Technische Universität München

Zentrum Mathematik

Estimating LTC Premiums using GEEs for Pseudo-Values

Diplomarbeit

von

Florian Helms

Themensteller:	Prof. Dr. Claudia Czado
Betreuer:	Prof. Dr. Claudia Czado
	Dipl.-Math. oec. Susanne Gschlößl
Abgabetermin:	18. Dezember 2003

Hiermit erkläre ich, dass ich die Diplomarbeit selbständig angefertigt und nur die angegebenen Quellen verwendet habe.

München, 18. Dezember 2003

Acknowledgement

For the work completed in this diploma thesis I am indebted to quite a number of different people, that I want to thank at this point:

Firstly, I would like to thank my supervisor Prof. Dr. Claudia Czado for the great ongoing encouragement, as well as the helpful discussions and sound advice, that made this diploma thesis possible.

Secondly, I am grateful for the great support and review of all my work by my advisor Susanne Gschlößl. Special thanks for helping me with the dataset, the programs and the software involved, and for the rapid replies on all of my questions, which allowed me to pursue my goals in this thesis effectively.

Thirdly, sincere thanks to my student colleagues, especially Katharina Meyer, for reviewing the thesis, and my work colleagues that supported and encouraged me from the beginning on despite their own high workload.

I would also like to thank my family, friends and housemates for their technical and emotional support during the time, when long hours were required to be spent on this thesis.

Finally, I would like to thank the mathematical community, working in this field, that did not hesitate to answer questions on published work. I just want to mention Professor Per Kragh Andersen and Professor Richard Gill on behalf of many.

florian helms

Contents

Chapter 1

Introduction

One issue, that has been discussed for years and years and in fact is still discussed, is the effect of increasing life expectancy on society. Not only newborn children, also aged people, have today, compared to the past, a longer life expectancy. For example a 60-year old male can still expect to live for another 19.5 years, a female of the same age even for another 23.7 years according to the Federal Statistical Office Germany. Some people argue that increasing life expectancy is due to decreasing morbidity at all ages (compressed morbidity hypothesis); others claim that gained years of life expectancy are spent entirely in illness (expanded morbidity hypothesis). It seems as often right and equitable to assume that the reality is somewhere in between.

But not only the life expectancy or the so often quoted demographical development changes the life of older people. There are various reasons: Due to a more individualized society and a different family structure people live often alone and require especially at older ages external assistance to manage their daily life. The question arises, how should the individual, how should the society react on these issues?

More and more industrialized countries added the so-called long term care insurance (LTCI) to their social insurance system, to guarantee a basic cover in situations where external assistance and care is needed to cope with daily life. Insurance companies also have developed full or additional insurance cover for long term care (LTC). It might be necessary to note at this point that LTC is not only a problem of aging; also in the childhood, mainly due to impairments at birth, children can become or even be LTC patients and stay that for a lifetime. These children still have, if accurate medical care is given to them, a life expectancy of 50 years and more, thanks to medical progress. Young- and middle-aged people also can become in need of LTC, mainly because of accidents.

In Germany today around 4% of the people aged between 60 and 80 years and around 30% of the people older than 80 years require some kind of external assistance. To relief them from the financial burden involved a compulsory LTCI was introduced in 1995 as the so-called fitfth column of the welfare system beside pension, health, accident and unemployment insurance. The leading idea was that LTCI follows health insurance. Every citizen in the social health insurance system was given LTCI cover for 1.7% of its income up to a certain ceiling. Citizen in the private insurance system, had to be granted LTCI cover by their private health insurer without any medical examination. Consequently the private insurance companies had to take over a large claim portfolio right at the start. Today, out of approximately 80 million people in Germany, around 1.84 millions in the social security system and another 107,000 in the private insurance system (BMGS (2003)) receive benefits from this compulsory LTCI.

We are going to analyze a representative sample of claim-records in the private sector within the compulsory LTCI in this diploma thesis.

The life-history of an individual can be interpreted as a multi-state model. During its lifetime an individual spends varying times in different states, that for example can correspond to its health status. Transitions to other states occur over time. Objective of our analysis is to determine the probabilities of such transitions for any individual. Usually Cox's proportional hazard-model can be used to asses the influence of age, gender, place and severeness of care on the individual's health status and thus on these transition probabilities. In Cox's model the hazard-rate function $\lambda(t)$ is assumed to be the product of a baseline hazard $\lambda_0(t)$ and $exp\{\mathbf{Z}^T\boldsymbol{\beta}\}$, where $\mathbf{Z}$ is a vector of covariates and $\boldsymbol{\beta}$ the parameter vector to be estimated. A regression analysis produces an estimate for $\boldsymbol{\beta}$, and the hazard-rate, also known as the transition intensity, can be calculated. For further details see Czado and Rudolph (2002).

However, for actuarial purposes less the transition intensities, but more the transition probabilities are the quantities of interest. In the case of Cox's proportional hazard-model the transition probabilities are calculated from the transition intensities using a relationship given by a set of differential equations, if the intensities sum up to zero, that is $\sum_{h\in\mathcal{S}} \lambda_{gh} = 0 \; \forall g \in \mathcal{S}$. Thus the transition probabilities are complex non-linear functions of the intensity regression coefficients.

Against this background Andersen, Klein, and Rosthøj (2003) developed a method that models the transition probabilities directly. This method calculates pseudo-values based on the Aalen-Johansen estimator, an almost unbiased estimator of the transition matrix of an Markovian multi-state model. These pseudo-values are then used in Generalized Estimating Equations (GEEs), that take, in contrast to the Generalized Linear Models (GLMs), correlation between observations into account, to estimate the parameters of the model.

We want to apply this method to a set of data, containing the claim-records of LTC claimants, and derive the necessary transition probabilities in order to calculate the insurance premiums required for a given LTC-plan.

In the next chapter we introduce the basic quantities of survival analysis such as the survival function and the hazard-rate function, as well as the special features of survival data, namely censoring and truncation, to emphasize the problems that need to be considered in the statistical analysis of survival data. In the chapter thereafter, the discrete and continuous-time Markovian multi-state model is explained, since we are going to use such a model for our analysis. The quantities of interest, the transition probabilities and transition intensities, are defined and their relationship given by the Kolmogorov differential equations is derived.

Beside other non-parametric estimators used in survival analysis such as the Nelson-Aalen estimator for the cumulative hazard-rate and the Kaplan-Meier estimator for the survival distribution function, we present the Aalen-Johansen estimator, an almost unbiased estimator for the transition matrix in a Markovian multi-state model, in the fourth chapter. We will see that the Aalen-Johansen estimator reduces in a two-state model to the Kaplan-Meier estimator and we will give an algorithm to compute the Aalen-Johansen estimator in a multi-state model. In the chapter that follows, the concept of pseudo-values form jackknife methodology is introduced. Since the Aalen-Johansen estimator does not depend on covariates, pseudo-values based on the Aalen-Johansen estimator are calculated, that link the i^{th} pseudo-value with the covariates of the i^{th} observation and thus generate the data required for a regression analysis.

We calculate pseudo-values at different points in time using the same observations, therefore pseudo-values are not independent as required for GLMs. Thus we introduce GEEs in the sixth chapter, that in contrast to GLMs take correlation into account. Beside the Maximum Likelihood Estimation (MLE) for GLMs, the Maximum Quasi-Likelihood Estimation (MQLE) is presented in this chapter, as well, since the GEEs can be seen as an extension of quasi-likelihood to longitudinal data analysis. The solution of the GEEs is denoted by $\hat{\boldsymbol{\beta}}_G$ and we will show that the GEEs produce consistent and asymptotically Gaussian distributed estimates for the true value of $\boldsymbol{\beta}$ under mild conditions, even when the correlation structure is misspecified.

In the seventh chapter we apply the tools introduced to the data set containing the claim-history of LTC patients mentioned above: We calculate the Aalen-Johansen estimator of a three-state model, derive the pseudo-values and thus generate the data for a regression analysis using GEEs, where we specify the logit as link function. With the estimates obtained from this regression analysis we calculate finally the one-year transition probabilities of our model. These transition probabilities are used in the chapter that follows to calculate the actuarial values for a given LTC-plan and thus derive the necessary premiums. A summary with a comparison of our premiums with premiums offered by a German health insurer finalizes the analysis.

Chapter 2

Survival Analysis

In this chapter we are going to introduce the basic concepts of survival analysis based on survival data. Survival data concern the time of different lives to death. We are interested in deriving the probability that an individual survives beyond a certain age. Therefore a sample of individuals with their corresponding lifetimes is needed in order to calculate these probabilities, e.g. the survival function or other related quantities like the so-called hazard-rate function. The hazard-rate function gives the probability of death over a small interval and is related with the survival function in a certain way, that will be explained in this chapter.

For practical reasons individuals can only be observed for a certain period of time and not over their whole lifetime. "Censoring" and "Truncation" are the characteristic features of survival surveys accounting for this. Another characteristic feature is "conditioning" on survival up to a certain time. Given an z-year old individual, its survival to age $z + t$ is conditional on having survived to time z. Because of these special features modified methods have to be used, compared to standard statistical practice, to take this into account. For example, constructing the Maximum Likelihood for survival data, one has to correct for the missing information due to censoring or truncation.

2.1 The Survival Function

We are interested in the future lifetime of an individual. In our analysis, following MacDonald (1996) and Klein and Moeschberger (1997), we treat the future lifetime as a random variable, denoted by T, that takes values in $[0, \omega)$, where ω is the limiting age, i.e. survival beyond ω is not possible. We define the survival function of T as the probability of surviving beyond time t.

$$S(t) := P(T > t)$$

This is a non-increasing function, where we require $S(0) = 1$ and $S(\omega) = 0$. In the following we have to distinguish between T as a continuous random variable and a discrete one. However, in both cases we require the following properties to hold for the survival function:

- $S(t)$ is a monotone function
- $S(t)$ is a non-increasing function
- $S(0) = 1$
- $S(\omega) = 0$, where ω is the limiting age

Continuous case

For a continuous random variable T the survival function $S(t)$ is a strictly decreasing function and the following relation holds:

$$S(t) := P(T > t) = 1 - P(T \leq t) = 1 - F(t) = 1 - \int_0^t f(u)du = \int_t^\infty f(u)du$$

$F(t) := P(T \leq t)$ is the cumulative distribution function and $f(t)$ is the density function of the random variable T, thus $dF(t) = f(t)dt$. Further we obtain:

$$f(t) = -\frac{dS(t)}{dt} \qquad \Rightarrow \qquad -dS(t) = f(t)dt$$

which is the probability of death at time t.

Discrete case

For a discrete random variable T the survival function $S(t)$ is a step-function with jumps at the times of death and constant in between. If we define the quantity $p_j := P(T = t_j)$, it follows for $t_j \leq t < t_{j+1}$ that

$$S(t) = S(t_j) = P(T > t_j) = \sum_{j:t_j>t} p_j \qquad j = 1, \ldots, k$$

where $t_1 < t_2 < \ldots$ are the times deaths are observed.

As an example we plotted in the following figure the survival curve $S(t)$ based on one-year mortality-rates for z-year old females and males according to the Bavarian life tables 1986-1988 (see Appendix B.2):

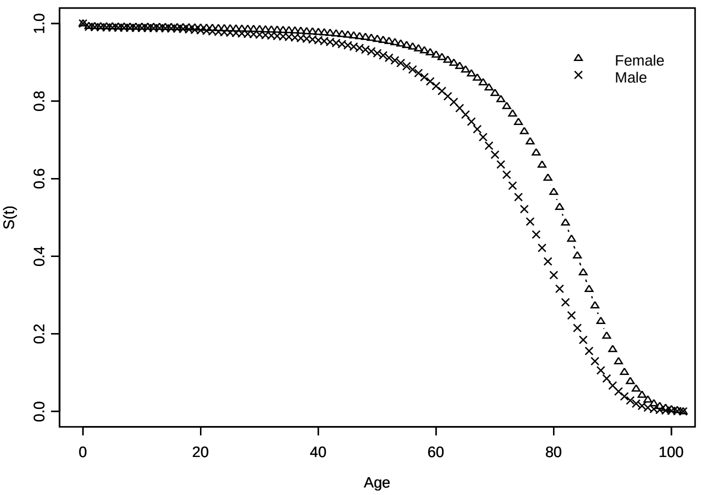

Figure 2.1: Example: Survival Curve $S(t)$

2.2 The Hazard-Rate Function

The hazard-rate function, that we will denote by $\lambda(t)$, describes the change of the chance of death over time. Generally it is assumed that the probability of death increases with age, i.e. the hazard-rate function increases with time. A constant hazard-rate function for all times t is not assumed to be adequate. Further we require the hazard-rate function $\lambda(t)$ to be non-negative, that is $\lambda(t) \geq 0$ for all t. "Force of Mortality", "Force of Transition" and "Transition Intensity" are synonymously used for "Hazard-Rate" in the literature.

Continuous case

For a continuous random variable T we define the hazard-rate function as

$$\lambda(t) := \lim_{\triangle t \to 0} \frac{P(t < T \leq t + \triangle t | T > t)}{\triangle t}$$

assuming that this limit exists. We can write for $\lambda(t)$

$$\begin{aligned} \lambda(t) &= \lim_{\triangle t \to 0} \frac{P(< T \leq t + \triangle t | T > t)}{\triangle t} \\ &= \lim_{\triangle t \to 0} \frac{P(t < T \leq t + \triangle t)}{\triangle t \quad P(T > t)} \\ &= \lim_{\triangle t \to 0} \frac{P(t < T \leq t + \triangle t)}{\triangle t} \cdot \frac{1}{S(t)} \\ &= \frac{dF(t)}{S(t)} = -\frac{dS(t)}{S(t)} = -d \ln S(t) \end{aligned} \tag{2.1}$$

The hazard-rate function at time t can be interpreted for small $\triangle t$ as the probability of dying in the interval $[t, t + \triangle t)$, i.e.

$$\lambda(t) \cdot \triangle t \approx P(\text{an individual of } t \text{ years dies in the small interval } [t, t + \triangle t)).$$

The cumulative hazard-rate function $\Lambda(t)$ is defined as the integral of the hazard-rate function over the interval $[0, t]$, that is

$$\Lambda(t) := \int_0^t \lambda(u) du.$$

Using (2.1) the following relation holds:

$$\Lambda(t) = \int_0^t \lambda(u) du = -\int_0^t d \ln S(u) du = -\ln S(t) + \ln \underbrace{S(0)}_{=1} = -\ln S(t)$$

Thus we derived the well known relationship between the survival function and the cumulative hazard-rate function or hazard-rate function, respectively:

$$S(t) = \exp\{-\Lambda(t)\} = \exp\left\{-\int_0^t \lambda(u) du\right\}$$

Discrete case

For a discrete random variable T with values $t_1 < t_2 < \ldots$ we define the hazard-rate function on $[0, t_1)$ as zero and on $[t_j, t_{j+1})$, $j = 1, 2, \ldots$, as

$$\lambda(t_j) := P(T = t_j | T \geq t_j) \tag{2.2}$$

Using $P(T = t_j) = P(T \geq t_j) - P(T \geq t_{j+1}) = S(t_{j-1}) - S(t_j)$ the hazard-rate function can be represented, as

$$\begin{aligned} \lambda(t_j) &= P(T = t_j | T \geq t_j) = \frac{P(T = t_j)}{P(T \geq t_j)} \\ &= \frac{S(t_{j-1}) - S(t_j)}{S(t_{j-1})} = 1 - \frac{S(t_j)}{S(t_{j-1})} \end{aligned} \tag{2.3}$$

The hazard-rate is constant on the interval $[t_j, t_{j+1})$. It can be interpreted as the probability of dying in the small interval $[t_j, t_{j+1})$, i.e.

$$\lambda(t_j) \approx P(\text{an individual of } t_j \text{ years dies in the small interval } [t_j, t_{j+1})).$$

The cumulative hazard-rate function $\Lambda(t)$ is defined as the sum of the hazard-rate function at times $t_j \leq t$, that is

$$\Lambda(t) := \sum_{j:t_j \leq t} \lambda(t_j).$$

With (2.3) the following relationship holds:

$$\Lambda(t) = \sum_{j:t_j \leq t} \lambda(t_j) = \sum_{j:t_j \leq t} \left(1 - \frac{S(t_j)}{S(t_{j-1})}\right)$$

Since $S(t_0) = 1$, we derive, using (2.3), for $t_j \leq t < t_{j+1}$ the following relationship:

$$S(t) = S(t_j) = \prod_{j:t_j \leq t} \frac{S(t_j)}{S(t_{j-1})} = \prod_{j:t_j \leq t} (1 - \lambda(t_j))$$

If we can decompose the discrete hazard-rate function $\lambda(t_j)$ into a sum of a continuous function $\lambda_c(t)$ and a step-function with mass $\lambda_d(t_j)$ at times $t_1 < t_2 < \ldots$, i.e.

$$\lambda(t) = \lambda_c(t) + \lambda_d(t),$$

we can represent the survival function, according to Klein (1996), using the product-integral representation (see Section 4.2.3) as follows:

$$S(t) = \prod_{j:t_j \leq t} (1 - \lambda_d(t_j)) \cdot \exp\left\{-\int_0^t \lambda_c(u)du\right\}$$

Further details on product-integration, the definition of a product-integral and its properties, can be found in the book by Andersen, Borgan, Gill, and Keiding (1993).

2.3 The Future Lifetime of an z-year old Individual: T_z

We can extend the random variable T, i.e. the future lifetime of a new-born, to ages $z > 0$. We define T_z as the future lifetime of a z-year old individual after age z, where $T_0 = T$. Note that this is conditional on surviving up to age z. We denote by $F_z(t)$ the cumulative distribution function and by $S_z(t)$ the survival function of the random variable T_z.

$$F_z(t) := P\left(T \leq z+t \middle| T > z\right) \qquad S_z(t) := P\left(T > z+t \middle| T > z\right)$$

The relationship between the cumulative distribution function of the future lifetime of an z-year old individual and a newborn one is the following:

$$F_z(t) = P\left(T \leq z+t \middle| T > z\right) = \frac{P\left(z < T \leq z+t\right)}{P\left(T > z\right)} = \frac{F(z+t) - F(z)}{1 - F(z)}$$

Therefore it follows for the survival function of an z-year old individual that

$$\begin{aligned} S_z(t) &= 1 - F_z(t) = 1 - \frac{F(z+t) - F(z)}{1 - F(z)} \\ &= \frac{1 - F(z) - F(z+t) + F(z)}{S(z)} \\ &= \frac{S(z+t)}{S(z)} \end{aligned}$$

The hazard-rate function for age $z+t$ becomes

$$\begin{aligned} \lambda_{z+t} := \lambda_z(t) &= \lim_{\triangle t \to 0} \frac{P\left(t < T_z \leq t + \triangle t \middle| T_z > t\right)}{\triangle t} \\ &= \lim_{\triangle t \to 0} \frac{P\left(t < T_z \leq t + \triangle t\right)}{\triangle t \cdot P\left(T_z > t\right)} \\ &= \lim_{\triangle t \to 0} \frac{F_z(t + \triangle t) - F_z(t)}{\triangle t \cdot (1 - F_z(t))} \\ &= \lim_{\triangle t \to 0} \frac{\frac{F(z+t+\triangle t) - F(z)}{1-F(z)} - \frac{F(z+t)-F(z)}{1-F(z)}}{\triangle t \cdot \frac{1-F(z+t)}{1-F(z)}} \\ &= \lim_{\triangle t \to 0} \frac{F(z+t+\triangle t) - F(z+t)}{\triangle t \cdot (1 - F(z+t))} \\ &= \lim_{\triangle t \to 0} \frac{F(z+t+\triangle t) - F(z+t)}{\triangle t} \cdot \frac{1}{S(z+t)} \end{aligned}$$

In actuarial notation we have $F_z(t) := {}_tq_z$ and $S_z(t) := {}_tp_z$. We denote the probability density function of T_z by $f_z(t)$. It can be represented in terms of the hazard-rate function:

$$\begin{aligned} f_z(t) &= \frac{\partial}{\partial t} F_z(t) = \lim_{\triangle t \to 0} \frac{F_z(t+\triangle t) - F_z(t)}{\triangle t} \\ &= \frac{S(z+t)}{S(z)} \cdot \lim_{\triangle t \to 0} \frac{F(z+t+\triangle t) - F(z+t)}{\triangle t} \cdot \frac{1}{S(z+t)} = {}_tp_z \lambda_{z+t} \end{aligned}$$

We noted before that $F_z(t) = {}_tq_z$; thus it follows that $F_z(t) = 1 - {}_tp_z$ and we can rewrite the above formula as

$$\frac{\partial}{\partial t}F_z(t) = \frac{\partial}{\partial t}\left(1 - {}_tp_z\right) = -\frac{\partial}{\partial t}{}_tp_z = {}_tp_z\lambda_{z+t} \qquad \Rightarrow \qquad \frac{\partial}{\partial t}{}_tp_z = -{}_tp_z\lambda_{z+t}$$

Thus it follows

$$-\lambda_{z+t} = \frac{1}{{}_tp_z}\cdot\frac{\partial}{\partial t}{}_tp_z = \frac{\partial}{\partial t}\ln{}_tp_z$$

Integrating both sides and using the boundary condition ${}_0p_z = P(T_z > 0) = 1$ we derive

$$\ln{}_tp_z - \ln\underbrace{{}_0p_z}_{=1} = \int_0^t \frac{\partial}{\partial t}\ln{}_rp_z dr = -\int_0^t \lambda_{z+r}dr$$

gives us

$${}_tp_z = \exp\left\{-\int_0^t \lambda_{z+r}dr\right\},$$

which is the well-known relationship in actuarial sciences between hazard-rate function and survival function. It has been used by Gompertz (1825) and Makeham (1860) to derive their laws of mortality, i.e. a representation of the hazard-rate function as a mathematical functions of age z (see Trachtenberg (1924)):

$$\text{Gompertz' law} : \quad \lambda_z = B\cdot c^z \qquad\qquad \text{Makeham's law} : \quad \lambda_z = A + B\cdot c^z$$

Using above formula for ${}_tp_z$ we can write in the case of Makeham's law:

$$\begin{aligned}
{}_tp_z &= \exp\left\{-\int_0^t \lambda_{z+r}dr\right\}\\
&= \exp\left\{-\int_0^t A + B\cdot c^{z+r}dr\right\}\\
&= \exp\left\{-A\cdot t - B\cdot c^z\int_0^t c^r dr\right\}\\
&= \exp\{-A\cdot t\}\cdot\exp\left\{-\frac{B}{\ln c}\cdot c^z\cdot\left(c^t - 1\right)\right\}\\
&= \exp\{-A\}^t\cdot\exp\left\{-\frac{B}{\ln c}\right\}^{c^z\cdot\left(c^t-1\right)}\\
&= s^t\cdot g^{c^z\cdot(c^t-1)}
\end{aligned}$$

where $s := \exp\{-A\}$ and $g := \exp\{-B/\ln c\}$.

Since this is an equation with three unknowns, three values of ${}_tp_z$ at different ages are sufficient to calculate the parameters A, B and c, and thus the hazard-rate function for all ages z.

In the case of Gompertz' law the factor s^t vanishes and we are left with an equation with two unknowns. Thus only two values of ${}_tp_z$ are required to calculate the parameters B and c for Gompertz' law. Of course, an inapt choice of the ${}_tp_z$'s leads to a wrong survival probability and spurious results in both cases.

2.4 Censoring

One problem in survival analysis is the estimation of the survival distribution function $S(t)$ of a population. For a human population we could observe n newborn lives until death, to get an estimate for $S(t)$. A very obvious problem of such an experiment is the very long duration. Further many lives would be lost during the experiment due to migration or withdrawal and bias the result. Thus additional methods have to be used to account for this.

Therefore in survival analysis experiments are designed for a shorter period of time only, and secondly, have to account for the lost observations. If we observe a sample only for a short period of time, we only know that some individuals were alive at the end of the survey and no information on their exact time of death is available. If observations are lost during the experiment, all we know is that these individuals were still alive at some stage and no information on their exact time of death is available.

