the elderly: are Austrian seniors adequately protected by vaccinations? *Wiener Klinische Wochenschrift*, **114**(5-6), 167–170.

Heinz, F.X. 2006. Etiology. *Chap. 2, pages 7–16 of: Management of Tick-borne Encephalitis. Compendium of scientific literature.* Institute for Social Medicine of the University of Vienna, Rooseveltplatz 3, A-1090 Vienna, Internet: `http://www.isw-tbe.info/tbe.aspx_param_target_is_51410_and_l_is_2.v.aspx`: Prof. Ursula Kunze.

Holzmann, Heidemarie. 2003. Diagnosis of tick-borne encephalitis. *Vaccine*, **21**, 36–40.

Jezyna, C. et al. 1984. Epidemiologic and clinical studies of patients with tick-borne encephalitis from northeastern Poland. *Zentralblatt für Bakteriologie, Mikrobiologie und Hygiene [B]*, **178**(5-6), 510–521.

Kaiser, Reinhard. 1999. The clinical and epidemiological profile of tick-borne encephalitis in southern Germany 1994Ű98. A prospective study of 656 patients. *Brain. A journal of neurology*, **122**(11), 2067–2078.

Kaiser, Reinhard. 2002. Tick-borne encephalitis (TBE) in Germany and clinical course of the disease. *International journal of medical microbiology*, **291**(33).

Kaiser, Reinhard. 2006. Clinical Description. *Chap. 4, pages 33–41 of: Management of Tick-borne Encephalitis. Compendium of scientific literature.* Institute for Social Medicine of the University of Vienna, Rooseveltplatz 3, A-1090 Vienna, Internet: `http://www.isw-tbe.info/tbe.aspx_param_target_is_51410_and_l_is_2.v.aspx`: Prof. Ursula Kunze.

Köck, T. et al. 1992. Zur Klinik der Frühsommermeningoenzephalitis (FSME) in der Steiermark (On the clinical course of tick-borne encephalitis (TBE) in Styria). *Nervenarzt*, **63**(4), 205–208.

Kind, Albert. 2004. Wie viele Auffrischungsimpfungen sind notwendig gegen die Zeckenenzephalitis FSME (Frühsommermeningoenezephalitis)? *Schweizerische Ärztezeitung*, **85**(16), 844–848.

Kunz, Christian. 1992. Tick-borne encephalitis in Europe. *Acta Leiden*, **60**(2), 1–14.

Kunz, Christian. 2003. TBE vaccination and the Austrian experience. *Vaccine*, **21**, 50–55.

Labuda, M. et al. 1993. Tick-borne encephalitis virus activity in Styria, Austria. *Acta virologica*, **37**(2-3), 187–190.

Löser, C., Mehlhorn, H., & Schill, W.-B. 2002. Die Zecke sticht! Betrachtungen über ein Misnomer. *Zeitschrift Der Hausarzt*, **53**.

McCullagh, Peter, & Nelder, John A. 1989. *Generalized Linear Models. Second Edition.* Chapmann & Hall.

Mickiene, A. 2006. Diagnosis. *Chap. 6, pages 43–45 of: Management of Tick-borne Encephalitis. Compendium of scientific literature.* Institute for Social Medicine of the University of Vienna, Rooseveltplatz 3, A-1090 Vienna, Internet: `http://www.isw-tbe.info/tbe.aspx_param_target_is_51410_and_l_is_2.v.aspx`: Prof. Ursula Kunze.

Mickiene, A. et al. 2002. Tick-borne encephalitis in a area of high endemicity in Lithuania: disease severity and long-term prognosis. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, **35**(6), 650–658.

Molenberghs, Geert, & Verbeke, Geert. 2005. *Models for Discrete Longitudinal Data.* Springer New York.

Montgomery, Douglas C. 2001. *Design and Analysis of Experiments, 5th edition.* John Wiley & Sons.

Olsson, Ulf. 2002. *Generalized Linear Models. An Applied Approach.* Studentlitteratur, Lund.

Patterson, H. D., & Thompson, R. 1971. Recovery of inter-block information when block sizes are unequal. *Biometrika*, **58**.

P.N., Barrett, S., Schober-Bendixen, & Ehrlich, H.J. 2003. History of TBE vaccines. *Vaccine*, **21**(2), 41–49.

Rendi-Wagner, Pamela. 2004a. Risk and Prevention of Tick-borne Encephalitis in Travelers. *Journal of Travel Medicine*, **11**(5), 307–312.

Rendi-Wagner, Pamela et al. 2004b. Persistence of protective immunity following vaccination against tick-borne-encephalitis - longer than expected? *Vaccine*, **22**, 2743–2749.

SAS Institute Inc. 2003. *SAS OnlineDoc ® 9.1.* Internet: `http://support.sas.com/91doc/docMainpage.jsp`.

SAS Institute Inc. 2006 (June). *The GLIMMIX Procedure.* Internet: `http://www2.sas.com/proceedings/sugi30/196-30.pdf`.

Schabenberger, Oliver. 2005. Introducing the GLIMMIX Procedure for Generalized Linear Mixed Models. *In: Proceedings of the Thirtieth Annual SAS ®Users Group International Conference.* SAS Institute Inc., Cary, NC, USA, SAS Institute Inc., SAS Campus Drive, Cary, NC 27513 Email: oliver.schabenberger@sas.com.

Schneider, H. 1931. Über epidemische akute 'Meningitis serosa'. *Klinische Wochenschrift Wien*, **44**, 350–352.

Süss, J. 2006. Epidemiology. *Chap. 3, pages 17–32 of: Management of Tick-borne Encephalitis. Compendium of scientific literature.* Institute for Social Medicine of the University of Vienna, Rooseveltplatz 3, A-1090 Vienna, Internet: `http://www.isw-tbe.info/tbe.aspx_param_target_is_51410_and_l_is_2.v.aspx`: Prof. Ursula Kunze.

Stanek, Gerold, & Hofmann, Hanns. 1994. *Krank durch Zecken.* Maudrich, Wien.

Verbeke, Geert, & Molenberghs, Geert. 2000. *Linear Mixed Models for Longitudinal Data.* Springer New York.

Vittinghoff, Eric, Glidden, David V., Shiboski, Stephen C., & McCulloch, Charles E. 2005. *Regression Methods in Biostatistics.Linear, Logistic, Survival, and Repeated Measures Models.* Springer Science+Business Media.

Zeger, Scott L., Liang, Kung-Yee, & Albert, Paul S. 1988. Models for Longitudinal Data: A Generalized Estimating Equation Approach. *Biometrics*, **44**.

Zoehrer, B., Spork, D., & Zenz, W. 2003. Frühsommermeningoenzephalitis. Klinik und Therapie. *Monatsschrift Kinderheilkunde*, 1163–1169.