Compared with the usual statistical analysis in survival analysis only limited information is available. The process accounting for this is called censoring. In the following sections we focus on this "key feature of survival data analysis", as noted by MacDonald (1996). He considers survival analysis as the "analysis of censored data". The different types of censoring are explained in the following sections, where we follow MacDonald (1996) and Klein and Moeschberger (1997).

Right-Censoring

Data are called right-censored, if the current survey ends at a fixed date known in advance. If the event of interest happens after this date, the observation is censored. All we know in this case is that the event might have happened after the end of the survey.

For example, if we are interested in the lifetime of individuals, we only know that the censored observations were alive at the end of our survey and nothing about their times of death after the end of the survey. Only for the uncensored observations we know the exact lifetimes.

We introduce the following notation: Denote the lifetime of each individual by the random variable X and define the censoring time by $C^{(r)}$. If the lifetime is smaller or equal than the censoring time, we know the exact lifetime of this individual, whereas if the lifetime is greater than the censoring time, we only know that this individual survived its censoring time. We denote the actual observed lifetime of an individual by T.

$$\begin{array}{lllll} X \leq C^{(r)} & \Rightarrow & T = X & \Rightarrow & \text{observation is not censored} \\ X > C^{(r)} & \Rightarrow & T = C^{(r)} & \Rightarrow & \text{observation is censored} \end{array}$$

Thus the observed lifetime T can be calculated as

$$T = min\left(X, C^{(r)}\right) = \begin{cases} X & \text{if the exact lifetime is observed} \\ C^{(r)} & \text{if the exact lifetime is not observed} \end{cases}$$

Further we introduce an indicator δ, called the "death indicator". This indicator is a random variable, indicating if the observation has been censored or not:

$$\delta = \begin{cases} 1 & \text{if the individual is not censored, i.e. the exact lifetime is observed} \\ 0 & \text{if the individual is censored, i.e. the exact lifetime is not observed} \end{cases}$$

The pair of random variables (T, δ) represents now all available information in the data. As an example consider the case described in Figure 2.2: Death of observations 1 and 3 are observed within the survey, thus their exact lifetimes X_1 and X_3 are known. Observations 2 and 4 survive beyond the end of the survey. Their exact lifetimes are unknown and $C_2^{(r)}$ and $C_4^{(r)}$ are their censoring times, respectively.

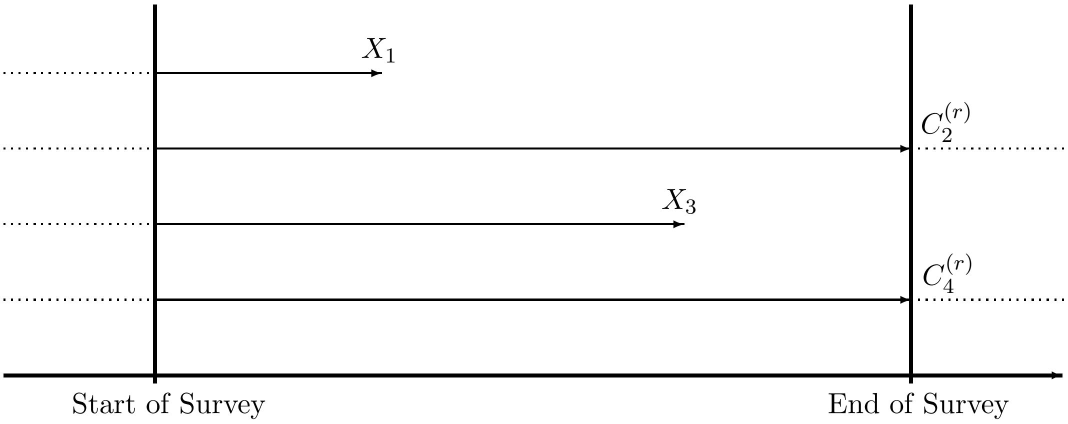

Figure 2.2: Right-Censoring

Further one distinguishes Type I and Type II censoring: In the case of Type I censoring we specify for each individual a fixed point in time until this individual is observed. The censoring time $C^{(r)}$ is known in advance. If the event did not occur, the observation is censored. Therefore the number of individuals in the survey is random, whereas the observation time is fixed.

In the case of Type II censoring, this is the other way round: We start with n individuals and observe these individuals until $r < n$ events, are observed. Thus this observation consists of the r smallest lifetimes and the theory of order statistics can be applied. As already mentioned, in this case the number of individuals in the survey is fixed and the observation time is random.

Further special cases of Type I censoring are the so-called progressive and generalized Type I censoring. In the first case we have different, fixed-sacrifice censoring times, and in the latter the individuals enter the survey at different times and are then observed for a certain predetermined time. Progressive censoring is also a special case of Type II censoring, where the first $r_1, r_2, \dots$ events are observed, giving random observation times $T_{r_1}, T_{r_1+r_2}, \dots$.

Left-Censoring

Data are called left-censored, if no information on the date, at which the event of interest occurred, is available. An example for left-censored data are medical studies, where patients are examined and we only know that a certain disease occurred before the examination. Thus the event of interest has already occurred before time $C^{(l)}$ and the exact time of occurrence is unknown. The exact time of occurrence X is less than the censoring time $C^{(l)}$. We only know the exact time of occurrence, if X is greater than or equal to $C^{(l)}$. Analogue to the case of right-censored observations we can represent the data by a pair of random variables (T, ε):

$$\begin{array}{lllll} X \geq C^{(l)} & \Rightarrow & T = X & \Rightarrow & \text{observation is not censored} \\ X < C^{(l)} & \Rightarrow & T = C^{(l)} & \Rightarrow & \text{observation is censored} \end{array}$$

Thus the observed time of occurrence T can be calculated as

$$T = max\left(X, C^{(l)}\right) = \begin{cases} X & \text{if the exact time of occurrence is observed} \\ C^{(l)} & \text{if the exact time of occurrence is not observed} \end{cases}$$

The indicator ε is a random variable indicating, if the observation has been censored or not:

$$\varepsilon = \begin{cases} 1 & \text{if the individual is not censored, i.e. the exact lifetime is observed} \\ 0 & \text{if the individual is censored, i.e. the exact lifetime is not observed} \end{cases}$$

Interval-Censoring

If we generalize left and right censoring and combine both this leads us to Interval-Censoring. Data are called interval-censored, if we only know that the event of interest fell within an interval of time $(L_i, R_i]$, where L_i denotes the left and R_i the right endpoint of the interval.

An example for interval-censored data are actuarial studies, where only the calendar year of death is known or medical studies, where periodic follow-up takes place and a disease is only known to have occurred in the time between the last and current follow-up.

Consequently, the data are left-censored, if we use the interval $(0, C^{(l)}]$, and right-censored, if the interval is $(C^{(r)}, \infty]$. In the first case the event of interest occurred before $C^{(l)}$ and in the last case, the event of interest takes place after time $C^{(r)}$.

Random-Censoring

We call a censoring mechanism random-censoring, if the censoring time C_i is a random variable. This is opposed to right-censoring, where the censoring time $C^{(r)}$ is known in advance.

The analogue to right-censoring is, if we censor an observation i, in the case that C_i is smaller than its lifetime X_i, the analogue to left-censoring is, if we censor an observation i, in the case that C_i is greater than its lifetime X_i.

2.5 Truncation

Truncation is defined as a condition which screens certain subjects so that the investigator will not be aware of their existence. We only include individuals in our survey that fulfill a certain condition. This condition might be the occurrence of an event prior to the actual event of interest. In this case the data are left-truncated.

For example if we observe in a medical survey only individuals that have been exposed to a special disease, or if we observe individuals above a certain age. Individuals that experience the event of interest earlier, e.g. at a younger age, are not included in the survey. If Y is the time used for truncation, we only observe individuals, which fulfill $X \geq Y$. This is in contrast to left-censoring, where this individual would be included and we would make use of the information that this individual experienced the event of interest prior to time $C^{(l)}$.

Individuals are right-truncated if we include only individuals into our survey that already have fulfilled a certain condition, i.e. they have already experienced the event of interest. An example for right-truncated data is a mortality survey based on death records.

In the case of truncated data we have to use a conditional distribution. This has to be accounted for when constructing likelihood functions.

2.6 Likelihood Construction

The different censoring mechanisms mentioned in Section 2.4 affect also the way in which the likelihood function has to be constructed, since we do not know the exact survival time of each individual. The independence of individuals and the independence of censoring times are important assumptions that should be carefully considered. If we observe the exact lifetime of an individual, the contribution to the likelihood function is of course $f(x_i)$.

For right-censored individuals, we only know that they survived at least to time $C^{(r)}$. Therefore the contribution of these individuals has to be $S(C^{(r)})$. Similarly, left-censored individuals, contribute $1 - S(C^{(l)})$, as we only know that the event of interest has already happened before time $C^{(l)}$. As interval-censoring is a combination of right- and left-censoring and therefore we obtain a contribution of $S(L_i) - S(R_i)$.

In the case of truncated data, as already mentioned in Section 2.5, we have to use conditional probabilities. Thus a left-truncated observation contributes $f(x)/S(Y)$ and a right-truncated observation contributes $f(Y)/(1-S(Y))$ to the likelihood function. In the first case observations have to survive to time Y without experiencing the event of interest and in the second case they have to experience the event at time Y to be included into the survey, for examples we include only deaths into the survey.

If censoring and left-truncation is combined and the truncation time Y is independent from the death time, we replace $f(x_i)$ by $f(x_i)/S(Y)$, $S(C^{(l)})$ by $S(C^{(l)})/S(Y)$ and $S(C^{(r)})$ by $S(C^{(r)})/S(Y)$. Collecting the observation with exact death times in the set D, the right-censored observations in the set R, the left-censored observations in the set L and the interval-censored observations in the set I,we obtain for the likelihood function combining all elements:

$$L = \prod_{i \in D} f(x_i) \cdot \prod_{i \in R} S(C^{(r)}) \cdot \prod_{i \in L} \left(1 - S(C^{(l)})\right) \cdot \prod_{i \in I} (S(L_i) - S(R_i)) \tag{2.4}$$

In the following we show, how we get to this formula in the case of right-censored data. We represent right-censored data by pairs of random variables T_i, δ_i, where T_i indicates the lifetime of individual i and the indicator variable δ_i tells us, if the observation i has been censored at time T_i ($\delta_i = 0$ and consequently $T_i = C^{(r)}$) or died at time T_i ($\delta_i = 1$ and $T_i = x_i$). Denoting X_i as the actual lifetime of the individual i, we can write $T_i = min(X_i, C^{(r)})$. Assuming that the individual is censored ($\delta_i = 0$) we obtain:

$$\begin{aligned} P(T_i, \delta_i = 0) &= P(T_i = C^{(r)} \Big| \delta_i = 0) \cdot P(\delta_i = 0) \\ &= P(\delta_i = 0) = P(X_i > C^{(r)}) = S(C^{(r)}) \end{aligned}$$

If the individual is not censored ($\delta_i = 1$), we have

$$\begin{aligned} P(T_i, \delta_i = 1) &= P(T_i = C^{(r)} \Big| \delta_i = 1) \cdot P(\delta_i = 1) \\ &= P(C^{(r)} = T_i \Big| X_i \leq C^{(r)}) \cdot P(X_i \leq C^{(r)}) \\ &= \frac{f(t_i)}{1 - S(C^{(r)})} \cdot \left(1 - S(C^{(r)})\right) = f(t_i) \end{aligned}$$

We can summarize both cases in one single expression:

$$P(t_i, \delta_i) = f(t_i)^{\delta_i} \cdot S(t_i)^{(1-\delta_i)}$$

This leads us for n independent individuals (T_i, δ_i) to the likelihood function fulfilling equation (2.4), that is in the case of right-censored data

$$L = \prod_{i=1}^{n} P(t_i, \delta_i) = \prod_{i=1}^{n} f(t_i)^{\delta_i} \cdot S(t_i)^{(1-\delta_i)}$$

In the case of left-censored data represented by pairs of random variables T_i, ε_i, where T_i indicates the lifetime of individual i and the indicator variable ε_i tells us now, if the observation i has been left-censored at time T_i ($\varepsilon_i = 0$ and $T_i = C^{(l)}$) or died at time T_i ($\varepsilon_i = 1$ and $T_i = X_i$). Denoting the X_i as the actual lifetime of the individual i, we can write $T_i = max(X_i, C^{(l)})$. Assuming that the individual is left-censored ($\varepsilon_i = 0$) we obtain now:

$$\begin{aligned} P(T_i, \varepsilon_i = 0) &= P(T_i = C^{(l)} \Big| \varepsilon_i = 0) \cdot P(\varepsilon_i = 0) \\ &= P(\varepsilon_i = 0) = P(X_i < C^{(l)}) = 1 - S(C^{(l)}) \end{aligned}$$

If the individual is not censored ($\varepsilon_i = 1$), we have

$$\begin{aligned} P(T_i, \varepsilon_i = 1) &= P(T_i = C^{(l)} \Big| \varepsilon_i = 1) \cdot P(\varepsilon_i = 1) \\ &= P(C^{(l)} = T_i \Big| X_i \geq C^{(l)}) \cdot P(X_i \geq C^{(l)}) \\ &= \frac{f(t_i)}{S(C^{(l)})} \cdot S(C^{(l)}) = f(t_i) \end{aligned}$$

We can summarize both cases in one single expression:

$$P(t_i, \varepsilon_i) = f(t_i)^{\varepsilon_i} \cdot (1 - S(t_i))^{(1-\varepsilon_i)}$$

This leads us for n independent individuals (T_i, ε_i) to the likelihood function fulfilling equation (2.4), which is in the case of left-censored data:

$$L = \prod_{i=1}^{n} P(t_i, \varepsilon_i) = \prod_{i=1}^{n} f(t_i)^{\varepsilon_i} \cdot (1 - S(t_i))^{(1-\varepsilon_i)}$$

Chapter 3

Markovian Multi-State Models

In this chapter we explain the setup of discrete and continuous-time Markovian multi-state models. Multi-state models are a common tool to describe the life-history of an individual and thus ideal for our purposes. Through its lifetime, from birth to death, each individual visits different states. Some might be visited only once, some more frequently and even some not at all. Certainly the state "Dead" is eventually visited. We have to choose the states of the model in such a way that at each point in time an individual can be contained in exactly one state. Individuals are observed over time and we record their transition between states.

For example in the Illness-Death model (Figure 3.1) there are three states, "Disease-Free", "Diseased" or "Dead". Generally we label the states from $1, \dots, K$ and collect them in the state space, denoted by $\mathcal{S}$. The set $\mathcal{S}$ is supposed to be finite. The states should correspond to the events we want to observe that result then in transitions between different states.

These transitions between different states can be described by transition probabilities or transition intensities. We are going to define these quantities and derive their properties such as the Chapman-Kolmogorov equations, as well as the relationship between the transition probabilities and transition intensities which is given by the Kolmogorov differential equations. The transition probabilities are collected in the transition matrix. This matrix is responsible for the development of of the Markov chain in discrete-time or the Markov process in continuous-time, and thus the quantity of interest. In Chapter 4 we are then going to define a non-parametric estimator for this transition matrix.

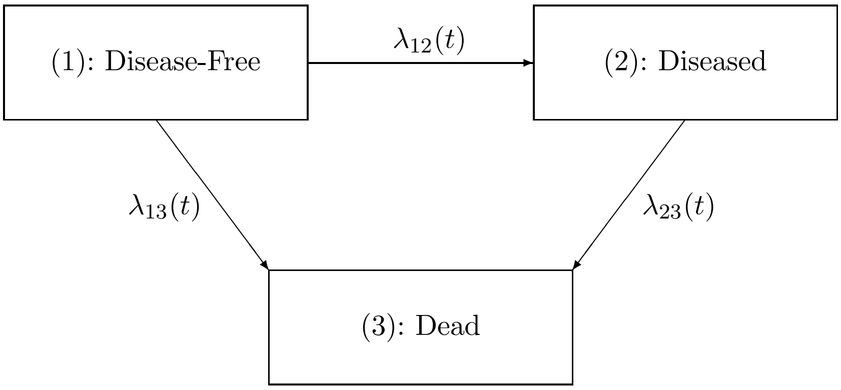

Figure 3.1: Illness-Death model

In the Illness-Death model the events "Recover from Disease", "Disease" and "Death" are the events of interest. We want to calculate probabilities like the probability of being "Disease-Free", "Diseased" or "Dead" at time t.

In Figure 3.2 we sketched the possible life-history for three individuals: The first individual (Person A) might start in state "Disease-Free", get ill after some time and recover again. Another individual (Person B) dies after being diseased for some time and a third one (Person C) dies without being diseased at all.

It is obvious that, already in this three-state model, various states have different properties. Whereas transitions in both directions are possible between the states "Disease-Free" and "Diseased", there is only the transition from "Disease-Free" to "Dead" or "Diseased" to "Dead" possible. We call the states "Disease-Free" and "Diseased" transient states and the state "Dead" an absorbing state, as no transition out of this state is possible.

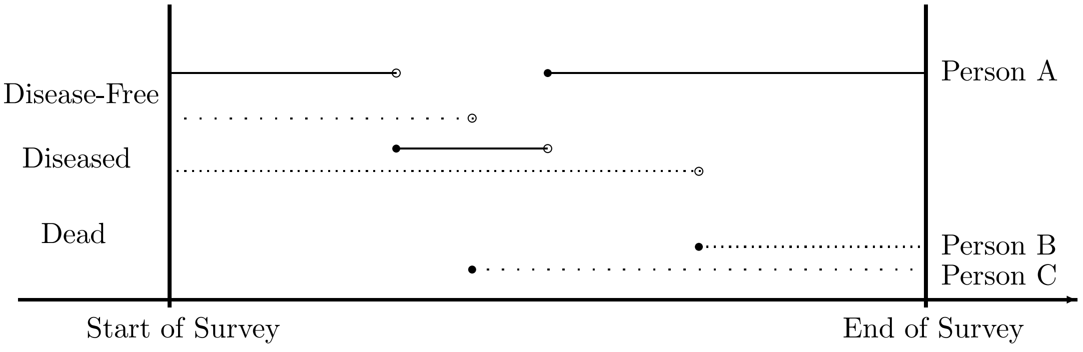

Figure 3.2: Life-history in the Illness-Death model

Definition 3.1 (States of a Multi-State Model) *Haberman and Pitacco (1999) defined the different states, that might occur in a multi-state model, as follows:*

- *transient state: it is possible to leave and to re-enter this state;*
- *strictly transient state: it is not possible to enter this state once it has been left;*
- *absorbing state: it is not possible to leave this state once it has been entered;*

Generally the set of possible transitions between states, we want allow for in our model, is a subset of the set of pairs (i, j). We call this set $\mathcal{F}$ and write

$$\mathcal{F} \subseteq \{(i,j)|\, i \neq j; i,j \in \mathcal{S}\}$$

The pair $(\mathcal{S}, \mathcal{F})$ is called a multi-state model. It describes the different states, an individual can be observed in, and determines the possible transitions between the states, that can be visited by an individual. We assume that from the initial state all states $j \in \mathcal{S}$ can be reached. This is in contrast to the general Markovian multi-state model, where no-accessible states are possible.

3.1 The time-discrete Markov Model

Definition 3.2 (Time-discrete Markov Chain) *Consider a time-discrete stochastic process $S(t)$, $t = 0, 1, \ldots \in T$, with a finite state space $\mathcal{S}$ We say that $S(t)$, $t = 0, 1, \ldots$, is a time-discrete Markov chain if, for any n and each finite set of integer times $0 \leq t_0 < \ldots < t_n < t < u \in T$ and corresponding set of states $i_0, \ldots, i_n, i, j$ in $\mathcal{S}$ and $T \in \mathbb{R}$ with*

$$P(S(t_0) = i_0, \ldots, S(t_n) = i_n, S(t) = i, S(u) = j) > 0,$$

the Markov property holds:

$$P(\,S(u) = j|\,S(t_0) = i_0, \ldots, S(t_n) = i_n, S(t) = i) = P(\,S(u) = j|\,S(t) = i) \tag{3.1}$$

Thus for a Markov chain the probability of being in state j in the future, i.e. at time u, only depends on the exact present, i.e. the state occupied at time t. The history or past of the process from time t_0 up to time t_n is irrelevant. One also can say that, when the present is fixed, the future and the past of the process are conditionally independent.

Theorem 3.1 *A stochastic process $\{S(t),\, t \in T\}$, is a Markov Chain if, and only if the following holds for all $n \geq 1$, $t_0 < \ldots < t_n \in T$ and $i_0, \ldots, i_n \in \mathcal{S}$:*

$$P(S(t_0) = i_0, \ldots, S(t_n) = i_n) = P(S(t_0) = i_0) \cdot \prod_{k=0}^{n-1} p_{i_k, i_{k+1}}(t_k, t_{k+1}) \tag{3.2}$$

where $p_{i_k, i_{k+1}}(t_k, t_{k+1})$ is the probability of transferring to state i_{k+1} by time t_{k+1} given that the Markov Chain has been in state i_k at time t_k (see Definition 3.4).

Proof:

"$\Longrightarrow$"
Assume $S(t)$, $t \in T$, is a Markov Chain with $P(S(t_0) = i_0, \ldots, S(t_n) = i_n) > 0$. Using the Markov property (3.1) it follows that

$$P(S(t_0) = i_0, \ldots, S(t_n) = i_n) = P(S(t_0) = i_0, \ldots, S(t_{n-1}) = i_{n-1}) \cdot p_{i_{n-1}, i_n}(t_{n-1}, t_n)$$

Using complete induction we derive above equation (3.2).

"$\Longleftarrow$"
Assume equation (3.2) holds, then we have for $P(S(t_1) = i_1, \ldots, S(t_n) = i_n) > 0$

$$\begin{aligned}
&P(\,S(t_n) = i_n|\,S(t_0) = i_0, \ldots, S(t_{n-1}) = i_{n-1}) = \\
&\frac{P(S(t_n) = i_n, S(t_0) = i_0, \ldots, S(t_{n-1}) = i_{n-1})}{P(S(t_0) = i_0, \ldots, S(t_{n-1}) = i_{n-1})} = \\
&\frac{P(S(t_0) = i_0) \cdot \prod_{k=0}^{n-1} p_{i_k, i_{k+1}}(t_k, t_{k+1})}{P(S(t_0) = i_0) \cdot \prod_{k=0}^{n-2} p_{i_k, i_{k+1}}(t_k, t_{k+1})} = \\
&p_{i_{n-1}, i_n}(t_{n-1}, t_n) = P(\,S(t_n) = i_n|\,S(t_{n-1}) = i_{n-1})
\end{aligned}$$

which is nothing else than the Markov property (3.1). Thus according to the definition it follows that $S(t)$, $t \in T$, is a Markov chain. □

Definition 3.3 (Time-Homogeneous Markov Chain) *We call a Markov chain $S(t)$, $t \in T$, time-homogeneous, if for all $s, t \in \mathbb{R}$, $i, j \in \mathcal{S}$ with $P(S(s) = i) > 0$ and $P(S(t) = i) > 0$ and $h > 0$ the following property holds:*

$$P(\,S(s+h) = j|\,S(s) = i) = P(\,S(t+h) = j|\,S(t) = i)$$

Otherwise the process is said to be time-inhomogeneous.

3.1.1 Transition Probabilities

Definition 3.4 (Transition Probabilities) *The conditional probability of being in state j at time u given having been in state i by time t, i.e. $P(\,S(u) = j|\,S(t) = i)$, is called transition probability and denoted by $p_{ij}(t,u)$.*

$$p_{ij}(t,u) := P(\,S(u) = j|\,S(t) = i) \qquad 0 \le t < u \in T \qquad i, j \in \mathcal{S}$$

It is worth mentioning that this definition applied to $p_{ii}(t,u)$ does not require the process to stay in state i in the interval (t,u). Excursions to other states are possible. The individual starts at time t in state i and has to be back in state i by u. Transitions to other states in (t,u) are not excluded. If we do not want to allow for transition to other states in (t,u), we have to introduce a modified transition probability $p_{\underline{ii}}(t,u)$; this will be done in Section 3.1.2.

The transition probabilities are the quantities that determine the behavior of the Markov chain. For a time-homogeneous Markov chain the transition probabilities do not depend on the time and therefore we define

$$p_{ij}(h) := p_{ij}(s, s+h)$$

This means that for a time-homogeneous Markov chain the transition probabilities depend only on the time difference and not from the time-level reached. For example in a simple two-state model the probability of dying in the next ten years for a 20-year old live would be the same as for a 50-year old live.

The conditional probabilities also satisfy the so-called Chapman-Kolmogorov equations, which are written in the discrete case as

Theorem 3.2 (Chapman-Kolmogorov Equations)

$$p_{ij}(t,u) = \sum_{k \in \mathcal{S}} p_{ik}(t,w) \cdot p_{kj}(w,u) \qquad t \le w \le u \qquad \forall i, j \in \mathcal{S} \tag{3.3}$$

Proof:

For $w = t$ or $w = u$ this is obviously true (see Definition 3.5).

Let $t < w < u$ and define using the assumption $P(S(t) = i) > 0$

$$\begin{aligned} \mathcal{S}^* &:= \{\,k \in \mathcal{S}|\,p_{ik}(t,w) \ne 0\} \\ &= \{\,k \in \mathcal{S}|\,P(S(w) = k|S(t) = i) \ne 0\} \\ &= \{\,k \in \mathcal{S}|\,P(S(w) = k, S(t) = i) \ne 0\} \end{aligned}$$

We can then write $p_{ij}(t,u)$ as follows:

$$\begin{aligned} p_{ij}(t,u) &= P(\,S(u)=j|\,S(t)=i) \\ &= \sum_{k\in\mathcal{S}^*} P(\,S(u)=j, S(w)=k|\,S(t)=i) \\ &= \sum_{k\in\mathcal{S}^*} P(\,S(w)=k|\,S(t)=i)\cdot P(\,S(u)=j|\,S(w)=k, S(t)=i) \\ &= \sum_{k\in\mathcal{S}^*} p_{ik}(t,w)\cdot p_{kj}(w,u) \\ &= \sum_{k\in\mathcal{S}} p_{ik}(t,w)\cdot p_{kj}(w,u) \end{aligned}$$

□

Haberman and Pitacco (1999) use a special case of the Chapman-Kolmogorov equations, setting $w = t+1$. The equations (3.3) can then be rewritten as

$$p_{ij}(t,u) = \sum_{k\in\mathcal{S}} p_{ik}(t,t+1)\cdot p_{kj}(t+1,u)$$

This means that we can derive any conditional probability $p_{ij}(t,u)$ from the set of one-year transition probabilities $p_{ik}(t,t+1)$. We denote the one-year transition probabilities by $p_{ik}(z) := p_{ik}(z,z+1)$, where $z = 0,1,\ldots$. If $p_{ik}(z)$ is independent of z, the corresponding process is time-homogeneous. That is at each age the probability of transition is the same.

Definition 3.5 (Transition Matrix) *A family $(p_{ij}(t,u))_{i,j}$ is called a transition matrix, if the following properties are fulfilled:*

- $p_{ij}(t,u) \geq 0 \qquad \forall i,j \in \mathcal{S}$ *and* $t \leq u \in T$
- $\sum_{j\in\mathcal{S}} p_{ij}(t,u) = 1 \qquad \forall i \in \mathcal{S}$
- *For* $P(S(t)=i) > 0$ *the following holds:*

$$p_{ij}(t,t) = \begin{cases} 1 & i=j \\ 0 & i\neq j \end{cases} \qquad \forall i,j\in\mathcal{S}$$

- $p_{ij}(t,u) = \sum_{k\in\mathcal{S}} p_{ik}(t,w)\cdot p_{kj}(w,u)$ *for* $t \leq w \leq u \in T$, $P(S(t)=i) > 0$ *and* $i,j \in \mathcal{S}$

Theorem 3.3 *For a Markov chain $S(t)$, $t \in T$, $(p_{ij}(t,u))_{i,j}$ is a transition matrix.*

Proof:

This theorem follows directly from the Chapman-Kolmogorov equations. □

3.1.2 Occupancy Probabilities

Definition 3.6 (Occupancy Probabilities) *The following probabilities are called the occupancy probabilities:*

$$p_{\underline{ii}}(t,u) = P(\,S(z) = i \textit{ for all } z \in [t,u] |\, S(t) = i)$$

where $t < u$. In contrast to the transition probabilities the process stays in the interval $[t,u]$ in state i. In case of a transition probability it would be sufficient to be in state i at time t and after an excursion to other states to be back in state i by time u.

Because of their similar nature to the transition probabilities, the occupancy equations satisfy also a set of Chapman-Kolmogorov equations, however, a different one.

Theorem 3.4 (Chapman-Kolmogorov)

$$p_{\underline{ii}}(t,u) = p_{\underline{ii}}(t,w) \cdot p_{\underline{ii}}(w,u) \qquad t \leq w \leq u$$

The proof of this theorem is analogue to the continuous case as in Theorem 3.8, which can be found in Section 3.2.5.

3.2 The time-continuous Markov Model

The definitions of the time-continuous case are very similar to the discrete one. Therefore we only quote the necessary definitions in this section and highlight, where additional requirements are needed. Further we introduce transition intensities and derive their relationship to transition probabilities given by the Kolmogorov differential equations.

Definition 3.7 (Time-continuous Markov Chain) *Consider a time-continuous stochastic process $S(t)$, $t \geq 0$, with a finite state space $\mathcal{S}$. We say that $S(t)$, $t \geq 0$, is a time-continuous Markov chain if, for any n and each finite set of integer times $0 \leq t_0 < \ldots < t_n < t < u \in \mathbb{R}$ and corresponding set of states $i_0, \ldots, i_n, i, j \in \mathcal{S}$ with*

$$P(S(t_0) = i_0, \ldots, S(t_n) = i_n, S(t) = i, S(u) = j) > 0,$$

the Markov property holds, that is

$$P(\, S(u) = j |\, S(t_0) = i_0, \ldots, S(t_n) = i_n, S(t) = i) = P(\, S(u) = j |\, S(t) = i).$$

3.2.1 Transition Probabilities

Definition 3.8 (Transition Probabilities) *The conditional probability of being in state j at time u given having been in state i by time t, i.e. $P(\, S(u) = j |\, S(t) = i)$, is called transition probability and denoted by $p_{ij}(t, u)$:*

$$p_{ij}(t, u) := P(\, S(u) = j |\, S(t) = i) \qquad 0 \leq t < u \qquad i, j \in \mathcal{S}$$

Definition 3.9 (Time-Homogeneous Markov Chain) *We call a Markov chain S_t; $t \geq 0$, time-homogeneous, if for all $s, t \in \mathbb{R}$, $i, j \in \mathcal{S}$ with $P(S(s) = i) > 0$ and $P(S(t) = i) > 0$ and $h > 0$ the following property holds:*

$$P(\, S(s + h) = j |\, S(s) = i) = P(\, S(t + h) = j |\, S(t) = i)$$

Theorem 3.5 (Chapman-Kolmogorov Equations)

$$p_{ij}(t, u) = \sum_{k \in \mathcal{S}} p_{ik}(t, w) \cdot p_{kj}(w, u) \qquad t \leq w \leq u \qquad i, j \in \mathcal{S} \tag{3.4}$$

This is intuitively clear: We are starting in state i at time t, visit some state k at time w before we arrive in state j by time u. Summing over all possible states k, visited at time w, we obtain the probability $p_{ij}(t, u)$.

Proof:

$$\begin{aligned}
p_{ij}(t, u) &= P(\, S(u) = j |\, S(t) = i) \\
&= \sum_{k \in \mathcal{S}} P(\, S(u) = j, S(w) = k |\, S(t) = i) \\
&= \sum_{k \in \mathcal{S}} P(\, S(w) = k |\, S(t) = i) \cdot P(\, S(u) = j) |\, S(w) = k, S(t) = i) \\
&= \sum_{k \in \mathcal{S}} P(\, S(w) = k |\, S(t) = i) \cdot P(\, S(u) = j) |\, S(w) = k) \\
&= \sum_{k \in \mathcal{S}} p_{ik}(t, w) \cdot p_{kj}(w, u)
\end{aligned}$$

The unconditional probabilities $P(S(u)=j)$ can be easily computed using the initial distribution $a_i := P(S(0)=i)$ and the conditional probabilities:

$$P(S(u)=j)=\sum_{i\in\mathcal{S}} a_i \cdot p_{ij}(0,u)$$

Proof:

$$\begin{aligned} P(S(u)=j) &= \sum_{i\in\mathcal{S}} P(S(u)=j, S(0)=i) \\ &= \sum_{i\in\mathcal{S}} P(\,S(u)=j|\,S(0)=i)\cdot P(S(0)=i) \\ &= \sum_{i\in\mathcal{S}} a_i \cdot p_{ij}(0,u) \end{aligned}$$

□

Definition 3.10 (Transition Matrix) *A family $p_{ij}(t,u)$ is called a transition matrix, if the following properties are fulfilled:*

- $p_{ij}(t,u)\geq 0 \qquad \forall i,j\in\mathcal{S}$ *and* $0\leq t\leq u$
- $\sum_{j\in\mathcal{S}} p_{ij}(t,u)=1 \qquad \forall i\in\mathcal{S}$
- *For $P(S(t)=i)>0$ the following holds:*

  $$p_{ij}(t,t)=\begin{cases} 1 & i=j \\ 0 & i\neq j \end{cases} \qquad \forall i\in\mathcal{S}$$

  This is to ensure that the stochastic process is at any time in exactly one state and therefore well-defined.

- $p_{ij}(t,u)=\sum_{k\in\mathcal{S}} p_{ik}(t,w)\cdot p_{kj}(w,u)$ *for $0\leq t\leq w\leq u$, $P(S(t)=i)>0$ and $i,j\in\mathcal{S}$*

3.2.2 Transition Intensities

Definition 3.11 (Transition Intensities) *We define the transition intensities for all $i\neq j$ and $t\geq 0$ as follows:*

$$\mu_{ij}(t) := \lim_{dt\to 0}\frac{p_{ij}(t,t+dt)}{dt} = \lim_{dt\to 0}\frac{P(\,S(t+dt)=j|\,S(t)=i)}{dt} \qquad (3.5)$$

where we assume that this limit exists and that the intensities are integrable on compact intervals.

Further we define for all $i\in\mathcal{S}$ and $t\geq 0$:

$$\mu_{ii}(t) := -\sum_{j\neq i}\mu_{ij}(t)$$

For a time-homogeneous process $\mu_{ij}(t)$ reduces to time-constant function μ_{ij}:

$$\mu_{ij}(t) := \lim_{dt\to 0} \frac{p_{ij}(t, t+dt)}{dt} = \mu_{ij}$$

From this definition we can interpret the product $\mu_{ij}(t)dt$ as the conditional probability of making a transition from state i to state j in the small interval $[t, t+dt)$ after having been in state i at time t.

Definition 3.12 (Intensity of Decrement) *Summing up the transition intensities out of state i, we define the total intensity of decrement from state i as*

$$\mu_i(t) := \sum_{j\neq i} \mu_{ij}(t)$$

We can interpret $\mu_i(t)$ as the conditional probability of leaving state i in the small interval $[t, t+dt)$ to any other state, given to be in state i at time t. We can re-write $\mu_i(t)$ in terms of the transition probabilities using (3.5):

$$\begin{aligned}
\mu_i(t) &= \sum_{j\neq i} \lim_{dt\to 0} \frac{p_{ij}(t, t+dt)}{dt} \\
&= \lim_{dt\to 0} \sum_{j\neq i} \frac{p_{ij}(t, t+dt)}{dt} \\
&= \lim_{dt\to 0} \frac{1 - p_{ii}(t, t+dt)}{dt}
\end{aligned}$$

In the following we will derive two important differential equations, that are commonly used in Markovian multi-state models, the Kolmogorov forward and Kolmogorov backward differential equations, which can be proved using the Chapman-Kolmogorov equations.

Both equations give the relationship between the transition probabilities and the transition intensities in a Markovian multi-state model. As we will see for the case of a three-state model in Example 3.1 and Example 3.2 both differential equations lead to the same results.

3.2.3 Kolmogorov Forward Differential Equations

Theorem 3.6 (Kolmogorov Forward Differential Equations) *The Kolmogorov forward differential equations are given for all states i, j and $0 \leq z \leq t$ with the boundary condition $p_{ij}(z,z) = \delta_{ij}$ by*

$$\begin{aligned}\frac{d}{dt}p_{ij}(z,t) &= \sum_{k\neq j} p_{ik}(z,t)\cdot\mu_{kj}(t) - p_{ij}(z,t)\cdot\mu_j(t)\\ &= \sum_{k\neq j} p_{ik}(z,t)\cdot\mu_{kj}(t) - p_{ij}(z,t)\cdot\sum_{k\neq j}\mu_{jk}(t) \qquad (3.6)\end{aligned}$$

Proof:

We can write the Chapman-Kolmogorov equations as

$$p_{ij}(z,t+dt) = \sum_{k\neq j} p_{ik}(z,t)\cdot p_{kj}(t,t+dt) + p_{ij}(z,t)\cdot p_{jj}(t,t+dt)$$

This gives:

$$\frac{p_{ij}(z,t+dt) - p_{ij}(z,t)}{dt} = \sum_{k\neq j} p_{ik}(z,t)\cdot\frac{p_{kj}(t,t+dt)}{dt} + p_{ij}(z,t)\cdot\frac{p_{jj}(t,t+dt)-1}{dt}$$

Since the transition probabilities are probabilities, the sum over the $p_{ik}(t,t+dt)$ should be equal to one for fixed t and $t+dt$. Summation is done over all possible states k, that can be reached. In other words, the probability of staying in state i and the probabilities of leaving state i to any other state k, has to be equal to one (see Definition 3.10).

$$\sum_{k\in\mathcal{S}} p_{jk}(t,t+dt) = 1 \qquad \Rightarrow \qquad 1 - p_{jj}(t,t+dt) = \sum_{k\neq j} p_{jk}(t,t+dt)$$

Inserting this in above equation we obtain:

$$\frac{p_{ij}(z,t+dt) - p_{ij}(z,t)}{dt} = \sum_{k\neq j} p_{ik}(z,t)\cdot\frac{p_{kj}(t,t+dt)}{dt} - p_{ij}(z,t)\cdot\sum_{k\neq j}\frac{p_{jk}(t,t+dt)}{dt}$$

Letting $dt \to 0$ the final result is:

$$\lim_{dt\to 0}\frac{p_{ij}(z,t+dt) - p_{ij}(z,t)}{dt} = \sum_{k\neq j} p_{ik}(z,t)\cdot\lim_{dt\to 0}\frac{p_{kj}(t,t+dt)}{dt} - p_{ij}(z,t)\cdot\lim_{dt\to 0}\sum_{k\neq j}\frac{p_{jk}(t,t+dt)}{dt}$$

$$\frac{d}{dt}p_{ij}(z,t) = \sum_{k\neq j} p_{ik}(z,t)\cdot\mu_{kj}(t) - p_{ij}(z,t)\cdot\sum_{k\neq j}\mu_{jk}(t)$$

□

Haberman and Pitacco (1999) interpreted this as follows: The transition probabilities start in state i at time z. The left-hand side of the Kolmogorov forward differential equations represents the change in the probability of entering state j over the small interval $[t, t+dt)$, whereas the right-hand side represents the probability of entering state j starting from any state k, $k \neq j$ minus the probability of leaving state j in the small interval $[t, t+dt)$.

Example 3.1

Consider in the following the Illness-Death model (Figure 3.1) with states ”Disease-Free”, ”Diseased” and ”Dead”, labeled 1, 2 and 3, respectively. In the following we are going to write down the Kolmogorov forward differential equations (3.6) and obtain their solutions. First notice that we exclude transitions from state 2 to 1, that is a transition from state ”Diseased” to ”Disease-Free”; thus $\mu_{21} = 0$. The same holds for transitions from state 3 to 1 and 3 to 2 for obvious reasons. Recall that the Kolmogorov forward differential equations have the form

$$\frac{d}{dt}p_{ij}(z,t) = \sum_{k \neq j} p_{ik}(z,t) \cdot \mu_{kj}(t) - p_{ij}(z,t) \cdot \sum_{k \neq j} \mu_{jk}(t)$$

This gives for the conditional probability from state 1 to state $j \in \mathcal{S}$

$$\begin{aligned}
\frac{d}{dt}p_{11}(z,t) &= p_{12}(z,t) \cdot \mu_{21}(t) + p_{13}(z,t) \cdot \mu_{31}(t) - p_{11}(z,t) \cdot (\mu_{12}(t) + \mu_{13}(t)) \\
&= -p_{11}(z,t) \cdot (\mu_{12}(t) + \mu_{13}(t)) \\
\frac{d}{dt}p_{12}(z,t) &= p_{11}(z,t) \cdot \mu_{12}(t) + p_{13}(z,t) \cdot \mu_{32}(t) - p_{12}(z,t) \cdot (\mu_{21}(t) + \mu_{23}(t)) \qquad (3.7) \\
&= p_{11}(z,t) \cdot \mu_{12}(t) - p_{12}(z,t) \cdot \mu_{23}(t) \\
\frac{d}{dt}p_{13}(z,t) &= p_{11}(z,t) \cdot \mu_{13}(t) + p_{12}(z,t) \cdot \mu_{23}(t) - p_{13}(z,t) \cdot (\mu_{31}(t) + \mu_{32}(t)) \\
&= p_{11}(z,t) \cdot \mu_{13}(t) + p_{12}(z,t) \cdot \mu_{23}(t)
\end{aligned}$$

For transitions out of state 2 we have

$$\begin{aligned}
\frac{d}{dt}p_{21}(z,t) &= p_{22}(z,t) \cdot \mu_{21}(t) + p_{23}(z,t) \cdot \mu_{31}(t) - p_{21}(z,t) \cdot (\mu_{12}(t) + \mu_{13}(t)) \\
&= 0 \\
\frac{d}{dt}p_{22}(z,t) &= p_{21}(z,t) \cdot \mu_{12}(t) + p_{23}(z,t) \cdot \mu_{32}(t) - p_{22}(z,t) \cdot (\mu_{21}(t) + \mu_{23}(t)) \\
&= -p_{22}(z,t) \cdot \mu_{23}(t) \\
\frac{d}{dt}p_{23}(z,t) &= p_{21}(z,t) \cdot \mu_{13}(t) + p_{22}(z,t) \cdot \mu_{23}(t) - p_{23}(z,t) \cdot (\mu_{31}(t) + \mu_{32}(t)) \\
&= p_{22}(z,t) \cdot \mu_{23}(t)
\end{aligned}$$

Finally the equations for transitions out of state 3 are the following:

$$\begin{aligned}
\frac{d}{dt}p_{31}(z,t) &= p_{32}(z,t) \cdot \mu_{21}(t) + p_{33}(z,t) \cdot \mu_{31}(t) - p_{31}(z,t) \cdot (\mu_{12}(t) + \mu_{13}(t)) \\
&= 0 \\
\frac{d}{dt}p_{32}(z,t) &= p_{31}(z,t) \cdot \mu_{12}(t) + p_{33}(z,t) \cdot \mu_{32}(t) - p_{32}(z,t) \cdot (\mu_{21}(t) + \mu_{23}(t)) \\
&= 0 \\
\frac{d}{dt}p_{33}(z,t) &= p_{31}(z,t) \cdot \mu_{13}(t) + p_{32}(z,t) \cdot \mu_{23}(t) - p_{33}(z,t) \cdot (\mu_{31}(t) + \mu_{32}(t)) \\
&= 0
\end{aligned}$$

Note that a transition intensity of zero implies a transition probability of zero, as well. Further it is important to note the boundary conditions $p_{ij}(z,z) = \delta_{ij}$, which are in our case

$$\begin{aligned} p_{11}(z,z) &= 1 \\ p_{22}(z,z) &= 1 \\ p_{33}(z,z) &= 1 \end{aligned}$$

All other transition probabilities are zero, e.g. $p_{12}(z,z) = 0$. Using these boundary conditions we can solve the above equations and obtain for p_{11} and similar for p_{22} the following solutions:

$$\begin{aligned} \frac{d}{dt} p_{11}(z,t) &= -p_{11}(z,t) \cdot (\mu_{12}(t) + \mu_{13}(t)) \\ \frac{1}{p_{11}(z,t)} \cdot \frac{d}{dt} p_{11}(z,t) &= -(\mu_{12}(t) + \mu_{13}(t)) \\ \frac{d}{dt} \ln p_{11}(z,t) &= -(\mu_{12}(t) + \mu_{13}(t)) \\ \int_z^t \frac{d}{du} \ln p_{11}(z,u) du &= -\int_z^t (\mu_{12}(u) + \mu_{13}(u))\, du \\ \ln p_{11}(z,t) - \ln \underbrace{p_{11}(z,z)}_{=1} &= -\int_z^t (\mu_{12}(u) + \mu_{13}(u))\, du \\ p_{11}(z,t) &= \exp\left\{ -\int_z^t (\mu_{12}(u) + \mu_{13}(u))\, du \right\} \end{aligned}$$

The solution for p_{22} can be derived in an analogue way:

$$p_{22}(z,t) = \exp\left\{ -\int_z^t \mu_{23}(u) du \right\} \tag{3.8}$$

Using the properties of the transition matrix we get

$$\begin{aligned} p_{13}(z,t) &= 1 - p_{11}(z,t) - p_{12}(z,t) \\ p_{23}(z,t) &= 1 - p_{21}(z,t) - p_{22}(z,t) = 1 - p_{22}(z,t) \end{aligned}$$

which leaves us to solve (3.7) for $p_{12}(z,t)$. This differential equation is a first order linear differential equation and can be solved using the variation of constants method. Given a differential equation of the form $y' = a(x)y + b(x)$ it follows that $y' = a(x)y$ is its homogeneous differential equation with solution

$$y = c \cdot \exp\left\{ \int a(u) du \right\}$$

for $c \in \mathbb{R}$ and the solution of the original differential equation is given by

$$y = \left(\int \left[b(s) \cdot \exp\left\{ -\int a(u) du \right\} \right] ds + c \right) \cdot \exp\left\{ \int a(u) du \right\}$$

Applying these results to equation (3.7) we obtain as its homogeneous differential equation

$$\frac{d}{dt}p_{12}(z,t) = -p_{12}(z,t) \cdot \mu_{23}(t)$$

which has similar to (3.8) a solution of the form

$$p_{12}(z,t) = c \cdot \exp\left\{-\int_z^t \mu_{23}(s)ds\right\}$$

Thus the solution for $p_{12}(z,t)$ is

$$\begin{aligned}
p_{12}(z,t) &= \left(\int_z^t p_{11}(z,u) \cdot \mu_{12}(u) \cdot \exp\left\{\int_z^u \mu_{23}(s)ds\right\} du + c\right) \cdot \exp\left\{-\int_z^t \mu_{23}(s)ds\right\} \\
&= \int_z^t p_{11}(z,u) \cdot \mu_{12}(u) \cdot \exp\left\{\int_z^u \mu_{23}(s)ds\right\} \cdot \exp\left\{-\int_z^t \mu_{23}(s)ds\right\} du + c \cdot \exp\left\{-\int_z^t \mu_{23}(s)ds\right\} \\
&= \int_z^t p_{11}(z,u) \cdot \mu_{12}(u) \exp\left\{-\int_u^t \mu_{23}(s)ds\right\} du + c \cdot \exp\left\{-\int_z^t \mu_{23}(s)ds\right\} \\
&= \int_z^t p_{11}(z,u) \cdot \mu_{12}(u) \cdot p_{22}(u,t)du + c \cdot \exp\left\{-\int_z^t \mu_{23}(s)ds\right\}
\end{aligned}$$

Taking into account that $p_{12}(z,z) = 0$, as required by the boundary condition, we have to choose $c = 0$, and the solution of (3.7) is therefore

$$p_{12}(z,t) = \int_z^t p_{11}(z,u) \cdot \mu_{12}(u) \cdot p_{22}(u,t)du$$

Summarizing the solutions of the three-state model we have

$$\begin{aligned}
p_{11}(z,t) &= \exp\left\{-\int_z^t (\mu_{12}(u) + \mu_{13}(u))\, du\right\} \\
p_{12}(z,t) &= \int_z^t p_{11}(z,u) \cdot \mu_{12}(u) \cdot p_{22}(u,t)du \\
p_{13}(z,t) &= 1 - p_{11}(z,t) - p_{12}(z,t) \\
p_{21}(z,t) &= 0 \\
p_{22}(z,t) &= \exp\left\{-\int_z^t \mu_{23}(u)du\right\} \\
p_{23}(z,t) &= 1 - p_{22}(z,t) \\
p_{31}(z,t) &= 0 \\
p_{32}(z,t) &= 0 \\
p_{33}(z,t) &= 0
\end{aligned}$$

3.2.4 Kolmogorov Backward Differential Equations

Theorem 3.7 (Kolmogorov Backward Differential Equations) *The Kolmogorov backward differential equations are given for all states i, j and $0 \le z \le t$ with the boundary condition $p_{ij}(t,t) = \delta_{ij}$ by*

$$\begin{aligned}
\frac{d}{dz} p_{ij}(z,t) &= p_{ij}(z,t) \cdot \mu_i(z) - \sum_{k \neq i} \mu_{ik}(z) \cdot p_{kj}(z,t) \\
&= p_{ij}(z,t) \cdot \sum_{k \neq i} \mu_{ik}(z) - \sum_{k \neq i} \mu_{ik}(z) \cdot p_{kj}(z,t) \qquad (3.9)
\end{aligned}$$

Proof:

We can write the Chapman-Kolmogorov equations as

$$p_{ij}(z,t) = p_{ii}(z,z+dz) \cdot p_{ij}(z+dz,t) + \sum_{k \neq i} p_{ik}(z,z+dz) \cdot p_{kj}(z+dz,t)$$

Again using Definition 3.10 for $p_{ik}(z,z+dz)$ it follows that

$$\sum_{k \in \mathcal{S}} p_{ik}(z,z+dz) = 1 \qquad \Rightarrow \qquad p_{ii}(z,z+dz) = 1 - \sum_{k \neq i} p_{ik}(z,z+dz)$$

We obtain:

$$p_{ij}(z,t) = \left(1 - \sum_{k \neq i} p_{ik}(z,z+dz)\right) \cdot p_{ij}(z+dz,t) + \sum_{k \neq i} p_{ik}(z,z+dz) \cdot p_{kj}(z+dz,t)$$

This leads to

$$\begin{aligned}
\frac{p_{ij}(z+dz,t) - p_{ij}(z,t)}{dz} &= p_{ij}(z+dz,t) \cdot \sum_{k \neq i} \frac{p_{ik}(z,z+dz)}{dz} \\
&- \sum_{k \neq i} \frac{p_{ik}(z,z+dz)}{dz} \cdot p_{kj}(z+dz,t)
\end{aligned}$$

Letting $dz \to 0$ we finally obtain:

$$\begin{aligned}
\lim_{dz \to 0} \frac{p_{ij}(z+dz,t) - p_{ij}(z,t)}{dz} &= \lim_{dz \to 0} p_{ij}(z+dz,t) \cdot \sum_{k \neq i} \frac{p_{ik}(z,z+dz)}{dz} \\
&- \sum_{k \neq i} \lim_{dz \to 0} \frac{p_{ik}(z,z+dz)}{dz} \cdot p_{kj}(z+dz,t)
\end{aligned}$$

$$\frac{d}{dz} p_{ij}(z,t) = p_{ij}(z,t) \cdot \sum_{k \neq i} \mu_{ik}(z) - \sum_{k \neq i} \mu_{ik}(z) \cdot p_{kj}(z,t)$$

□

Example 3.2

It should be clear that we arrive at the same solutions as in Example 3.1 starting with the set of Kolmogorov backward differential equations (3.9), that are

$$\frac{d}{dz}p_{ij}(z,t) = p_{ij}(z,t) \cdot \sum_{k \neq i} \mu_{ik}(z) - \sum_{k \neq i} p_{kj}(z,t) \cdot \mu_{ik}(z)$$

For our three-state model, in addition to the boundary conditions $p_{ij}(t,t) = \delta_{ij}$, the following equations hold:

$$\begin{aligned}
\frac{d}{dz}p_{11}(z,t) &= p_{11}(z,t) \cdot (\mu_{12}(z) + \mu_{13}(z)) - (p_{21}(z,t) \cdot \mu_{12}(z) + p_{31}(z,t) \cdot \mu_{13}(z)) \\
&= p_{11}(z,t) \cdot (\mu_{12}(z) + \mu_{13}(z)) \\
\frac{d}{dz}p_{12}(z,t) &= p_{12}(z,t) \cdot (\mu_{12}(z) + \mu_{13}(z)) - (p_{22}(z,t)\mu_{12}(z) + p_{32}(z,t) \cdot \mu_{13}(z)) \\
&= p_{12}(z,t) \cdot (\mu_{12}(z) + \mu_{13}(z)) - p_{22}(z,t) \cdot \mu_{12} \\
\frac{d}{dz}p_{13}(z,t) &= p_{13}(z,t) \cdot (\mu_{12}(z) + \mu_{13}(z)) - (p_{23}(z,t) \cdot \mu_{12}(z) + p_{33}(z,t) \cdot \mu_{13}(z)) \\
&= p_{13}(z,t) \cdot (\mu_{12}(z) + \mu_{13}(z)) - (p_{23}(z,t) \cdot \mu_{12}(z) + p_{33}(z,t) \cdot \mu_{13}(z)) \\
\frac{d}{dz}p_{21}(z,t) &= p_{21}(z,t) \cdot (\mu_{21}(z) + \mu_{23}(z)) - (p_{11}(z,t) \cdot \mu_{21}(z) + p_{31}(z,t) \cdot \mu_{23}(z)) \\
&= 0 \\
\frac{d}{dz}p_{22}(z,t) &= p_{22}(z,t) \cdot (\mu_{21}(z) + \mu_{23}(z)) - (p_{12}(z,t) \cdot \mu_{21}(z) + p_{32}(z,t) \cdot \mu_{23}(z)) \\
&= p_{22}(z,t) \cdot \mu_{23} \\
\frac{d}{dz}p_{23}(z,t) &= p_{23}(z,t) \cdot (\mu_{21}(z) + \mu_{23}(z)) - (p_{13}(z,t) \cdot \mu_{21}(z) + p_{33}(z,t) \cdot \mu_{23}(z)) \\
&= p_{23}(z,t) \cdot \mu_{23} - p_{33}(z,t) \cdot \mu_{23} \\
\frac{d}{dz}p_{31}(z,t) &= p_{31}(z,t) \cdot (\mu_{31}(z) + \mu_{32}(z)) - (p_{11}(z,t) \cdot \mu_{31}(z) + p_{21}(z,t) \cdot \mu_{32}(z)) \\
&= 0 \\
\frac{d}{dz}p_{32}(z,t) &= p_{32}(z,t) \cdot (\mu_{31}(z) + \mu_{32}(z)) - (p_{12}(z,t) \cdot \mu_{31}(z) + p_{22}(z,t) \cdot \mu_{32}(z)) \\
&= 0 \\
\frac{d}{dz}p_{33}(z,t) &= p_{33}(z,t) \cdot (\mu_{31}(z) + \mu_{32}(z)) - (p_{13}(z,t) \cdot \mu_{31}(z) + p_{23}(z,t) \cdot \mu_{32}(z)) \\
&= 0
\end{aligned}$$

We obtain again the following solutions for $p_{11}(z,t)$ and $p_{22}(z,t)$:

$$\begin{aligned}
p_{11}(z,t) &= \exp\left\{-\int_z^t (\mu_{12}(u) + \mu_{13}(u))\, du\right\} \\
p_{22}(z,t) &= \exp\left\{-\int_z^t \mu_{23}(u) du\right\}
\end{aligned}$$

Since in the case of $p_{11}(z,t)$

$$\begin{aligned}
\frac{d}{dz}p_{11}(z,t) &= p_{11}(z,t)\cdot(\mu_{12}(z)+\mu_{13}(z))\\
\frac{1}{p_{11}(z,t)}\cdot\frac{d}{dz}p_{11}(z,t) &= \mu_{12}(z)+\mu_{13}(z)\\
\frac{d}{dz}\ln p_{11}(z,t) &= \mu_{12}(z)+\mu_{13}(z)\\
\int_z^t \frac{d}{du}\ln p_{11}(u,t)du &= \int_z^t (\mu_{12}(u)+\mu_{13}(u))\,du\\
\underbrace{\ln p_{11}(t,t)}_{=0}-\ln p_{11}(z,t) &= \int_z^t (\mu_{12}(u)+\mu_{13}(u))\,du\\
p_{11}(z,t) &= \exp\left\{-\int_z^t (\mu_{12}(u)+\mu_{13}(u))\,du\right\}
\end{aligned}$$

Again using the properties of the transition matrix we get

$$\begin{aligned}
p_{13}(z,t) &= 1-p_{11}(z,t)-p_{12}(z,t)\\
p_{23}(z,t) &= 1-p_{21}(z,t)-p_{22}(z,t)=1-p_{22}(z,t)
\end{aligned}$$

Finally, using the variation of constants methods, we write down for $p_{12}(z,t)$ the homogeneous differential equation, that is

$$\frac{d}{dz}p_{12}(z,t)=p_{12}(z,t)\cdot(\mu_{12}(z)+\mu_{13}(z))$$

with solution

$$p_{12}(z,t)=c\cdot\exp\left\{-\int_z^t (\mu_{12}(u)+\mu_{13}(u))\,du\right\}$$

where $c\in\mathbb{R}$. Thus the solution of the inhomogeneous differential equation is

$$\begin{aligned}
p_{12}(z,t) &= \left(\int_z^t p_{22}(u,t)\cdot\mu_{12}(u)\cdot\exp\left\{\int_u^t(\mu_{12}(s)+\mu_{13}(s))\,ds\right\}du+c\right)\cdot\exp\left\{-\int_z^t(\mu_{12}(s)+\mu_{13}(s))\,ds\right\}\\
&= \int_z^t p_{22}(u,t)\cdot\mu_{12}(u)\cdot\exp\left\{\int_u^t(\mu_{12}(s)+\mu_{13}(s))\,ds\right\}\cdot\exp\left\{-\int_z^t(\mu_{12}(s)+\mu_{13}(s))\,ds\right\}du\\
&+ c\cdot\exp\left\{-\int_z^t(\mu_{12}(s)+\mu_{13}(s))\,ds\right\}\\
&= \int_z^t p_{22}(u,t)\cdot\mu_{12}(u)\cdot\exp\left\{-\int_z^u(\mu_{12}(s)+\mu_{13}(s))\,ds\right\}du+c\cdot\exp\left\{-\int_z^t(\mu_{12}(s)+\mu_{13}(s))\,ds\right\}\\
&= \int_z^t p_{22}(u,t)\cdot\mu_{12}(u)\cdot p_{11}(z,u)du+c\cdot\exp\left\{-\int_z^t(\mu_{12}(s)+\mu_{13}(s))\,ds\right\}
\end{aligned}$$

In order to fulfill the boundary condition $p_{12}(z,z)=0$, c is chosen to be equal zero and our solution is

$$p_{12}(z,t)=\int_z^t p_{11}(z,u)\cdot\mu_{12}(u)\cdot p_{22}(u,t)du$$

This shows that we derived the same solutions as in the case of the Kolmogorov forward differential equations as in Example 3.1.

3.2.5 Occupancy Probabilities

Definition 3.13 (Occupancy Probabilities) *The following probabilities are called the occupancy probabilities:*

$$p_{\underline{ii}}(t,u) = P(\, S(z) = i \textit{ for all } z \in [t,u] |\, S(t) = i) \qquad t < u$$

Further we assume that the probability of two or more transitions in the small interval $[t, t+dt)$ *is* $o(dt)$*, where* $\lim_{dt \to 0+} \frac{o(dt)}{dt} = 0$*.*

In contrast to the transition probabilities the process stays in state i during the time spent in the interval $[t,u]$. In case of a transition probability it would be sufficient to be in state i at time t and to be back in state i by time u.

Whereas the Chapman-Kolmogorov equations (3.4) hold for the transition probabilities, we have for the occupancy probabilities the following equations:

Theorem 3.8 (Chapman-Kolmogorov)

$$p_{\underline{ii}}(t,u) = p_{\underline{ii}}(t,w) \cdot p_{\underline{ii}}(w,u) \qquad t \leq w \leq u$$

This means, surviving in state i from time t to u can be seen as a two step process; first surviving in state i from time t to w, and secondly from time w to u.

Proof:

$$\begin{aligned}
p_{\underline{ii}}(t,u) &= P(\, S(z) = i \quad \forall z \in [t,u] |\, S(t) = i) \\
&= P(\, S(z) = i \quad \forall z \in [t,w], S(z) = i \quad \forall z \in [w,u] |\, S(t) = i) \\
&= \frac{P(S(z) = i \quad \forall z \in [t,w], S(z) = i \quad \forall z \in [w,u])}{P(S(t) = i)} \\
&= \frac{P(S(z) = i \quad \forall z \in [t,w])}{P(S(t) = i)} \cdot \frac{P(S(z) = i \quad \forall z \in [t,w], S(z) = i \quad \forall z \in [w,u])}{P(S(z) = i \quad \forall z \in [t,w])} \\
&= P(\, S(z) = i \quad \forall z \in [t,w] |\, S(t) = i) \cdot P(\, S(z) = i \quad \forall z \in [w,u] |\, S(z) = i \quad \forall z \in [t,w]) \\
&= P(\, S(z) = i \quad \forall z \in [t,w] |\, S(t) = i) \cdot P(\, S(z) = i \quad \forall z \in [w,u] |\, S(w) = i) \\
&= p_{\underline{ii}}(t,w) \cdot p_{\underline{ii}}(w,u)
\end{aligned}$$

□

Haberman and Pitacco (1999) use the transition and occupancy probabilities as a tool to formally label the states of a Markov process. For $0 \leq t \leq u$ we have the following:

- State i is a transient state, if $p_{ii}(t, +\infty) = 0$; once we have entered state i we can leave and re-enter state i as many times as we like.
- State i is a strictly transient state, if $p_{ii}(t,u) = p_{\underline{ii}(t,u)} < 1$; once we have entered state i, we can either stay in state i or leave, but not return to this state.
- State i is an absorbing state, if $p_{\underline{ii}}(t,u) = 1$; once we have entered state i, there is no way out of this state. We have to remain in state i, for example the state "Dead" in the Illness-Death model.

Theorem 3.9 (Differential Equations for Occupancy Probabilities) *For the occupancy probabilities the following differential equations holds:*

$$\frac{d}{dt} p_{\underline{ii}}(z,t) = -p_{\underline{ii}}(z,t) \cdot \mu_i(t) = -p_{\underline{ii}}(z,t) \cdot \sum_{k \neq i} \mu_{ik}(t)$$

The solution of this differential equations with boundary condition $p_{\underline{ii}}(z,z) = 1$ is

$$p_{\underline{ii}}(z,t) = \exp\left\{ -\int_z^t \sum_{k \neq i} \mu_{ik}(r) dr \right\}$$

Proof:

For the occupancy probabilities we have, using Theorem 3.8, the following relationship:

$$p_{\underline{ii}}(z,t+dt) = p_{\underline{ii}}(z,t) \cdot p_{\underline{ii}}(t,t+dt).$$

From here it follows that

$$\frac{p_{\underline{ii}}(z,t+dt) - p_{\underline{ii}}(z,t)}{dt} = \frac{p_{\underline{ii}}(z,t)\left(p_{\underline{ii}}(t,t+dt) - 1\right)}{dt}$$

The probability of two or more transitions in the interval $[t, t+dt)$ is by Definition 3.13 $o(dt)$:

$$p_{\underline{ii}}(t,t+dt) + o(dt) = p_{ii}(t,t+dt)$$

Thus we obtain

$$\begin{aligned}
\frac{p_{\underline{ii}}(z,t+dt) - p_{\underline{ii}}(z,t)}{dt} &= -p_{\underline{ii}}(z,t) \cdot \frac{1 - p_{ii}(t,t+dt) + o(dt)}{dt} \\
\lim_{dt \to 0} \frac{p_{\underline{ii}}(z,t+dt) - p_{\underline{ii}}(z,t)}{dt} &= -p_{\underline{ii}}(z,t) \cdot \lim_{dt \to 0} \frac{\sum_{k \neq i} p_{ik}(t,t+dt) + o(dt)}{dt} \\
\frac{d}{dt} p_{\underline{ii}}(z,t) &= -p_{\underline{ii}}(z,t) \cdot \sum_{k \neq i} \mu_{ik}(t)
\end{aligned}$$

To derive a solution of this differential equations we write further

$$\frac{1}{p_{\underline{ii}}(z,t)} \cdot \frac{d}{dt} p_{\underline{ii}}(z,t) = -\sum_{k \neq i} \mu_{ik}(t) \qquad \Rightarrow \qquad \frac{d}{dt} \ln p_{\underline{ii}}(z,t) = -\sum_{k \neq i} \mu_{ik}(t)$$

Integrating from z to t gives

$$\begin{aligned}
\ln p_{\underline{ii}}(z,t) - \ln p_{\underline{ii}}(z,z) &= \int_z^t \frac{d}{dr} \ln p_{\underline{ii}}(z,r) dr = -\int_z^t \sum_{k \neq i} \mu_{ik}(r) dr \\
p_{\underline{ii}}(z,t) &= \exp\left\{ -\int_z^t \sum_{k \neq i} \mu_{ik}(r) dr \right\}
\end{aligned}$$

□

Haberman and Pitacco (1999) note that under mild general conditions on the transition intensities, each set of simultaneous differential equations (Theorem 3.6) or (Theorem 3.9) uniquely determines the transition probabilities $p_{ij}(z,t)$. The transition probabilities satisfy then the definiton of a transition matrix (Definition 3.10) and the Chapman-Kolmogorov equations (3.4). In practice transition intensities are estimated from statistical data and are then used to derive the transition probabilities via the differential equations.

Chapter 4

Non-Parametric Estimators

In the situation of survival data we study the time to failure or death for a homogeneous population of n individuals with hazard-rate function $\lambda(t)$ and cumulative hazard-rate function $\Lambda(t) = \int_0^t \lambda(s)ds$ as introduced in Chapter 2. Another quantity of interest, the survival distribution function, that gives the distribution function of the time until death, i.e. $S(t) = \exp\{-\Lambda(t)\}$ has also been defined in Chapter 2.

In this chapter we are going to derive estimators for these quantities: The cumulative hazard-rate function can be estimated with the Nelson-Aalen estimator and the survival distribution function with the Kaplan-Meier estimator. We are going to show that both estimators are almost unbiased estimators, and derive estimators for their variances, e.g. Greenwood's formula, to calculate confidence intervals.

The Kaplan-Meier estimator can be used in a two-state model with states "Alive" and "Death" to estimate the survival distribution function, the probability of staying in state "Alive" beyond a certain time, and thus give an estimate for the transition matrix of a two-state model, where transitions are only allowed from state "Alive" to state "Death".

If we extend this two-state model to a multi-state model or allow all possible transitions between states, the Aalen-Johansen estimator is needed to produce estimates of the whole transition matrix. Looking at this the other way round, the Aalen-Johansen estimator reduces to the Kaplan-Meier estimator in above mentioned two-state model and the same values are estimated with both estimators.

Therefore beside the Nelson-Aalen estimator and the Kaplan-Meier estimator we define the Aalen-Johansen estimator, show above mentioned relationship between the Kaplan-Meier estimator and the Aalen-Johansen estimator and give an algorithm for computing the Aalen-Johansen estimator in different cases.

Before doing so, recall that censoring and truncation are important issues in survival analysis and have also to be taken in account constructing non-parametric estimators: A sample is subject to right censoring, that is for some individuals we only know that they survived beyond a certain censoring time t. As Borgan (1997) points out, we require independent censoring in the sense that the additional knowledge of censoring before time t does not alter the risk of failure at time t. Individuals who are censored are at the same risk of failure as those who are still alive and uncensored. In other words, the censoring process is independent of the survival time.

To calculate above mentioned estimators we observe a sample of individuals over time and record information on the times transitions occurred, how many individuals transferred from one state to another and how many individuals were contained in each state at this time.

Definition 4.1 *We define the following quantities:*

- $t_1 < t_2 < \ldots$ *are the times, when deaths are observed*
- d_j *is the number of observed individuals that die at* t_j
- r_j *is the number of observed individuals alive and uncensored just prior to* t_j

If, in addition to right-censoring, the sample is subject to left-truncation, we change the definition of r_j, the individuals at risk for dying at t_j, to

- r_j is the number of individuals that entered the study before t_j and are still alive and uncensored just prior to t_j

The quantities d_j and r_j are realizations of the random variables D_j and R_j, respectively. They will be investigated later. Estimates are only defined in the range where data are available. For values beyond the largest observation the estimates might be spurious. Further we denote by $\mathcal{F}_{t-}$ all information available just prior to time t.

In the following section three estimators, namely the Nelson-Aalen for the cumulative hazard-rate function, the Kaplan-Meier for the survival distribution function and the Aalen-Johansen for the matrix of transition probabilities, are introduced. We follow in our presentation mainly Klein and Moeschberger (1997), use the paper by Borgan (1997) and the three-part publication by MacDonald (1996).

4.1 The Nelson-Aalen Estimator for the Cumulative Hazard

4.1.1 Definition

The Nelson-Aalen estimator is a non-parametric estimator for the cumulative hazard-rate function $\Lambda(t) = \int \lambda(s)ds$ using censored survival data. It is zero for $0 \leq t < t_1$ and for $t \geq t_1$ we define it as an increasing right-continuous step-function with increments $\hat{\lambda}_j := d_j/r_j$ at the observed failure times t_j, that is

$$\hat{\Lambda}(t) := \sum_{j:t_j \leq t} \frac{d_j}{r_j}$$

Lemma 4.1 *The variance of the Nelson-Aalen estimator can, according to Borgan (1997), be estimated with the following formula:*

$$\hat{\sigma}^2 := V\hat{ar}\left[\hat{\Lambda}(t)\right] = \sum_{j:t_j \leq t} \frac{(r_j - d_j) \cdot d_j}{(r_j - 1) \cdot r_j^2} \tag{4.1}$$

Proof:

Given all information up to time t_j the random variable D_j is binomially distributed with parameters r_j and λ_j. The quantity λ_j is the true, but unknown hazard-rate function (2.2), that is $\lambda_j = P($ an individual of t_j years dies in the small interval $[t_j, t_{j+1})) = P(T_j = t_j | T_j \geq t_j)$.

$$E[D_j | \mathcal{F}_{t_j-}] = r_j \lambda_j \qquad Var[D_j | \mathcal{F}_{t_j-}] = r_j \lambda_j (1 - \lambda_j) \tag{4.2}$$

By definition, the random variable R_j is, given $\mathcal{F}_{t_j-}$, known at time t_j with value r_j. Further, the random variables D_j/R_j, $j = 1, \ldots, k$, are conditionally uncorrelated:

$$E\left[\frac{D_k}{R_k}\frac{D_l}{R_l}\middle| \mathcal{F}_{t_l-}\right] \overset{k \leq l}{=} \frac{d_k}{r_k} E\left[\frac{D_l}{R_l}\middle| \mathcal{F}_{t_l-}\right] = E\left[\frac{D_k}{R_k}\middle| \mathcal{F}_{t_l-}\right] E\left[\frac{D_l}{R_l}\middle| \mathcal{F}_{t_l-}\right]$$

Using the decomposition of the variance into conditional expectation and conditional variance (see the Springer "Formelsammlung" by Råde and Westergren (1997)) we obtain:

$$Var\left[\hat{\Lambda}(t)\right] = Var\left[E\left[\hat{\Lambda}(t)\middle| \mathcal{F}_{t_j-}\right]\right] + E\left[Var\left[\hat{\Lambda}(t)\middle| \mathcal{F}_{t_j-}\right]\right] \tag{4.3}$$

We calculate the first term:

$$E\left[\hat{\Lambda}(t)\middle| \mathcal{F}_{t_j-}\right] = \sum_{j:t_j \leq t} E\left[\frac{D_j}{R_j}\middle| \mathcal{F}_{t_j-}\right] \overset{(4.2)}{=} \sum_{j:t_j \leq t} \frac{r_j \lambda_j}{r_j} = \sum_{j:t_j \leq t} \lambda_j$$

Since this conditional expectation is constant, its variance is zero and the first term in equation (4.3) vanishes. Using the law of iterated conditional expectation, that is $E[X] = E[E[X|Y]]$, we also obtain that the Nelson-Aalen estimator $\hat{\Lambda}(t)$ is almost unbiased for $\Lambda(t)$:

$$E\left[\hat{\Lambda}(t)\right] = E\left[E\left[\hat{\Lambda}(t)\middle| \mathcal{F}_{t_j-}\right]\right] = E\left[\sum_{j:t_j \leq t} \lambda_j\right] = \sum_{j:t_j \leq t} \lambda_j \approx \int_0^t \lambda(s)ds = \Lambda(t)$$

when the length of the interval $[t_j, t_{j+1})$ converges to zero for all j.

For the second term of (4.3) we get:

$$Var\left[\hat{\Lambda}(t)\middle|\mathcal{F}_{t_j-}\right] = \sum_{j:t_j\leq t} Var\left[\frac{D_j}{R_j}\middle|\mathcal{F}_{t_j-}\right] \overset{(4.2)}{=} \sum_{j:t_j\leq t} \frac{1}{r_j^2} r_j \cdot \lambda_j(1-\lambda_j) = \sum_{j:t_j\leq t} \frac{1}{r_j} \cdot \lambda_j(1-\lambda_j)$$

Again this is a constant quantity and therefore its expectation is the quantity itself. Adding above terms together we obtain:

$$Var\left[\hat{\Lambda}(t)\right] = \sum_{j:t_j\leq t} \frac{1}{r_j} \cdot \lambda_j(1-\lambda_j) \approx \sum_{j:t_j\leq t} \frac{(r_j-d_j)\cdot d_j}{r_j^3} \tag{4.4}$$

This is slightly different to (4.1). But, in order to get an unbiased estimator we use (4.1). Thus it remains to show now that the estimate of the variance, defined in (4.1), is an unbiased estimator, i.e. $E[\hat{\sigma}^2] = E[\widehat{Var}[\hat{\Lambda}(t)]] = Var[\hat{\Lambda}(t)]$:

$$\begin{aligned}
E\left[\hat{\sigma}^2\right] &= E\left[\sum_{j:t_j\leq t} \frac{1}{R_j^2} \cdot \frac{D_j\cdot(R_j-D_j)}{R_j-1}\right] = \sum_{j:t_j\leq t} E\left[E\left[\frac{1}{R_j^2}\cdot\frac{D_j\cdot(R_j-D_j)}{R_j-1}\middle|\mathcal{F}_{t_j-}\right]\right] \\
&= \sum_{j:t_j\leq t} E\left[\frac{1}{r_j^2\cdot(r_j-1)} E\left[D_j\cdot(R_j-D_j)\middle|\mathcal{F}_{t_j-}\right]\right] \\
&= \sum_{j:t_j\leq t} E\left[\frac{1}{r_j^2\cdot(r_j-1)}\left(E\left[D_j\cdot R_j\middle|\mathcal{F}_{t_j-}\right] - E\left[D_j^2\middle|\mathcal{F}_{t_j-}\right]\right)\right] \\
&\overset{(4.2)}{=} \sum_{j:t_j\leq t} E\left[\frac{1}{r_j^2\cdot(r_j-1)}\left(r_j\cdot r_j\lambda_j - r_j\lambda_j(1-\lambda_j) - r_j^2\cdot\lambda_j^2\right)\right] \\
&= \sum_{j:t_j\leq t} E\left[\frac{1}{r_j^2\cdot(r_j-1)}\cdot\left(r_j^2\cdot(\lambda_j-\lambda_j^2) - r_j\lambda_j(1-\lambda_j)\right)\right] \\
&= \sum_{j:t_j\leq t} E\left[\frac{1}{r_j^2\cdot(r_j-1)}\cdot(r_j^2-r_j)\cdot\lambda_j\cdot(1-\lambda_j)\right] \\
&= \sum_{j:t_j\leq t} E\left[\frac{\lambda_j\cdot(1-\lambda_j)}{r_j}\right] = \sum_{j:t_j\leq t} \frac{\lambda_j\cdot(1-\lambda_j)}{r_j} \overset{(4.4)}{=} Var\left[\hat{\Lambda}(t)\right]
\end{aligned}$$

□

This shows that the Nelson-Aalen estimator and its variance estimate are almost unbiased. Further one can show that, for large samples, the Nelson-Aalen estimator is asymptotically normally distributed for fixed t, that is

$$\hat{\Lambda}(t) \sim AN\left(\Lambda(t), \hat{\sigma}^2\right)$$

For details see Andersen, Borgan, Gill, and Keiding (1993) Theorem IV. 1.2. pp. 191.

For survival data with no ties the formula for the variance can be reduced further. If we choose the intervals $[t_j, t_{j+1})$ sufficiently small such that only one jump occurs in $[t_j, t_{j+1})$, the assumption of no ties is reasonable. Then d_j takes only a value of one and we obtain:

$$\hat{\sigma}^2 = \sum_{j:t_j\leq t} \frac{(r_j-d_j)\cdot d_j}{(r_j-1)\cdot r_j^2} = \sum_{j:t_j\leq t} \frac{1}{r_j^2}$$

4.1.2 The Nelson-Aalen Estimator in a Multi-State Model

The above setup can be interpreted as a two-state model with states "Alive" and "Death". Extending this to a multi-state model we can assume that either individuals in each state are subject to more than one type of event, known as competing risk model (see Houggaard (2000) or Borgan (1997)), or events can happen to each individual more than once.

A well known example of the latter is the Illness-Death model, where individuals can recover from disease, and "Dead" is the only absorbing state. More generally we model the live-history of an individual using a Markovian process with a finite number of states. The transition intensity from state g to h is denoted by λ_{gh} for $g \neq h$. Modifying the definition of t_j, d_j and r_j the Nelson-Aalen estimator can be applied to the cumulative intensities in an analogues way. For the transitions from state g to h we define:

- $t_1 < t_2 < \ldots$ are the times, when transitions, regardless of the states involved, are observed
- d_{ghj} is the number of individuals that transfer from state g to h at t_j
- $d_{gj} := \sum_{h \neq g} d_{ghj}$ is the number of individuals that transfer out of state g at t_j.
- r_{gj} is the number of individuals in state g just prior to t_j

The Nelson-Aalen estimator for the cumulative transition intensity $\Lambda_{gh}(t) = \int_0^t \lambda_{gh}(s)ds$ from state g to state h is then given by

$$\hat{\Lambda}_{gh}(t) = \sum_{j:t_j \leq t} \frac{d_{ghj}}{r_{gj}}$$

4.2 The Kaplan-Meier Estimator for the Survival Distribution

4.2.1 Definition

The Kaplan-Meier estimator, also known as the product-limit estimator, is a non-parametric estimator for the survival distribution function $S(t)$ using censored survival data. It is one for $0 \leq t < t_1$ and for $t \geq t_1$ defined as

$$\hat{S}(t) := \prod_{j:t_j \leq t} \left(1 - \hat{\lambda}_j\right) \tag{4.5}$$

where $\hat{\lambda}_j = d_j / r_j$. This is a decreasing right-continuous step-function with jumps only at the death times. The size of the jumps is determined by the number of deaths at t_j as well as the number of censored observations in the interval $[t_{j-1}, t_j)$. In the case of no censoring it reduces to the empirical distribution function.

Lemma 4.2 (Greenwood's formula) *To estimate the variance of the Kaplan-Meier estimator Greenwood's formula (see MacDonald (1996)) can be used, which is defined as*

$$\hat{\sigma}^2 := \hat{S}^2(t) \cdot \sum_{j:t_j \leq t} \frac{d_j}{r_j \cdot (r_j - d_j)} \tag{4.6}$$

Proof:

Given a random variable $X_n \sim AN(\mu, \sigma^2)$ as $n \to \infty$ and some function f, the Delta-method states that under mild regularity conditions on f the transformed random variable $f(X_n)$ is asymptotically normally distributed with mean $E\left[f(X_n)\right] \approx f(\mu)$ and variance $Var\left[f(X_n)\right] \approx f'(\mu^2)\sigma^2$. A sequence of random variables X_n is asymptotically normally distributed with mean μ and variance σ^2, that is $X_n \sim AN(\mu, \sigma^2)$, if for sufficiently large n the quantity $(X_n - \mu)/\sigma^2$ converges in distribution against a standard normally distributed random variable. To understand the idea behind the Delta-Method consider the 1^{st} order Taylor approximation of f:

$$\begin{aligned}
f(X_n) &\approx f(\mu) + f'(\mu) \cdot (X_n - \mu) \\
\Rightarrow \quad E\left[f(X_n)\right] &\approx f(\mu) \quad \text{since} \quad E[X_n] \approx \mu \qquad (4.7) \\
\Rightarrow \quad Var\left[f(X_n)\right] &= E\left[(f(X_n) - E[f(X_n)])^2\right] \approx E\left[f'(\mu)^2 \cdot (X_n - \mu)^2\right] \\
&= f'(\mu)^2 \cdot E\left[(X_n - \mu)^2\right] = f'(\mu)^2 \cdot Var\left[X_n\right] \qquad (4.8)
\end{aligned}$$

Now we use that the random variable D_j given all information until time t_j- is binomially distributed with parameters r_j and λ_j; thus $E[\hat{\lambda}_j \big| \mathcal{F}_{t_{j-}}] = \lambda_j$ and $Var[\hat{\lambda}_j \big| \mathcal{F}_{t_{j-}}] = \lambda_j(1-\lambda_j)/r_j$.

$$\begin{aligned}
E\left[\hat{\lambda}_j\right] &= E\left[E\left[\hat{\lambda}_j \middle| \mathcal{F}_{t_{j-}}\right]\right] = E\left[E\left[\frac{D_j}{R_j} \middle| \mathcal{F}_{t_{j-}}\right]\right] \overset{(4.2)}{=} E\left[\frac{r_j \lambda_j}{r_j}\right] = E\left[\lambda_j\right] = \lambda_j \\
Var\left[\hat{\lambda}_j\right] &= E\left[Var\left[\hat{\lambda}_j \middle| \mathcal{F}_{t_{j-}}\right]\right] + Var\left[E\left[\hat{\lambda}_j \middle| \mathcal{F}_{t_{j-}}\right]\right] = E\left[Var\left[\frac{D_j}{R_j} \middle| \mathcal{F}_{t_{j-}}\right]\right] + \underbrace{Var\left[\lambda_j\right]}_{=0} \\
&\overset{(4.2)}{=} E\left[\frac{\lambda_j \cdot (1 - \lambda_j)}{r_j}\right] = \frac{\lambda_j \cdot (1 - \lambda_j)}{r_j}
\end{aligned}$$

We apply now in the following the Delta-method with the function $f(x) = \ln(1-x)$ to the Kaplan-Meier estimator:

$$\begin{aligned}
\ln(\hat{S}(t)) &= \ln\left(\prod_{j:t_j\le t}(1-\hat{\lambda}_j)\right) = \sum_{j:t_j\le t}\ln(1-\hat{\lambda}_j) && (4.9)\\
E\left[\ln\hat{S}(t)\right] &\overset{(4.9)}{=} E\left[\sum_{j:t_j\le t}\ln(1-\hat{\lambda}_j)\right] = \sum_{j:t_j\le t}E\left[\ln(1-\frac{D_j}{R_j})\right] \\
&\overset{(4.7)}{\approx} \sum_{j:t_j\le t}\ln E\left[(1-\frac{D_j}{R_j})\right] = \sum_{j:t_j\le t}\ln E\left[E\left[(1-\frac{D_j}{R_j})\middle|\mathcal{F}_{t_{j-}}\right]\right] \\
&\overset{(4.2)}{=} \sum_{j:t_j\le t}\ln E\left[(1-\frac{r_j\lambda_j}{r_j})\right] = \sum_{j:t_j\le t}\ln(1-\lambda_j) && (4.10)\\
Var\left[\ln\left(1-\hat{\lambda}_j\right)\right] &\overset{(4.8)}{\approx} \left(\frac{-1}{1-\lambda_j}\right)^2\cdot Var[1-\hat{\lambda}_j] = \left(\frac{1}{1-\lambda_j}\right)^2\cdot\frac{\lambda_j\cdot(1-\lambda_j)}{r_j} \\
&= \frac{\lambda_j}{(1-\lambda_j)\cdot r_j} \approx \frac{\hat{\lambda}_j}{(1-\hat{\lambda}_j)\cdot r_j} = \frac{d_j}{(r_j-d_j)\cdot r_j} \\
Var\left[\ln(\hat{S}(t))\right] &\overset{(4.9)}{=} Var\left[\sum_{j:t_j\le t}\ln(1-\hat{\lambda}_j)\right] = \sum_{j:t_j\le t}Var\left[\ln(1-\hat{\lambda}_j)\right] \\
&\approx \sum_{j:t_j\le t}\frac{d_j}{(r_j-d_j)\cdot r_j} && (4.11)
\end{aligned}$$

To obtain the asymptotic expectation and variance of $\hat{S}(t)$ we apply again the Delta-method with the function $\exp\{x\}$ to the above derived results, yielding to:

$$\begin{aligned}
E\left[\hat{S}(t)\right] &= E\left[\exp\left\{\ln\hat{S}(t)\right\}\right] \overset{(4.7),(4.10)}{\approx} \exp\left\{\sum_{j:t_j\le t}\ln(1-\lambda_j)\right\} \\
&= \prod_{j:t_j\le t}(1-\lambda_j) = S(t) \\
Var\left[\hat{S}(t)\right] &= Var\left[\exp\left\{\ln(\hat{S}(t))\right\}\right] \overset{(4.8)}{\approx} \hat{S}(t)^2\cdot Var\left[\ln\hat{S}(t)\right] \\
&\overset{(4.11)}{\approx} \hat{S}(t)^2\cdot\sum_{j:t_j\le t}\frac{d_j}{r_j\cdot(r_j-d_j)} && (4.12)
\end{aligned}$$

□

Thus we derived Greenwood's formula (4.6) and even proved that the Kaplan-Meier estimator is an almost unbiased estimator.

Lemma 4.3 *The variance estimate (4.6) reduces in the case of no censoring to the binomial variance, thus we obtain:*

$$\hat{\sigma}^2 = \frac{\hat{S}(t) \cdot (1 - \hat{S}(t))}{n} \tag{4.13}$$

Proof:

In the case of no-censoring the number of individuals at risk at time t_{j+1} is equal to the number of individuals at risk at time t_j minus the individuals that died at time t_j, that is $r_{j+1} = r_j - d_j$. For $t_k \leq t < t_{k+1}$ we have:

$$\begin{aligned}
\sum_{j:t_j \leq t} \frac{d_j}{(r_j - d_j) \cdot r_j} &= \sum_{j:t_j \leq t} \frac{r_j - r_{j+1}}{r_{j+1} \cdot r_j} \\
&= \sum_{j:t_j \leq t} \left(\frac{1}{r_{j+1}} - \frac{1}{r_j} \right) \\
&= \frac{1}{r_{k+1}} - \frac{1}{r_1} = \frac{r_1 - r_{k+1}}{r_1 \cdot r_{k+1}} \\
&= \frac{1 - \frac{r_{k+1}}{r_1}}{r_1 \cdot \frac{r_{k+1}}{r_1}} = \frac{1 - \hat{S}(t)}{r_1 \cdot \hat{S}(t)}
\end{aligned}$$

Thus it follows:

$$\begin{aligned}
Var[\hat{S}(t)] &\overset{(4.12)}{\approx} \hat{S}(t)^2 \cdot \sum_{j:t_j \leq t} \frac{d_j}{(r_j - d_j) \cdot r_j} \\
&= \hat{S}(t)^2 \cdot \frac{1 - \hat{S}(t)}{r_1 \cdot \hat{S}(t)} \\
&= \frac{\hat{S}(t) \cdot (1 - \hat{S}(t))}{r_1}
\end{aligned}$$

□

This is nothing else than a binomial variance, as in the case of no censoring r_1 is equal to n and the relation $\hat{S}(t) = r_{k+1}/r_1$ holds for all k. If $S(t)$ is the true survival probability at time t, the distribution of survivors to time t, denoted by r_{k+1}, is binomially distributed with parameters r_1 and $S(t)$. Again, for a large sample $\hat{S}(t)$ is approximately normally distributed.

In the case of left-truncated observations we use above mentioned modified r_j. Using this quantity all estimation procedures are still applicable, but the result needs a different interpretation, as probabilities have to be replaced by conditional probabilities.

The Kaplan-Meier estimator is now the probability of survival beyond t, conditional on survival to the smallest entry time, denoted by $C^{(l)}$, that is $S(t)/S(C^{(l)})$. In a similar way the Nelson-Aalen estimator estimates the integral of the hazard-rate function over the interval $C^{(l)}$ to t, that is $\Lambda(t) = \int_{C^{(l)}}^{t} \lambda(s) ds$.

4.2.2 The Kaplan-Meier Estimator as Maximum Likelihood Estimator

In the Case of No Censoring

In this section we want to derive the Kaplan-Meier estimator as the maximum likelihood estimator of the survival distribution function, following MacDonald (1996). We assume that n lives under observation die at times $t_1, \ldots, t_k$ where $k \leq n$ so that multiple deaths are possible. The quantities d_j and r_j are defined as in Definition 4.1 and depend on each other through the relation $r_{j+1} = r_j - d_j$. We assume that $r_{k+1} = 0$, which means that all n individuals die eventually. Consequently $n = \sum_{j=1}^{k} d_j$.

Let denote by F the discrete lifetime distribution with jumps at $t_1, \ldots, t_k$, then the probability of dying at time t_j is given by $F(t_j) - F(t_{j-})$, where the random variable T denotes again the future lifetime of an individual. Thus the likelihood can be written as

$$L(F) = \prod_{j=1}^{k} P\left(T = t_j\right)^{d_j} = \prod_{j=1}^{k} P\left(T = t_j | T \geq t_j\right)^{d_j} \cdot P\left(T \geq t_j\right)^{d_j} \tag{4.14}$$

We want to find a distribution function $\hat{F}$ that estimates the true, but unknown distribution function F and maximizes L, i.e. $\hat{F}$ is the maximum likelihood estimate of F. We define the discrete hazard-rate at time t_j as

$$\lambda_j := P\left(T = t_j | T \geq t_j\right) = \frac{P\left(T = t_j\right)}{P\left(T \geq t_j\right)} = \frac{F(t_j) - F(t_{j-})}{1 - F(t_{j-})} \tag{4.15}$$

Thus we can calculate with (4.15) one minus the discrete hazard, i.e.

$$1 - \lambda_j = \frac{P\left(T \geq t_j\right) - P\left(T = t_j\right)}{P\left(T \geq t_j\right)} = \frac{P\left(T > t_j\right)}{P\left(T > t_{j-1}\right)} \tag{4.16}$$

where we interpret t_0 as zero and define $P(T > t_0) = 1$. Multiplying the factors from (4.16) for $j = 1, \ldots, r \leq k$ we get:

$$\prod_{j=1}^{r} (1 - \lambda_j) = (1 - \lambda_1) \cdot \ \ldots \ \cdot (1 - \lambda_r) = \frac{P\left(T > t_1\right)}{\underbrace{P\left(T > t_0\right)}_{=1}} \cdot \ \ldots \ \cdot \frac{P\left(T > t_r\right)}{P\left(T > t_{r-1}\right)} = P\left(T > t_r\right)$$

From here it follows:

$$F(t_r) = P\left(T \leq t_r\right) = 1 - P\left(T > t_r\right) = 1 - \prod_{j=1}^{r} (1 - \lambda_j) \tag{4.17}$$

More generally we get for $t_j \leq t \leq t_{j+1}$ the following:

$$F(t) = 1 - \prod_{j: t_j \leq t} (1 - \lambda_j)$$

Since all lives die, we have $r_k = d_k$. Thus $(1 - \lambda_k)^{r_k - d_k} = 1$ and

$$r_i = \sum_{j=i}^{k} d_j \qquad \Rightarrow \qquad r_i - d_i = \sum_{j=i+1}^{k} d_j \tag{4.18}$$

Using the property (4.18) for $1 - F(t_{j-})$ we can write

$$\begin{aligned}
\prod_{j=1}^{k} (1 - F(t_{j-}))^{d_j} &= \prod_{j=1}^{k} (1 - F(t_{j-1}))^{d_j} \overset{(4.17)}{=} \prod_{j=1}^{k} \prod_{i=1}^{j-1} (1 - \lambda_i)^{d_j} \\
&= \prod_{i=1}^{k-1} \prod_{j=i+1}^{k} (1 - \lambda_i)^{d_j} = \prod_{i=1}^{k-1} (1 - \lambda_i)^{\sum_{j=i+1}^{k} d_j} \\
&\overset{(4.18)}{=} \prod_{i=1}^{k-1} (1 - \lambda_i)^{r_i - d_i} = \prod_{i=1}^{k} (1 - \lambda_i)^{r_i - d_i} \qquad (4.19)
\end{aligned}$$

Using (4.15) and (4.19) we can calculate the likelihood function (4.14) as

$$\begin{aligned}
L(F) &= \prod_{j=1}^{k} \left(\frac{F(t_j) - F(t_{j-})}{1 - F(t_{j-})} \right)^{d_j} \cdot (1 - F(t_{j-}))^{d_j} \\
&\overset{(4.15)}{=} \prod_{j=1}^{k} \lambda_j^{d_j} \cdot (1 - F(t_{j-}))^{d_j} \\
&\overset{(4.19)}{=} \prod_{j=1}^{k} \lambda_j^{d_j} \cdot (1 - \lambda_j)^{r_j - d_j}
\end{aligned}$$

Setting the derivative of the log-likelihood function with respect to λ_j zero we obtain:

$$\begin{aligned}
\frac{\partial \ln L}{\partial \lambda_j} &= \frac{\partial}{\partial \lambda_j} \left(\sum_{j=1}^{k} d_j \cdot \ln \lambda_j + (r_j - d_j) \cdot \ln (1 - \lambda_j) \right) \\
&= \frac{d_j}{\lambda_j} - \frac{r_j - d_j}{1 - \lambda_j} = 0
\end{aligned}$$

Thus the maximum likelihood estimate for λ_j, denoted by $\hat{\lambda}_j$, is given by

$$\hat{\lambda}_j := \frac{d_j}{r_j}$$

Therefore the maximum likelihood estimate $\hat{S}(t)$ of the survival function $S(t)$ is given by

$$\hat{S}(t) = 1 - \hat{F}(t) = 1 - \left(1 - \prod_{j:t_j \leq t} \left(1 - \hat{\lambda}_j \right) \right) = \prod_{j:t_j \leq t} \left(1 - \frac{d_j}{r_j} \right)$$

This is nothing else than the Kaplan-Meier estimator as defined in (4.5).

In the Case of Censoring

We want to allow now for censoring: As before we have n independent lives that die at times $t_1 < \ldots < t_k$. Again d_j is the number of deaths at time t_j, but in contrast to the above, $\sum_{j=1}^{k} d_j = m \neq n$. Let c_j be the number of censored lives between times t_j and t_{j+1} at the censoring times $t_{j_1}, \ldots, t_{j_{c_j}}$. In total we have $n - m$ censored observations. Therefore we get the likelihood from (4.14) as

$$L(F) = \prod_{j=1}^{k} (F(t_j) - F(t_{j-}))^{d_j} \cdot \prod_{j=0}^{k} \prod_{l=1}^{c_j} (1 - F(t_{jl}))$$

Since $F(t)$ is a distribution function, it is is non-decreasing. Therefore $F(t_j) \leq F(t_{jl})$ and we will bound each factor $(1 - F(t_{jl}))$, if we take the smaller value $F(t_j)$ for all $l = 1, \ldots, c_j$, that is $F(t_{jl}) = F(t_j)$. We also use $1 - F(t_0) = 1$ and $r_j = \sum_{i=j}^{k} d_i + \sum_{i=j}^{k} c_i$.

$$\begin{aligned} L(F) &= \prod_{j=1}^{k} \left(\frac{F(t_j) - F(t_{j-})}{1 - F(t_{j-})} \right)^{d_j} \cdot \prod_{j=0}^{k} \left((1 - F(t_{j-}))^{d_j} \cdot \prod_{l=1}^{c_j} (1 - F(t_{jl})) \right) \\ &\leq \prod_{j=1}^{k} \lambda_j^{d_j} \cdot \prod_{j=0}^{k} (1 - F(t_{j-}))^{d_j} \cdot (1 - F(t_j))^{c_j} \\ &= \prod_{j=1}^{k} \lambda_j^{d_j} \cdot (1 - F(t_{j-}))^{\sum_{i=j+1}^{k} d_i} \cdot (1 - F(t_j))^{\sum_{i=j}^{k} c_i} \\ &= \prod_{j=1}^{k} \lambda_j^{d_j} \cdot (1 - \lambda_j)^{r_j - d_j} \end{aligned}$$

Thus the likelihood for censored observations is bounded by the one derived for no-censored observations. Thus we obtain the same estimator $\hat{\lambda}_j$ for λ_j in both cases. The maximum likelihood estimate for $S(t)$ is therefore the Kaplan-Meier estimator (4.5), as well:

$$\hat{S}(t) = \prod_{j: t_j \leq t} \left(1 - \frac{d_j}{r_j} \right)$$

4.2.3 Relationship between Kaplan-Meier and Nelson-Aalen Estimator

From survival analysis we know that between the survival distribution function and the cumulative hazard-rate function the following relation holds:

$$S(t) = \exp\{-\Lambda(t)\}$$

Given the Nelson-Aalen estimator it might be tempting to use this relationship as an estimate for the survival distribution function. Using the approximation $\exp\{x\} \approx 1 + x$ for small x one could approximate the Kaplan-Meier estimator setting $x := -\hat{\lambda}_j$:

$$\hat{S}(t) = \prod_{j:t_j \le t} (1 - \hat{\lambda}_j) \approx \prod_{j:t_j \le t} \exp\left\{-\hat{\lambda}_j\right\} = \exp\left\{-\sum_{j:t_j \le t} \hat{\lambda}_j\right\} = \exp\left\{-\hat{\Lambda}(t)\right\} := \hat{S}_{NA}(t)$$

One might receive similar behavior for $\hat{S}_{NA}(t)$ and $\hat{S}(t)$, but it is important (see Borgan (1997)) that $\hat{S}_{NA}(t)$ is only an approximation of the Kaplan-Meier estimator $\hat{S}(t)$. It is not the canonical estimator of the survival distribution function. The approximation is justified in the sense that the quantity $\hat{\lambda}_j$ is a small number for large samples and sufficient small chosen intervals $[t, t+dt)$. But, especially in the lower and upper age ranges, where observations are rare, the result might be spurious. Further it is worth mentioning that above relation is only valid for the continuous case. Therefore we introduce a notation that unifies both, the continuous and discrete-time case:

$$\Lambda(t) = -\int_0^t \frac{dS(u)}{S(u-)} \tag{4.20}$$

where $S(u-)$ is the left-hand limit of the survival distribution function.

For an absolute continuous distribution this becomes

$$\Lambda(t) = -\int_0^t \frac{dS(u))}{S(u-)} = -\int_0^t d\ln S(u) = -\ln S(t) = \int_0^t \lambda(u)du$$

For a discrete distribution we have:

$$\Lambda(t) = -\int_0^t \frac{dS(u))}{S(u-)} = \sum_{j:t_j \le t} \lambda_j$$

The above statements follow from the definition of the hazard-rate function and the relationship $F(t) = 1 - S(t)$, that gives us $dF(t) = -dS(t)$.

$$\begin{aligned}
\lambda(t) &:= \lim_{dt \to 0^+} \frac{P(t < T \le t + dt | T > t)}{dt} \\
&= \lim_{dt \to 0^+} \frac{F(t + dt) - F(t)}{dt \quad (1 - F(t))} \\
&= \frac{dF(t)}{1 - F(t)} \\
&= \frac{-dS(t)}{S(t)}
\end{aligned}$$

Using a product-integral, that is the limit of approximating finite products in a similar manner as the ordinary integral is defined as the limit of approximating finite sums, for details see Gill (2001), we can express the survival distribution function as

$$S(t) = \prod_{0 \leq t} (1 - d\Lambda(u)) \tag{4.21}$$

The actual definition of the product-integral is

$$S(t) = \prod_{0 \leq t} (1 - d\Lambda(u)) := \lim_{max|t_i - t_{i+1}| \to 0} \prod_{0 \leq t} (1 - (\Lambda(t_i) - \Lambda(t_{i-1})))$$

where the limit is taken over a sequence of finer and finer partitions $0 < t_0 < t_1 < \ldots < t_k = t$ of the time interval $[0, t]$.

The hazard $d\Lambda(u)$ can be interpreted as the probability of dying in the interval $[u, u + du)$ given survival to time u. Consequently $1 - d\Lambda(u)$ is the probability of surviving the interval $[u, u + du)$ given survival to time u. Multiplying over all small intervals $[u, u+du)$, that make up the interval $[0, t)$, is then the unconditional probability of surviving up to time t. Therefore equation (4.21) is for $t_k \leq t < t_{k+1}$ the limiting case of

$$S(t) = P(T > t) = \prod_{i=1}^{k} P(T > t_i | T > t_{i-1}) = \prod_{i=1}^{k} (1 - P(T \leq t_i | T > t_{i-1}))$$

For a continuous distribution expression (4.21) is equal to

$$S(t) = \prod_{0 \leq t} (1 - d\Lambda(u)) = \exp\{-\Lambda(t)\}$$

For a discrete distribution we obtain:

$$S(t) = \prod_{0 \leq t} (1 - d\Lambda(u)) = \prod_{j:t_j \leq t} (1 - \lambda_j)$$

The Nelson-Aalen estimator for the cumulative hazard-rate function was derived as

$$\hat{\Lambda}(t) = \sum_{j:t_j \leq t} \frac{d_j}{r_j}$$

As already noted, this is an increasing right-continuous step-functions, which means that all probability mass concentrates at the observed times of deaths t_j, and with discrete hazard-rate $\hat{\lambda}_j = d_j / r_j$. Using (4.21) we obtain:

$$\hat{S}(t) = \prod_{0 \leq t} (1 - d\hat{\Lambda}(u)) = \prod_{j:t_j \leq t} (1 - \frac{d_j}{r_j})$$

which is the Kaplan-Meier estimator. Comparing this with equation (4.20) and (4.21) we can see that both estimators are related like the survival function and the cumulative hazard-rate function. Therefore they are the canonical non-parametric estimators.

Example 4.1

Suppose that we have a sample of 16 individuals from the population under study. Out of these 16 individuals 12 die at times t_j (0.75, 0.91, 1.32, 1.70, 2.15, 2.76, 2.88, 2.98, 4.51, 6.23, 8.57, 10.23) and we have four censored lifetimes c_j, namely 0.5, 0.8, 1.70 and 2.08. The following table summarizes the data and gives the Kaplan-Meier estimate for the survival function:

j	$[t_j, t_{j+1})$	n_j	d_j	$(n_j - d_j)/n_j$	$\hat{S}(t_j)$	$\hat{F}(t_j)$
0	[0, 0.75)	16	0	1	1	0
1	[0.75, 0.91)	15	1	0.9333	0.9333	0.0667
2	[0.91, 1.32)	13	1	0.9231	0.8615	0.1385
3	[1.32, 1.70)	12	1	0.9167	0.7897	0.2103
4	[1.70, 2.15)	11	1	0.9091	0.7179	0.2821
5	[2.15, 2.76)	8	1	0.8750	0.6282	0.3718
6	[2.76, 2.88)	7	1	0.8571	0.5385	0.4615
7	[2.88, 2.98)	6	1	0.8333	0.4487	0.5513
8	[2.98, 4.51)	5	1	0.8000	0.3590	0.5410
9	[4.51, 6.23)	4	1	0.7500	0.2692	0.7308
10	[6.23, 8.57)	3	1	0.6667	0.1795	0.8205
11	[8.57, 10.23)	2	1	0.5000	0.0897	0.9103
12	[10.23, ω)	1	1	0.0000	0.0000	1.0000

Table 4.1: Development of the Kaplan-Meier Estimate of the Survival Function for Example 4.1

Figure 4.1 gives the Kaplan-Meier estimate of the survival distribution together with the confidence intervals (CI) using Greenwood's formula (4.6):

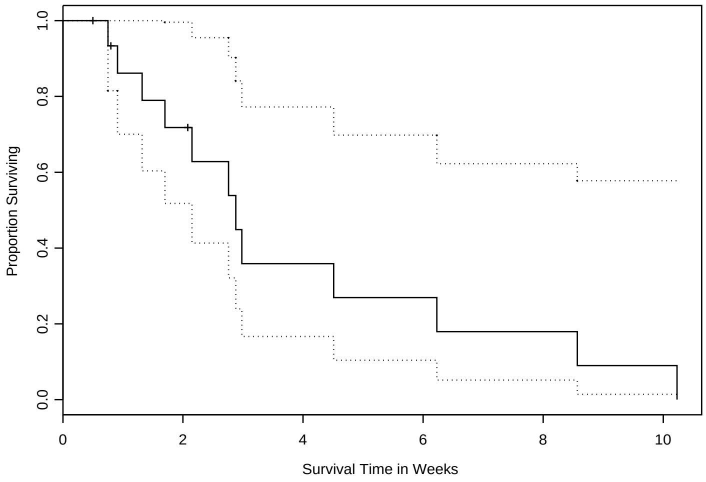

Figure 4.1: Kaplan Meier Estimate of the Survival Function together with CIs for Example 4.1

4.3 The Aalen-Johansen Estimator for the Transition Matrix

4.3.1 Definition

The Aalen-Johansen Estimator is a non-parametric estimator for the matrix of transition probabilities, the so-called transition matrix, of a given Markovian process from censored survival data. We follow in our presentation Borgan (1997), a more detailed description can be found in Andersen, Borgan, Gill, and Keiding (1993).

Assume that the life-history of an individual is described by a Markovian process with a finite number of states $S = \{1, \dots, K\}$. While the hazard-rate function $\lambda(t)$ describes the instantaneous risk of death, the transition intensities describe the instantaneous risk of transition between states, i.e. $\alpha_{gh}(t)$ is the transition intensity from state g to h for $g \neq h$. Similar to the quantity $\lambda(t)dt$, $\alpha_{gh}(t)dt$ describes the probability that an individual in state g just prior to time t will make a transition to state h in the small interval $[t, t+dt)$. The transition probability is denoted by $p_{gh}(s,t)$ and describes the probability that an individual in state g at time s to transfer into state h by t. The $K \times K$ matrix $\mathbf{P}(s,t)$ summarizes the transition probabilities of above given Markovian process.

We extend the definitions of the quantities d_j and r_j (see Definition 4.1) to a multi-state model as already done for the Nelson-Aalen estimator in Section 4.1.2, and define additionally the new quantity d_{gj}, the number of transitions out of state g at time t_j. Rewriting Definition 4.1 we obtain the following new definition:

Definition 4.2 *We define the following quantities:*

- $t_1 < t_2 < \dots$ *are the times, when transitions are observed*
- d_{ghj} *is the number of individuals that transfer from state* g *to* h *at time* t_j
- $d_{gj} = \sum_{h \neq g} d_{ghj}$ *is the number of transitions out of state* g *at* t_j
- r_{gj} *is the number of individuals in state* g *just prior to* t_j

Then the Aalen-Johansen estimator for the transition matrix $\mathbf{P}(s,t)$ takes the following form, if only one transition takes place at the same time t_j. This assumption of no ties is relaxed later.

$$\hat{\mathbf{P}}(s,t) = \prod_{j: s < t_j \leq t} (\mathbf{I} + \hat{\boldsymbol{\alpha}}_j) \tag{4.22}$$

where $\hat{\boldsymbol{\alpha}}_j$ is a $K \times K$ matrix with entry (g,h) equal to $\hat{\alpha}_{ghj} = d_{ghj}/r_{gj}$, entry (g,g) equal to $\hat{\alpha}_{ggj} = -d_{gj}/r_{gj}$ and all other entries are zero. $\mathbf{I}$ is the identity matrix. The product is taken in the order of increasing t_j's. From this definition it becomes clear that the Aalen-Johansen estimator is a product of stochastic matrices.

One can show using the theory of counting-processes, more precisely Duhamel's equation, that the Aalen-Johansen estimator is almost unbiased under similar conditions as the Nelson-Aalen and Kaplan-Meier estimator, it is even unbiased if the probability that $r_{gj} = 0$ is equal to zero for all times t_j and states g. Since this is, given a large sample of observations, usually the case we refer in the following to the Aalen-Johansen estimator as an unbiased estimator. The necessary theory and a proof of this result can be found in the book by Andersen, Borgan, Gill, and Keiding (1993) on pp. 287.

4.3.2 In Case of a Two-State Model

The Aalen-Johansen estimator can be seen as a matrix version of the Kaplan-Meier estimator. This can be understood, if one takes a two-state model with states "Alive" and "Death". The Aalen-Johansen estimator reduces then to a 2×2 matrix. The probability to be at time t still alive, $P_{11}(0,t)$, is the survival probability. The Aalen-Johansen estimator gives the same estimate as obtained by the Kaplan-Meier estimator.

Assume we observe n individuals in the interval $[0,T]$. The two states are "Alive" and "Death", where a transition is only possible from state "Alive" to "Death" and denoted by $\alpha_{01}(t)$.

- $t_1 < t_2 < \ldots$ are the times, when deaths are observed
- d_j is the number of individuals that die at t_j
- r_j is the number of individuals alive and uncensored at t_j

The Aalen-Johnsen estimator reduces then to

$$\begin{aligned}
\mathbf{P}(\hat{0},t) &= \prod_{j:t_j \le t} (\mathbf{I} + \hat{\boldsymbol{\alpha}}_j) = \prod_{j:t_j \le t} \left[\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \begin{pmatrix} -\frac{d_j}{r_j} & \frac{d_j}{r_j} \\ 0 & 0 \end{pmatrix} \right] = \prod_{j:t_j \le t} \begin{pmatrix} 1 - \frac{d_j}{r_j} & \frac{d_j}{r_j} \\ 0 & 1 \end{pmatrix} \\
&= \begin{pmatrix} \prod_{j:t_j \le t} (1 - \frac{d_j}{r_j}) & 1 - \prod_{j:t_j \le t} (1 - \frac{d_j}{r_j}) \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} \hat{S}(t) & 1 - \hat{S}(t) \\ 0 & 1 \end{pmatrix}
\end{aligned}$$

Example 4.2

Now we apply the Aalen-Johansen estimator to the data from Example 4.1. We should receive the same results, modeling survival as a two-state model, as obtained from the Kaplan-Meier estimator in Section 4.2:

$$\hat{\mathbf{P}}(0,t_1) = (\mathbf{I} + \hat{\boldsymbol{\alpha}}_1) = \begin{pmatrix} \frac{14}{15} & \frac{1}{15} \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0.9333 & 0.0667 \\ 0 & 1 \end{pmatrix}$$

$$\hat{\mathbf{P}}(0,t_2) = \hat{\mathbf{P}}(0,t_1)\,(\mathbf{I} + \hat{\boldsymbol{\alpha}}_2) = \begin{pmatrix} \frac{14}{15} & \frac{1}{15} \\ 0 & 1 \end{pmatrix} \times \begin{pmatrix} \frac{12}{13} & \frac{1}{13} \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} \frac{56}{65} & \frac{9}{65} \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0.8615 & 0.1385 \\ 0 & 1 \end{pmatrix}$$

$$\hat{\mathbf{P}}(0,t_3) = \hat{\mathbf{P}}(0,t_2)\,(\mathbf{I} + \hat{\boldsymbol{\alpha}}_3) = \begin{pmatrix} \frac{56}{65} & \frac{9}{65} \\ 0 & 1 \end{pmatrix} \times \begin{pmatrix} \frac{11}{12} & \frac{1}{12} \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} \frac{154}{195} & \frac{41}{195} \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0.7897 & 0.2103 \\ 0 & 1 \end{pmatrix}$$

$$\hat{\mathbf{P}}(0,t_4) = \hat{\mathbf{P}}(0,t_3)\,(\mathbf{I} + \hat{\boldsymbol{\alpha}}_4) = \begin{pmatrix} \frac{254}{195} & \frac{41}{195} \\ 0 & 1 \end{pmatrix} \times \begin{pmatrix} \frac{10}{11} & \frac{1}{11} \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} \frac{28}{39} & \frac{11}{39} \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0.7179 & 0.2821 \\ 0 & 1 \end{pmatrix}$$

etc.

4.3.3 In Case of a Multi-State Model

In the following we extend the two-state model to a multi-state model, but still assume that only one transition takes place at a time. This assumption however, is relaxed later.

$$
\begin{aligned}
\hat{\mathbf{P}}(0,t) &= \prod_{j:t_j\leq t} (\mathbf{I} + \hat{\boldsymbol{\alpha}}_j) \\
&= \prod_{j:t_j\leq t} \left[\begin{pmatrix} 1 & \dots & 0 \\ & \ddots & \\ 0 & \dots & 1 \end{pmatrix} + \begin{pmatrix} 0 & \dots & \dots & \dots & \dots & \dots & 0 \\ 0 & \ddots & \dots & \dots & \dots & \dots & 0 \\ 0 & \dots & -\frac{d_{ghj}}{r_{gj}} & \dots & \frac{d_{ghj}}{r_{gj}} & \dots & 0 \\ 0 & \dots & \dots & \ddots & \dots & \dots & 0 \\ 0 & \dots & \dots & \dots & \ddots & \dots & 0 \\ 0 & \dots & \dots & \dots & \dots & \ddots & 0 \\ 0 & \dots & \dots & \dots & \dots & \dots & 0 \end{pmatrix} \right] \\
&= \prod_{j:t_j\leq t} \begin{pmatrix} 1 & \dots & \dots & \dots & \dots & \dots & 0 \\ 0 & 1 & \dots & \dots & \dots & \dots & 0 \\ 0 & \dots & 1-\frac{d_{ghj}}{r_{gj}} & \dots & \frac{d_{ghj}}{r_{gj}} & \dots & 0 \\ 0 & \dots & \dots & 1 & \dots & \dots & 0 \\ 0 & \dots & \dots & \dots & 1 & \dots & 0 \\ 0 & \dots & \dots & \dots & \dots & 1 & 0 \\ 0 & \dots & \dots & \dots & \dots & \dots & 1 \end{pmatrix}
\end{aligned}
$$

We can notice here that the off-diagonal element of this matrix (g,h) is equal to the Nelson-Aalen estimator in the small interval $[t_j; t_{j+1})$, i.e. $d\Lambda_{gh}(t)$, whereas the diagonal element (g,g) is equal to $1 - d\Lambda_{gh}(t)$.

4.3.4 Algorithm for Computing the Aalen-Johansen Estimator (Forwards)

Aalen and Johansen (1978) suggested an algorithm for computing their estimator at time t_{i+1} based on the result at time t_i, where no transitions occur between those two time-points. We denote the Aalen-Johansen estimator at time t_{i+1} by $\hat{\mathbf{P}}(s, t_{i+1})$ and at time t_i by $\hat{\mathbf{P}}(s, t_i)$. If a transition occurs at time t_{i+1} from state g to state h the following calculations have to be performed to obtain $\hat{\mathbf{P}}(s, t_{i+1})$:

- The $g'th$ column of $\hat{\mathbf{P}}(s, t_i)$ is multiplied by $\left(1 - \frac{1}{r_{gj}}\right)$, where r_{gj} are the individuals at risk at time $t_j = t_{i+1}$
- The $g'th$ column of $\hat{\mathbf{P}}(s, t_i)$ is multiplied by $\frac{1}{r_{gj}}$ and the result is added to the $h'th$ column
- All other columns stay unchanged

The resulting matrix is the Aalen-Johansen estimator at time t_{i+1}. To see the equivalence between the above introduced calculation and this algorithm we examine the case of a three-state Markovian model, where we observe a jump from state 1 to 2 at time t_{i+1}. The elements of the matrix $\hat{\mathbf{P}}(s, t_i)$ are denoted by $(p_{ij})_{i,j=1,\dots,3}$ that is:

$$\hat{\mathbf{P}}(s, t_i) = \begin{pmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{pmatrix}$$

If a jump occurs from state g to state h at time t_{i+1}, the matrix $\hat{\mathbf{P}}(s, t_i)$ has to be multiplied, according to (4.22), with the following matrix, where we have chosen $g = 1$ and $h = 2$:

$$\begin{pmatrix} 1 - \frac{1}{r_{gj}} & \frac{1}{r_{gj}} & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Thus we obtain:

$$\hat{\mathbf{P}}(s, t_{i+1}) = \begin{pmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{pmatrix} \times \begin{pmatrix} 1 - \frac{1}{r_{gj}} & \frac{1}{r_{gj}} & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} p_{11}(1 - \frac{1}{r_{gj}}) & p_{11}\frac{1}{r_{gj}} + p_{12} & p_{13} \\ p_{21}(1 - \frac{1}{r_{gj}}) & p_{21}\frac{1}{r_{gj}} + p_{22} & p_{23} \\ p_{31}(1 - \frac{1}{r_{gj}}) & p_{31}\frac{1}{r_{gj}} + p_{32} & p_{33} \end{pmatrix}$$

Applying the algorithm in the example of the two-state model to $\hat{\mathbf{P}}(0, t_2)$ we have:

$$\hat{\mathbf{P}}(0, t_1) = \begin{pmatrix} \frac{14}{15} & \frac{1}{15} \\ 0 & 1 \end{pmatrix}$$

Further we have $1 - 1/r_{gj} = 1 - 1/13 = 12/13$. The jump we observe at time t_2 goes from state 1 to 2, thus we multiply the first column of $\hat{\mathbf{P}}(0, t_1)$ by 12/13 and the second column by 1/13 and add it to the second column. This leads us to

$$\hat{\mathbf{P}}(0, t_2) = \left(\begin{pmatrix} \frac{14}{15} \\ 0 \end{pmatrix} \times \frac{12}{13}\ , \begin{pmatrix} \frac{14}{15} \\ 0 \end{pmatrix} \times \frac{1}{13} + \begin{pmatrix} \frac{1}{15} \\ 1 \end{pmatrix} \right) = \begin{pmatrix} \frac{56}{65} & \frac{9}{65} \\ 0 & 1 \end{pmatrix}$$

which is the same result as derived before in Example 4.1.

4.3.5 Algorithm for Computing the Aalen-Johansen Estimator (Backwards)

Aalen and Johansen (1978) also suggested an algorithm for computing their estimator backwards. Given $\hat{\mathbf{P}}(u_{i+1}, t)$, the value for the Aalen-Johansen estimator at time u_{i+1}, we can compute the Aalen-Johansen estimator at time $u_i < u_{i+1}$, if no transitions occur between those two timepoints. Assume the jump at time u_i goes from state g to state h then the following calculations have to be performed to obtain $\hat{\mathbf{P}}(u_i, t)$:

- The $g'th$ row of $\hat{\mathbf{P}}(u_{i+1}, t)$ is replaced by a convex combination of the $g'th$ and $h'th$ row with weights $\left(1 - \frac{1}{r_{gj}}\right)$ and $\frac{1}{r_{gj}}$, respectively
- All other rows stay unchanged

The resulting matrix is the Aalen-Johansen estimator at time u_i. To see the equivalence between the above introduced calculation and this algorithm we examine the case of a three-state Markovian model, where we observe a jump from state 1 to 2 at time u_i. The elements of the matrix $\hat{\mathbf{P}}(u_{i+1}, t)$ are denoted by $(p_{ij})_{i,j=1,\ldots,3}$ that is:

$$\hat{\mathbf{P}}(u_{i+1}, t) = \begin{pmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{pmatrix}$$

We have to multiply $\hat{\mathbf{P}}(u_{i+1}, t)$ according to (4.22) with the following matrix, to obtain the Aalen-Johansen estimator, if the jump occurs from state state $g = 1$ to state $h = 2$:

$$\mathbf{A} := \begin{pmatrix} 1 - \frac{1}{r_{gj}} & \frac{1}{r_{gj}} & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{4.23}$$

Thus we obtain

$$\begin{aligned} \hat{\mathbf{P}}(u_i, t) &= \underbrace{\begin{pmatrix} 1 - \frac{1}{r_{gj}} & \frac{1}{r_{gj}} & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}}_{\mathbf{A}} \times \begin{pmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{pmatrix} \\ &= \begin{pmatrix} p_{11}(1 - \frac{1}{r_{gj}}) + \frac{p_{21}}{r_{gj}} & p_{12}(1 - \frac{1}{r_{gj}}) + \frac{p_{22}}{r_{gj}} & p_{13}(1 - \frac{1}{r_{gj}}) + \frac{p_{23}}{r_{gj}} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{pmatrix} \end{aligned}$$

This is basically nothing else than to write down the formula (4.22) for the Aalen-Johansen estimator and explicitly specifying all factors, that is all stochastic matrices at the times, where jumps are observed in $[u_{i+1}, t]$, that is at times $u_{i+1} < \ldots < u_{i+r}$, where at $u_{i+r} \leq u_{i+r+1} = t$ the last jump occurred.

$$\hat{\mathbf{P}}(u_i, t) = \prod_{j: u_i < t_j \leq t} (\mathbf{I} + \hat{\boldsymbol{\alpha}}_j) = \mathbf{A} \times \mathbf{A}_1 \times \ldots \times \mathbf{A}_r = \mathbf{A} \times \underbrace{\prod_{j: u_{i+1} < t_j \leq t} (\mathbf{I} + \hat{\boldsymbol{\alpha}}_j)}_{\hat{\mathbf{P}}(u_{i+1}, t)}$$

where $\mathbf{A} := \hat{\mathbf{P}}(u_i, u_{i+1})$ and $\mathbf{A}_k := \hat{\mathbf{P}}(u_{i+k}, u_{i+k+1})$, $k = 1, \ldots, r$. Since no jump occur between times u_i and u_{i+1}, the matrix $\mathbf{A}$ is nothing else than the above specified stochastic matrix (4.23).

4.3.6 In Case of a Multi-State Model with different Transitions at a Time

In this section we relax the assumption made so far that only one transition takes place at a given point in time. Questions, that arise in this situation, are, whether the resulting estimator still gives a stochastic matrix, as required by the nature of the transition matrix, whether the Chapman-Kolmogorov equations still hold, as required by the properties of a Markovian multi-state model and whether there is a similar algorithm available for computing the Aalen-Johansen estimator, as introduced in Section 4.3.4.

A positive answer for the first question can be given quite easily. We replace in the Aalen-Johansen estimator the $(g,g)^{th}$ diagonal element by $-d_{gj}/r_{gj}$, where $d_{gj} = \sum_{h \neq g} d_{ghj}$ according to Definition 4.2. If in addition to the jump from state g to h at time t_j, a jump from state g to k is observed at the same time, the $(g,k)^{th}$ diagonal element is equal to d_{gkj}/r_{gj} and the property of a stochastic matrix is preserved. This remains true if at time t_j a jump from state r to s takes place, as well. The $(r,r)^{th}$ diagonal element becomes $-d_{rj}/r_{rj}$ and the $(r,s)^{th}$ element becomes d_{rsj}/r_{rj}. As a stochastic matrix only requires that the sum over the elements in each row are equal to one, the just above outlined matrix is therefore a stochastic matrix and the Aalen-Johansen estimator remains a product of stochastic matrices.

$$
\begin{aligned}
\hat{\mathbf{P}}(0,t) &= \prod_{j:t_j \leq t} \begin{pmatrix}
1 & \dots & \dots & \dots & \dots & \dots & 0 \\
0 & 1-\frac{d_{rj}}{r_{rj}} & \dots & \frac{d_{rsj}}{r_{rj}} & \dots & \dots & 0 \\
0 & \dots & 1-\frac{d_{gj}}{r_{gj}} & \dots & \frac{d_{ghj}}{r_{gj}} & \frac{d_{gkj}}{r_{gj}} & 0 \\
0 & \dots & \dots & 1 & \dots & \dots & 0 \\
0 & \dots & \dots & \dots & 1 & \dots & 0 \\
0 & \dots & \dots & \dots & \dots & 1 & 0 \\
0 & \dots & \dots & \dots & \dots & \dots & 1
\end{pmatrix} \\
&= \prod_{j:t_j \leq t} \begin{pmatrix}
1 & \dots & \dots & \dots & \dots & \dots & 0 \\
0 & 1-\sum_{s \neq r} \frac{d_{rsj}}{r_{rj}} & \dots & \frac{d_{rsj}}{r_{rj}} & \dots & \dots & 0 \\
0 & \dots & 1-\sum_{h \neq g} \frac{d_{ghj}}{r_{gj}} & \dots & \frac{d_{ghj}}{r_{gj}} & \frac{d_{gkj}}{r_{gj}} & 0 \\
0 & \dots & \dots & 1 & \dots & \dots & 0 \\
0 & \dots & \dots & \dots & 1 & \dots & 0 \\
0 & \dots & \dots & \dots & \dots & 1 & 0 \\
0 & \dots & \dots & \dots & \dots & \dots & 1
\end{pmatrix}
\end{aligned}
$$

We show now that a product of stochastic matrices is again a stochastic matrix, the Aalen-Johansen estimator is a stochastic matrix and therefore fulfills the requirements of a transition matrix of a Markovian multi-state model. First we consider the case $n = 3$:

Let $\mathbf{P} = (p_{ij})_{ij}$ and $\mathbf{Q} = (q_{ij})_{ij}$ be two 3×3 stochastic matrices, that is the sum over each row is equal to one. The matrix product $\mathbf{P} * \mathbf{Q}$ is given by

$$
\begin{aligned}
\mathbf{P} \times \mathbf{Q} &= \begin{pmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{pmatrix} \times \begin{pmatrix} q_{11} & q_{12} & q_{13} \\ q_{21} & q_{22} & q_{23} \\ q_{31} & q_{32} & q_{33} \end{pmatrix} \\
&= \begin{pmatrix}
p_{11}q_{11} + p_{12}q_{21} + p_{13}q_{31} & p_{11}q_{12} + p_{12}q_{22} + p_{13}q_{32} & p_{11}q_{13} + p_{12}q_{23} + p_{13}q_{33} \\
p_{21}q_{11} + p_{22}q_{21} + p_{23}q_{31} & p_{21}q_{12} + p_{22}q_{22} + p_{23}q_{32} & p_{21}q_{13} + p_{22}q_{23} + p_{23}q_{33} \\
p_{31}q_{11} + p_{32}q_{21} + p_{33}q_{31} & p_{31}q_{12} + p_{32}q_{22} + p_{33}q_{32} & p_{31}q_{13} + p_{32}q_{23} + p_{33}q_{33}
\end{pmatrix}
\end{aligned}
$$

Taking sums over the elements of each row we obtain:

$$\begin{aligned}
(\mathbf{P} * \mathbf{Q})_{[1,]} &= (p_{11}q_{11} + p_{12}q_{21} + p_{13}q_{31}) + (p_{11}q_{12} + p_{12}q_{22} + p_{13}q_{32}) + (p_{11}q_{13} + p_{12}q_{23} + p_{13}q_{33}) \\
&= p_{11} \underbrace{(q_{11} + q_{12} + q_{13})}_{=1} + p_{12} \underbrace{(q_{21} + q_{22} + q_{23})}_{=1} + p_{13} \underbrace{(q_{31} + q_{32} + q_{33})}_{=1} \\
&= p_{11} + p_{12} + p_{13} = 1
\end{aligned}$$

as both, **P** and **Q** are stochastic matrices. It is obvious that the same holds for the second and third row. This proves that the product of two stochastic matrices is again a stochastic matrix. The approach can easily be extended to any $n \times n$ matrix.

□

The second question asked, whether the Chapman-Kolmogorov equations still hold, that is $\mathbf{P}(s,t) = \mathbf{P}(s,u)\mathbf{P}(u,t)$ for $s \leq u \leq t$. In Section 4.3.1 the Aalen-Johansen estimator was introduced such that it fulfills these equations. Since the modifications we introduced to the Aalen-Johansen estimator in this section do not affect the time structure, when jumps are observed, but only the jumps themselves at a given point in time, the Chapman-Kolmogorov equations still hold.

$$\hat{\mathbf{P}}(s,t) = \prod_{j:s<t_j\leq t} (\mathbf{I} + \hat{\boldsymbol{\alpha}}_j) = \underbrace{\prod_{j:s<t_j\leq u} (\mathbf{I} + \hat{\boldsymbol{\alpha}}_j)}_{\hat{\mathbf{P}}(s,u)} \times \underbrace{\prod_{j:u<t_j\leq t} (\mathbf{I} + \hat{\boldsymbol{\alpha}}_j)}_{\hat{\mathbf{P}}(u,t)} = \hat{\mathbf{P}}(s,u) \times \hat{\mathbf{P}}(u,t)$$

where $\hat{\boldsymbol{\alpha}}_j$ is now the modified matrix as explained in this section.

The third and last question, that arose, dealt with an algorithm for computing the Aalen-Johansen estimator. Obviously, in the case of a multi-state model with different transition at a time, the algorithm gets more complex and a calculation with direct specification of the different matrices of the estimator might be more preferable. We assume that transitions are observed at time t_j and distinguish three cases:

First case:

If there is no transition to and from state g, the g^{th} column stays unchanged.

Second case:

If there is a transition from state g to state $h, k, \ldots$ the following calculations have to be performed:

- The g^{th} column is multiplied by $(1 - d_{gj}/r_{gj})$
- The g^{th} column is multiplied by d_{ghj}/r_{gj} and added to the h^{th} column
- The g^{th} column is multiplied by d_{gkj}/r_{gj} and added to the k^{th} column
- ...

Third case:

If there is a transition from state g to state $h, k, \ldots$ and at the same time a transition from state i to state g, the necessary calculations are the following:

- The i^{th} column: Multiply the i^{th} column by $(1 - d_{ij}/r_{ij})$
- The g^{th} column: Multiply the i^{th} column by d_{igj}/r_{igj} and add $(1 - d_{ij}/r_{ij})$ of the g^{th} column
- The h^{th} column: Multiply the g^{th} column by d_{ghj}/r_{gj} and add the h^{th} column
- The k^{th} column: Multiply the g^{th} column by d_{gkj}/r_{gj} and add the k^{th} column
- ...

Chapter 5

Regression using Pseudo-Values

As we have seen in previous chapters, the Aalen-Johansen estimator is an unbiased estimator for the transition matrix of a Markovian multi-state process. Given a set of data containing information on transitions of observations as well as additional covariates of these observations, e.g. age, sex, . . . , we are now able to calculate the Aalen-Johansen estimator, that is the matrix for the transition probabilities of a specified Markovian multi-state model.

But, the whole data set produces only one single outcome that does not depend on covariates at all, it only uses the information on the transitions that occurred. In order to perform a regression analysis to investigate the effect of different covariates on the transition probabilities, we need to generate the necessary outcomes and link these outcomes with the covariates.

Andersen, Klein, and Rosthøj (2003) used so-called pseudo-values known from jackknife methodology (Efron and Tibshirani 1993) to overcome these problems. The i^{th} pseudo-value is obtained by calculating the estimator, e.g. the Aalen-Johansen estimator, using the whole data set and a data set with the i^{th} observation removed.

Usually these "leave-one-out diagnostics" are used to asses the bias and precision of the estimator by comparing the "leave-one-out diagnostics" with the estimator based on the entire sample. Here, we try to extract information on the way in which the covariates of each individual affect the estimator and perform a regression analysis with these pseudo-values.

Firstly we define the so-called i^{th} jackknife sample and the i^{th} jackknife replication of the estimator involved and derive the jackknife estimate of bias and standard error. Secondly we introduce pseudo-values, a different representation of the jackknife. Given n observations we are able to calculate n pseudo-values and thus generate the data required for a regression analysis.

Since the Aalen-Johansen estimator is an unbiased estimator, the expectation of the i^{th} pseudo-value is equal to the conditional expectation of the claim-history given the covariates of the i^{th} observation and we can match the i^{th} pseudo-value with the covariates of the i^{th} observation and thus construct a relationship between the pseudo-values and the covariates of the observations.

Finally we show how these tools are used to perform a regression analysis for a set of data containing the claim-history of LTC patients. We establish a quasi-likelihood model for the transition probabilities using the logit as link function. Further details on likelihood and quasi-likelihood methods can then be found in Chapter 6.

5.1 Jackknife

As already noted above, the jackknife is a method to estimate the bias and standard error of an estimate (Efron and Tibshirani 1993). In the following we have a sample of n observations $\mathbf{x} = (x_1, \ldots, x_n)$ and an estimator $\hat{\theta} = s(\mathbf{x})$ based on these observations. We define the i^{th} jackknife sample as

$$\mathbf{x}_{(-i)} := (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n),$$

that is the entire data set with the i^{th} observation removed. Further, the i^{th} jackknife replication of $\hat{\theta}$, the so-called "leave-one-out estimator", is defined as

$$\hat{\theta}_{-i} := s(\mathbf{x}_{(-i)}) \tag{5.1}$$

The estimator $\hat{\theta}_{-i}$ is the same estimator as $\hat{\theta}$, but based on the i^{th} jackknife sample and not on the whole data set. We are now going to define the jackknife estimate of bias and the jackknife estimate of standard error for an estimator $\hat{\theta}$ defined as above.

Before doing so, recall the two well-known quantities $\overline{x}$ and s^2, that are defined for n independent and identically distributed observations x_i, where $E[x_i] = \mu$ and $Var[x_i] = \sigma^2$:

$$\overline{x} := \frac{1}{n}\sum_{i=1}^{n} x_i \qquad\qquad s^2 := \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Definition 5.1 (Jackknife Estimate of Bias) *The jackknife estimate of bias for $\hat{\theta}$ is defined as the following quantity:*

$$\hat{bias}_{jack}(\hat{\theta}) := (n-1)\cdot\left(\hat{\theta}_{(\cdot)} - \hat{\theta}\right)$$

where $\hat{\theta}_{(\cdot)}$ is the average of the $\hat{\theta}_{-i}$'s, that is

$$\hat{\theta}_{(\cdot)} := \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{-i}.$$

Definition 5.2 (Jackknife Estimate of Standard Error) *The jackknife estimate of standard error for $\hat{\theta}$ is defined by*

$$\hat{se}_{jack}(\hat{\theta}) := \left(\frac{n-1}{n}\cdot\sum_{i=1}^{n}\left(\hat{\theta}_{-i} - \hat{\theta}_{(\cdot)}\right)^2\right)^{\frac{1}{2}} \tag{5.2}$$

An obvious question arises now: Why do we use the factors $(n-1)$ and $(n-1)/n$ in the jackknife estimate of bias and the jackknife estimate of standard error, respectively?

The answer given by Efron and Tibshirani (1993) is that this is "somewhat arbitrary convention": Using these factors ensures that in the case of iid data the jackknife estimate of bias for the sample variance $\hat{\sigma}^2 := (n-1)/n\, s^2$ is equal to the unbiased estimate of bias of $\hat{\sigma}^2$ and the jackknife estimate of standard error for the sample mean $\overline{x}$ is equal to the unbiased estimate of the standard error of $\overline{x}$.

We do not use the sample mean $\overline{x}$ in the first case, since $\overline{x}$ is an unbiased estimator for the mean of iid data and therefore $\hat{\theta}_{(\cdot)} - \hat{\theta} = 0$, that is

$$\begin{aligned}
\hat{\theta}_{(\cdot)} - \hat{\theta} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{-i} - \hat{\theta} &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{n-1}\sum_{j\neq i} x_j - \frac{1}{n}\sum_{j=1}^{n} x_j = \frac{1}{n}\cdot\frac{1}{n-1}\sum_{i=1}^{n}\sum_{j\neq i} x_j - \frac{1}{n}\sum_{j=1}^{n} x_j \\
&= \frac{1}{n(n-1)}\cdot\sum_{i=1}^{n}(x_1 + \ldots + x_{i-1} + \ldots + x_{i+1} + \ldots x_n) - \frac{1}{n}\sum_{j=1}^{n} x_j \\
&= \frac{1}{n(n-1)}\cdot\left((n-1)\cdot\sum_{i=1}^{n} x_i\right) - \frac{1}{n}\sum_{j=1}^{n} x_j = 0
\end{aligned}$$

Thus it follows:

$$\widehat{bias}_{jack}(\overline{x}) = (n-1)\cdot\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{-i} - \hat{\theta}\right) = 0$$

The jackknife estimate of bias for the sample variance

We have to calculate $\widehat{bias}_{jack}$ for the estimate $\hat{\sigma}^2 := 1/n\sum_{j=1}^{n}(x_i - \overline{x})^2 = 1/n\sum_{j=1}^{n}(x_i^2 - n\overline{x}^2)$; as a result we obtain

$$\begin{aligned}
\widehat{bias}_{jack}(\hat{\theta}) &:= (n-1)\cdot\left(\hat{\theta}_{(\cdot)} - \hat{\theta}\right) = (n-1)\cdot\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{-i} - \hat{\theta}\right) \\
\widehat{bias}_{jack}(\hat{\sigma}^2) &= (n-1)\cdot\left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{n-1}\cdot\left(\sum_{j\neq i} x_j^2 - \frac{1}{n-1}\cdot(n\overline{x} - x_i)^2\right) - \frac{1}{n}\left(\sum_{j=1}^{n} x_i^2 - n\overline{x}^2\right)\right) \\
&= (n-1)\cdot\left(\frac{1}{n(n-1)}\cdot\sum_{i=1}^{n}(n-1)\cdot x_i^2 - \frac{1}{n(n-1)^2}\cdot\sum_{i=1}^{n}(n\overline{x} - x_i)^2 - \frac{1}{n}\sum_{j=1}^{n} x_i^2 + \overline{x}^2\right) \\
&= (n-1)\cdot\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \frac{1}{n(n-1)^2}\cdot\left(n^3\overline{x}^2 - 2n^2\overline{x}^2 + \sum_{i=1}^{n} x_i^2\right) - \frac{1}{n}\sum_{j=1}^{n} x_i^2 + \overline{x}^2\right) \\
&= (n-1)\cdot\left(-\frac{n(n-2)}{(n-1)^2}\cdot\overline{x}^2 - \frac{1}{n(n-1)^2}\cdot\sum_{i=1}^{n} x_i^2 + \overline{x}^2\right) \\
&= (n-1)\cdot\left(-\frac{1}{n(n-1)^2}\cdot\sum_{i=1}^{n} x_i^2 - \left(\frac{n(n-2)}{(n-1)^2} - 1\right)\cdot\overline{x}^2\right) \\
&= (n-1)\cdot\left(-\frac{1}{n(n-1)^2}\cdot\sum_{i=1}^{n} x_i^2 - \frac{n^2 - 2n - n^2 + 2n - 1}{(n-1)^2}\cdot\overline{x}^2\right) \\
&= -\frac{1}{n(n-1)}\cdot\sum_{i=1}^{n} x_i^2 + \frac{1}{(n-1)}\cdot\overline{x}^2 = -\frac{1}{n(n-1)}\cdot\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right) \\
&= -\frac{1}{n(n-1)}\cdot\sum_{i=1}^{n}(x_i - \overline{x})^2
\end{aligned}$$

Thus we have the following result:

$$\hat{bias}_{jack}(\sigma^2) = -\frac{1}{n} \cdot \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2 = -\frac{s^2}{n}$$

This is equal to the unbiased estimate of the bias of $\hat{\sigma}^2$:

$$\hat{bias}_{jack}(\hat{\sigma}^2) = -\frac{s^2}{n} \qquad \text{where} \qquad s^2 := \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2$$

since the following holds for $\hat{\sigma}^2$ in the case of iid data:

$$E[\hat{\sigma}^2] = \sigma^2 - \frac{\sigma^2}{n} \qquad \Rightarrow \qquad bias(\hat{\sigma}^2) = E[\hat{\sigma}^2] - \sigma^2 = -\frac{\sigma^2}{n}$$

The jackknife estimate of standard error for the mean

We have to calculate $\hat{se}_{jack}$ for the estimate $\overline{x} = 1/n \sum_{i=1}^{n} x_i$:

$$\begin{aligned}
\hat{se}_{jack}(\hat{\theta}) &:= \left(\frac{n-1}{n} \cdot \sum_{i=1}^{n} \left(\hat{\theta}_{-i} - \hat{\theta}_{(\cdot)} \right)^2 \right)^{\frac{1}{2}} \\
&= \left(\frac{n-1}{n} \cdot \sum_{i=1}^{n} \left(\hat{\theta}_{-i} - \frac{1}{n} \sum_{l=1}^{n} \hat{\theta}_{-l} \right)^2 \right)^{\frac{1}{2}}
\end{aligned}$$

Now inserting $\overline{x}$ gives:

$$\begin{aligned}
\hat{se}_{jack}(\overline{x}) &= \left(\frac{n-1}{n} \cdot \sum_{i=1}^{n} \left(\frac{1}{n-1} \sum_{j \neq i} x_j - \frac{1}{n} \sum_{l=1}^{n} \frac{1}{n-1} \sum_{j \neq l} x_j \right)^2 \right)^{\frac{1}{2}} \\
&= \left(\frac{n-1}{n} \cdot \sum_{i=1}^{n} \left(\frac{1}{n-1} \right)^2 \cdot \left(\sum_{j \neq i} x_j - \frac{1}{n} \cdot (n-1) \cdot \sum_{j=1}^{n} x_j \right)^2 \right)^{\frac{1}{2}} \\
&= \left(\frac{1}{n(n-1)} \cdot \sum_{i=1}^{n} \left[\frac{1}{n} \cdot \left(\sum_{j \neq i} n x_j - (n-1) \cdot \sum_{j=1}^{n} x_j \right) \right]^2 \right)^{\frac{1}{2}} \\
&= \left(\frac{1}{n(n-1)} \cdot \sum_{i=1}^{n} \left[\frac{1}{n} \cdot \left(\sum_{j=1}^{n} x_j - n x_i \right) \right]^2 \right)^{\frac{1}{2}} \\
&= \left(\frac{1}{n(n-1)} \cdot \sum_{i=1}^{n} (x_i - \overline{x})^2 \right)^{\frac{1}{2}} \qquad (5.3)
\end{aligned}$$

This is equal to the unbiased estimate of the standard error of $\overline{x}$, if we use iid data, which is

$$\hat{se}(\overline{x}) = \left(\frac{s^2}{n} \right)^{\frac{1}{2}} \qquad \text{since} \qquad Var[\overline{x}] = \frac{\sigma^2}{n}$$

5.2 Pseudo-Values

A different representation of the jackknife are pseudo-values. They are defined in terms of the estimator $\hat{\theta}$ and the i^{th} jackknife replication of $\hat{\theta}$, that is

$$\tilde{\theta}_i := \hat{\theta} + (n-1) \cdot (\hat{\theta} - \hat{\theta}_{-i}) = n \cdot \hat{\theta} - (n-1) \cdot \hat{\theta}_{-i} \tag{5.4}$$

These pseudo-values are supposed to act as if they were n independent data values. This idea can be understood looking at the following lemma:

Lemma 5.1 *For any $\hat{\theta}$ the formula for $\widehat{se}_{jack}(\hat{\theta})$ can be expressed as*

$$\widehat{se}_{jack}(\tilde{\theta}) = \left(\frac{1}{n(n-1)} \cdot \sum_{i=1}^{n} \left(\tilde{\theta}_i - \tilde{\theta} \right)^2 \right)^{\frac{1}{2}}$$

where $\tilde{\theta} := 1/n \sum_{i=1}^{n} \tilde{\theta}_i$, i.e. it holds $\widehat{se}_{jack}(\hat{\theta}) = \widehat{se}_{jack}(\tilde{\theta})$.

This equation has a form similar to the jackknife estimate of the standard error for $\overline{x}$ derived in equation (5.3), where the data now is given by the $\tilde{\theta}_i$. This supports the idea looking at the pseudo-values as independent observations.

Proof of Lemma 5.1:

$$\begin{aligned}
\widehat{se}_{jack}(\tilde{\theta}) &= \left(\frac{1}{n(n-1)} \cdot \sum_{i=1}^{n} \left(\tilde{\theta}_i - \tilde{\theta} \right)^2 \right)^{\frac{1}{2}} \\
&= \left(\frac{1}{n(n-1)} \cdot \sum_{i=1}^{n} \left(n \cdot \hat{\theta} - (n-1) \cdot \hat{\theta}_{-i} - \frac{1}{n} \sum_{i=1}^{l} \tilde{\theta}_l \right)^2 \right)^{\frac{1}{2}} \\
&= \left(\frac{1}{n(n-1)} \cdot \sum_{i=1}^{n} \left(n \cdot \hat{\theta} - (n-1) \cdot \hat{\theta}_{-i} - \frac{1}{n} \sum_{i=l}^{n} (n \cdot \hat{\theta} - (n-1) \cdot \hat{\theta}_{-l}) \right)^2 \right)^{\frac{1}{2}} \\
&= \left(\frac{n-1}{n} \cdot \sum_{i=1}^{n} \left(-\hat{\theta}_{-i} + \frac{1}{n} \sum_{i=l}^{n} \hat{\theta}_{-l} \right)^2 \right)^{\frac{1}{2}} \\
&= \left(\frac{n-1}{n} \cdot \sum_{i=1}^{n} \left(\hat{\theta}_{-i} - \frac{1}{n} \sum_{i=1}^{l} \hat{\theta}_{-l} \right)^2 \right)^{\frac{1}{2}} \\
&= \left(\frac{n-1}{n} \cdot \sum_{i=1}^{n} \left(\hat{\theta}_{-i} - \hat{\theta}_{(\cdot)} \right)^2 \right)^{\frac{1}{2}} \\
&= \widehat{se}_{jack}(\hat{\theta})
\end{aligned}$$

which is the jackknife estimate of standard error $\widehat{se}_{jack}(\hat{\theta})$ as defined in (5.2). □

Example 5.1 (Application to regression data)

Let X_i be a random variable, e.g. the claim-history of the i^{th} individual. We assume that the X_i's are independent and identically distributed with expectation θ, and that an unbiased estimator $\hat{\theta}$ is available for θ, e.g. the Aalen-Johansen estimator, that is

$$E[\hat{\theta}] = \theta := E\left[X_i\right]$$

Further we introduce independent and identically distributed covariates $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{ip})^T$, that follow the distribution C. Thus we have:

$$\theta := E\left[X_i\right] = E\left[E\left[X_i \middle| \mathbf{Z}_i\right]\right] = \int E\left[X_i \middle| \mathbf{Z}_i = \mathbf{z}_i\right] dC(\mathbf{z}_i)$$

Since $\hat{\theta}$ is an unbiased estimator its expectation is equal to θ . But also the expectation of the quantity $E\left[X_i \middle| \mathbf{Z}_i\right]$ is equal to θ. We define θ_i to be this conditional expectation, i.e.

$$\theta_i := E\left[X_i \middle| \mathbf{Z}_i = \mathbf{z}_i\right].$$

If we estimate C by $\hat{C}$, the empirical distribution of $\mathbf{Z}_i$, it follows that the parameter θ might be interpreted as the simple average of the θ_i's, that is

$$E[\hat{\theta}] = \theta \approx \frac{1}{n} \sum_{i=1}^{n} \theta_i \tag{5.5}$$

Since the estimator $\hat{\theta}$ is an unbiased estimator for θ it follows that the "leave-one-out estimator", namely $\hat{\theta}_{-i}$, the i^{th} jackknife replication of θ as defined in (5.1), is also an unbiased estimator for θ and can be approximated in terms of the θ_i's. This can be understood, if we consider instead of the full data set the i^{th} jackknife sample, and apply the same reasoning as for $\hat{\theta}$:

$$E\left[\hat{\theta}_{-i}\right] \approx \frac{1}{n-1} \sum_{j \neq i} \theta_j \tag{5.6}$$

Since the data is only available given the covariates, we need a link between the estimator $\hat{\theta}$ and the quantity $\theta_i = E\left[X_i \middle| \mathbf{Z}_i = \mathbf{z}_i\right]$. Therefore Andersen, Klein, and Rosthøj (2001) defined the quantity $\tilde{\theta}_i$ as the summary statistic $\hat{\theta}$ based on the entire sample modified in the direction given by the "leave-one-out estimator" $(\hat{\theta} - \hat{\theta}_{-i})$, that we defined as the pseudo-values in (5.4):

$$\tilde{\theta}_i := n \cdot \hat{\theta} - (n-1) \cdot \hat{\theta}_{-i} \tag{5.7}$$

One can show, using the representation of θ as an average of the θ_i's that the expectation of the i^{th} pseudo-value $\tilde{\theta}_i$ is equal to the quantity $\theta_i = E\left[X_i \middle| \mathbf{Z}_i = \mathbf{z}_i\right]$:

$$E\left[\tilde{\theta}_i\right] = n \cdot E\left[\hat{\theta}\right] - (n-1) \cdot E\left[\hat{\theta}_{-i}\right] \overset{(5.5),(5.6)}{\approx} \sum_{j} \theta_j - \sum_{j \neq i} \theta_j = \theta_i$$

Therefore we can use $\tilde{\theta}_i$ as pseudo-values to fit the regression model

$$\tilde{\theta}_i = \theta_i + \varepsilon_i$$

where $\theta_i = E\left[X_i \middle| \mathbf{Z}_i = \mathbf{z}_i\right]$ is a specified regression mean.

It became clear in this example that we can use the pseudo-values to construct a relationship between an unbiased estimator $\hat{\theta}$, for example the Aalen-Johansen estimator $\hat{\mathbf{P}}(s,t)$, and the covariates of each single observation. The Aalen-Johansen itself does not depend on the covariates, thus we match the i^{th} pseudo-value with the covariates of the i^{th} observation. The pseudo-values allow us not only to generate the data required for a regression analysis, calculating the Aalen-Johansen estimator out of a given set of data would result in only one outcome, but also to construct a relationship between the pseudo-values and the covariates.

A regression analysis of above model produces estimates for the transition matrix over the interval $(s,t]$. But to calculate the necessary actuarial values, we need the one-year transition probabilities. Therefore we have to look at a multivariate case, where we calculate the Aalen-Johansen estimator in different intervals.

Example 5.2 (Fitting transition matrices which depend on covariates)

The above situation can be extended to the multivariate case: We consider a series of time-points $t_0, \ldots, t_k$ and define $\hat{\boldsymbol{\theta}} := (\hat{\boldsymbol{\theta}}(t_0), \ldots, \hat{\boldsymbol{\theta}}(t_k))$ calculating the pseudo-values analogue to (5.7) as

$$\tilde{\boldsymbol{\theta}}_{il} = n \cdot \hat{\boldsymbol{\theta}}(t_l) - (n-1) \cdot \hat{\boldsymbol{\theta}}_{-i}(t_l) \qquad i = 1, \ldots, n \qquad l = 0, \ldots, k \tag{5.8}$$

For our purpose we use the Aalen-Johansen estimator (4.22) and define the pseudo-values element-wise, that is for each element of the transition matrix,

$$\hat{\boldsymbol{\theta}}^{(gh)}(t_l) := \hat{\mathbf{P}}_{gh}(l, l+1) \qquad l = 0, \ldots, k$$

giving us above $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\theta}}(t_0), \ldots, \hat{\boldsymbol{\theta}}(t_k))$. We calculate then the Aalen-Johansen estimator based on the entire sample and the "leave-one-out estimator" for all $i = 1, \ldots, n$ and $l = 0, \ldots, k$ giving us the pseudo-values $\tilde{\boldsymbol{\theta}}_{il}$ as defined in (5.8).

Note that the Aalen-Johansen estimator is a matrix; therefore we obtain for each observation i one matrix of pseudo-values $\tilde{\boldsymbol{\theta}}_{il}$ at each time-point t_l, $l = 0, \ldots, k$.

We assume now a regression model for each element of $\tilde{\boldsymbol{\theta}}_{il}$ on $\mathbf{Z}_i$ to quantify the effect of the covariates on the transition probabilities. To perform this regression we assume that the relationship between $\tilde{\theta}_{il}^{(gh)}$, where $g, h \in \mathcal{S}$ denotes the possible entries of the transition matrix, and the covariates is given by a GLM with link function $g(\cdot)$, that is

$$g(\tilde{\theta}_{il}^{(gh)}) = \mathbf{Z}_i^T \boldsymbol{\beta}^{(gh)}$$

The first covariate Z_{i1} is usually assumed to be equal to one to include an intercept term for $\boldsymbol{\beta}$.

In the following we are going to specify the approach described above, to give an idea how it is used in our data application (see Chapter 7):

Assume a three-state model with states labeled 1, 2 and 3. We calculate the Aalen-Johansen estimator $\hat{\boldsymbol{\theta}}(t_l) \in \mathbb{R}^{3\times 3}$, $l = 0, \ldots, k$, for the transition matrix of this three-state model. Since this is an unbiased estimator, we can calculate with the "leave-one-out estimators", as seen above, our n matrices of pseudo-values $\tilde{\boldsymbol{\theta}}_{il}$ at times t_l, $l = 0, \ldots, k$. As covariates we include an intercept term, "Age" and "Sex", denoted by Z_{i1}, Z_{i2} and Z_{i3}.

Note that the Aalen-Johansen estimator is a matrix. Therefore we perform a regression analysis for each element of this matrix. Our model for the pseudo-values of the transition probability from state g to state h is the following:

$$\tilde{\theta}_{il}^{(gh)} = \frac{\exp\{\alpha_l^{(gh)} + Z_{i1}\beta_1^{(gh)} + Z_{i2}\beta_2^{(gh)} + Z_{i3}\beta_3^{(gh)}\}}{1 + \exp\{\alpha_l^{(gh)} + Z_{i1}\beta_1^{(gh)} + Z_{i2}\beta_2^{(gh)} + Z_{i3}\beta_3^{(gh)}\}} + \varepsilon_{il}^{(gh)} \qquad i = 1, \dots, n \quad l = 0, \dots, k$$

where $\varepsilon_{il}^{(gh)} \sim N(0, \sigma_{(gh)}^2)$ and $\tilde{\theta}_{il}^{(gh)}$ is according to the above an unbiased estimator for the transition probability $\mathbf{P}_{gh}(l, l+1)$, the quantity we are interested in. Further we define

$$\begin{aligned} \eta_{il}^{(gh)} &:= \alpha_d^{(gh)} + Z_{i1}\beta_1^{(gh)} + Z_{i2}\beta_2^{(gh)} + Z_{i3}\beta_3^{(gh)} \\ \mu_{il}^{(gh)} &:= \frac{\exp\{\alpha_l^{(gh)} + Z_{i1}\beta_1^{(gh)} + Z_{i2}\beta_2^{(gh)} + Z_{i3}\beta_3^{(gh)}\}}{1 + \exp\{\alpha_l^{(gh)} + Z_{i1}\beta_1^{(gh)} + Z_{i2}\beta_2^{(gh)} + Z_{i3}\beta_3^{(gh)}\}} = \frac{\exp\{\eta_{il}^{(gh)}\}}{1 + \exp\{\eta_{il}^{(gh)}\}} \in [0, 1]. \end{aligned}$$

Note that a regression analysis was performed for each element of the transition matrix. Therefore for each transition probability an own model will be used and different parameter estimates will be obtained. We use the logit as link function, our pseudo-values are restricted to the interval $(0, 1)$ due to the definition of the natural logarithm.

$$\begin{aligned} g(\mu_{il}^{(gh)}) = logit(\mu_{il}^{(gh)}) &= \ln \underbrace{\frac{\mu_{il}^{(gh)}}{1 - \mu_{il}^{(gh)}}}_{\in (0,1)} = \eta_{il}^{(gh)} \in \mathbb{R} \\ \mu_{il}^{(gh)} = g^{-1}(\eta_{il}^{(gh)}) &= \frac{\exp\{\eta_{il}^{(gh)}\}}{1 + \exp\{\eta_{il}^{(gh)}\}} \in [0, 1] \end{aligned}$$

The formula for the pseudo-values $\tilde{\boldsymbol{\theta}}_{il} = n \cdot \hat{\boldsymbol{\theta}}(t_l) - (n-1) \cdot \hat{\boldsymbol{\theta}}_{-i}(t_l)$ requires that for large n the "leave-one-out estimator" is very close to the full estimator. Since the data usually does not meet this accuracy, we observe values smaller than 0 or larger than 1, which cause problems for the choice of the link function. To avoid these problems, we assumed a normal error distribution for the pseudo-values and specify the link function with the logit and the variance function as constant, that is

$$\tilde{\theta}_{il}^{(gh)} \sim N(\mu_{il}^{(gh)}, \sigma_{(gh)}^2)$$

where $\mu_{il}^{(gh)}$ is the value of the inverse of the link function evaluated at $\eta_{il}^{(gh)}$.

Chapter 6

Parameter Estimation

In the last chapter we generated the data required for a regression analysis and provided the reasoning why we can match the i^{th} pseudo-value with the covariates of the i^{th} observation. We also presented the model we want to use for a regression analysis. Aim of this chapter is now to give an overview of the estimation methods available and introduce the one we used:

We start with a short introduction to Maximum Likelihood inference in general and show how these tools can be applied to a class of distribution functions, the exponential family, which leads to the generalized linear models (GLMs). Parameters can be estimated solving the so-called set of Score equations, that is the derivative of the log-likelihood function with respect to the parameters. Using the Newton-Raphson and Fisher-Scoring method, one can show that the Score equations can be solved using an iterative weighted least-squares algorithm.

In the case of GLMs an assumption on the whole distribution function of the outcomes is made. Generalizing this approach leads us to quasi-likelihood estimation, where only the relationship between the mean and the variance is specified. One can show that the quasi-likelihood function is nothing else than the log-likelihood function if the outcomes follow a distribution function from the exponential family fulfilling the relationship between the mean and the variance. Estimation is done using the so-called Score-like equations for quasi-likelihood. Similar to the case of GLMs an iterative weighted least-squares algorithm is available to solve these Score-like equations.

Both the GLMs and the quasi-likelihood approach relay on the assumption of independence between observations. In situations with longitudinal data, that is we record for example n observations at n_i different time-points, this assumption still holds for the outcomes of different observations but obviously not for the outcomes of the same observation at different points in time. Therefore we need a model that takes correlation into account. The Generalized Estimating Equations (GEEs) can be seen as an extension of quasi-likelihood to longitudinal data.

Firstly we introduce the Independence Estimating Equations that in contrast to the above assume that all outcomes of the set of longitudinal data are independent. Secondly we sacrifice this assumption and present the GEEs where we allow for correlation between the outcomes of the same observations at different points in time. We show in Theorem 6.5 that under mild conditions the estimate obtained by solving the GEEs is consistent and asymptotically multivariate Gaussian distributed even if the correlation structure is misspecified in the first place. Similar to the GLMs and quasi-likelihood methods, the GEEs can be solved using an iterative weighted least-squares algorithm, but in contrast, one has to iterate between Fisher-scoring for the parameter vector $\boldsymbol{\beta}$ and moment estimation of the correlation structure, as well.

6.1 Maximum Likelihood Inference

One objective of statistical analysis is, as Diggle, Liang, and Zeger (1996) point out, to estimate the unknown probability pattern underlying observed data. Let us assume we are given a set of observed data $\mathbf{y} = (y_1, \ldots, y_n)^T$ and want to find its underlying probability density function $f_i(y_i, \boldsymbol{\theta})$, depending on a $p \times 1$ vector of unknown parameters $\boldsymbol{\theta}$. This is generally done using maximum likelihood methods. Assuming independence of the n observations the joint density of $y_1, \ldots, y_n$ is given by

$$f(\mathbf{y}; \boldsymbol{\theta}) = \prod_{i=1}^{n} f_i(y_i, \boldsymbol{\theta})$$

This means that once the data are observed, only the quantity $\boldsymbol{\theta}$ is unknown and needs to be estimated. Therefore we define the likelihood function as a function of $\boldsymbol{\theta}$, while $\mathbf{y}$ is fixed.

$$L(\boldsymbol{\theta}; \mathbf{y}) := f(\mathbf{y}; \boldsymbol{\theta}) = \prod_{i=1}^{n} f_i(y_i; \boldsymbol{\theta}) = \prod_{i=1}^{n} L_i(\boldsymbol{\theta}; y_i)$$

The maximum likelihood estimate for $\boldsymbol{\theta}$ is usually denoted by $\hat{\boldsymbol{\theta}}$. This is the value that maximizes the likelihood function or equivalently the log-likelihood function denoted by

$$l(\boldsymbol{\theta}; \mathbf{y}) := \ln L(\boldsymbol{\theta}; \mathbf{y}) = \ln \prod_{i=1}^{n} L_i(\boldsymbol{\theta}; y_i) = \sum_{i=1}^{n} \ln L_i(\boldsymbol{\theta}; y_i) = \sum_{i=1}^{n} l_i(\boldsymbol{\theta}; y_i)$$

An estimate for $\boldsymbol{\theta}$ can be obtained by setting the Score function $S(\boldsymbol{\theta})$, the derivative of the log-likelihood function with respect to $\boldsymbol{\theta}$, equal to zero and solve these equations, the so-called Score equations.

$$S(\boldsymbol{\theta}) := \frac{l(\boldsymbol{\theta}; \mathbf{y})}{\partial \boldsymbol{\theta}} = 0 \tag{6.1}$$

The Score function evaluated at the true parameter value $\boldsymbol{\theta}$ has expectation zero and its covariance matrix is given by the information matrix $I(\boldsymbol{\theta}) := E[S(\boldsymbol{\theta})S(\boldsymbol{\theta})^T]$, which under mild regularity conditions can be obtained as minus the expected value of the second order derivatives of the log-likelihood function.

$$I(\boldsymbol{\theta}) = -E\left[\frac{\partial^2 l(\boldsymbol{\theta}; \mathbf{y})}{\partial \boldsymbol{\theta}^2}\right] \tag{6.2}$$

Usually numerical methods are required to solve (6.1) for $\hat{\boldsymbol{\theta}}$, e.g. the Newton-Raphson method.

6.1.1 Newton-Raphson Method

The Newton-Raphson Method is used for solving non-linear equations of the form $f(\mathbf{x}) = \mathbf{0}$. We assume that $\boldsymbol{\xi}$ solves the equation, and approximate $f(\boldsymbol{\xi})$ with a first order Taylor series at $\mathbf{x}_0$ with $\mathbf{x}_0$ close enough to $\boldsymbol{\xi}$. For further details see Stoer (1999).

$$\mathbf{0} = f(\boldsymbol{\xi}) \approx f(\mathbf{x}_0) + Df(\mathbf{x}_0)(\boldsymbol{\xi} - \mathbf{x}_0) \tag{6.3}$$

where $Df(\mathbf{x}_0)$ is the Jacobi-matrix. Solving (6.3) for $\boldsymbol{\xi}$ we obtain:

$$\boldsymbol{\xi} = \mathbf{x}_0 - [Df(\mathbf{x}_0)]^{-1} f(\mathbf{x}_0)$$

This equation gives us an iterative scheme to solve for $\mathbf{x}$:

$$\mathbf{x}^{(i+1)} = \mathbf{x}^{(i)} - \left[Df(\mathbf{x}^{(i)})\right]^{-1} f(\mathbf{x}^{(i)}) \tag{6.4}$$

Thus we can use the iterative procedure (6.4) to solve the Score equations (6.1) for the unknown parameter $\boldsymbol{\theta}$ and obtain an estimate $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$.

As the Score function is already the first derivative of the log-likelihood function, the Jacobi-matrix is simply the Hessian matrix of the log-likelihood function, denoted by $H(\boldsymbol{\theta})$. Using (6.4) our approach to solve the Score equations is the following:

$$\boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}^{(i)} - \left[H(\boldsymbol{\theta}^{(i)})\right]^{-1} S(\boldsymbol{\theta}^{(i)}) \tag{6.5}$$

6.1.2 Fisher-Scoring Method

As in many cases the Hessian matrix $H(\boldsymbol{\theta})$ in (6.5) still depends on the data, we use instead its expectation. This is the Fisher-Scoring Method as described by McCullagh and Nelder (1989). We define the matrix $A(\boldsymbol{\theta})$ as minus the information matrix, that is minus the expected value of the Hessian matrix.

$$A(\boldsymbol{\theta}) := -E[H(\boldsymbol{\theta})] = -E\left[\frac{\partial^2 l(\boldsymbol{\theta};\mathbf{y})}{\partial\boldsymbol{\theta}^2}\right]$$

Therefore (6.5) takes the form

$$\boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}^{(i)} + \left[A(\boldsymbol{\theta}^{(i)})\right]^{-1} S(\boldsymbol{\theta}^{(i)}) \tag{6.6}$$

Solutions of (6.6) can be obtained by iterative weighted least-squares, as we will show in the case of GLMs in Section 6.2.4.

6.1.3 Variance of the Maximum Likelihood Estimator $\hat{\boldsymbol{\theta}}$

Under mild regularity conditions one can show that in large samples $\hat{\boldsymbol{\theta}}$ is an asymptotically unbiased and efficient estimator (see Serfling (1980)). We call an estimator $\hat{\boldsymbol{\theta}}_n$ based on n observations asymptotically unbiased, if its expectation is equal to the true value, if n tends to infinity, that is

$$\lim_{n\to\infty} E[\hat{\boldsymbol{\theta}}_n] = \boldsymbol{\theta}$$

The estimator $\hat{\boldsymbol{\theta}}_n$ is said to be consistent, if it converges in probability to $\boldsymbol{\theta}$, that is

$$\lim_{n\to\infty} P(|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}| < \boldsymbol{\varepsilon}) = 1 \qquad \forall \boldsymbol{\varepsilon} > \mathbf{0}$$

An estimator $\hat{\boldsymbol{\theta}}$ is efficient, if we estimate the parameter $\hat{\boldsymbol{\theta}}$ with the asymptotically smallest possible covariance matrix V of $\hat{\boldsymbol{\theta}}$. This matrix is known as the inverse of the Fisher information matrix (McCulloch and Searle 2001):

$$V = -\left[E\frac{\partial^2 l(\boldsymbol{\theta};\mathbf{y})}{\partial\boldsymbol{\theta}^2}\right]^{-1} \tag{6.7}$$

It further follows that $\hat{\boldsymbol{\theta}}_n$ is asymptotically normally distributed, that is $\hat{\boldsymbol{\theta}}_n \sim AN(\boldsymbol{\theta}, V)$. We call a sequence of random variables $\boldsymbol{\theta}_n$ asymptotically normal with mean $\boldsymbol{\theta}$ and variance V, if $V > 0$ and for sufficiently large n the quantity $V^{-1}(\boldsymbol{\theta}_n - \boldsymbol{\theta})$ converges in distribution against a standard normally distributed random variable.

If we pursue the estimation of unknown probability patterns underlying observed data further and introduce covariates, this leads us to regression models. Aim here is, to describe the dependence structure of the mean response on the explanatory variables like in linear models or generalized linear models.

6.1.4 Generalized Linear Models (GLMs)

GLMs are a natural generalization of the classical linear models, originally introduced by Gauss and Legendre. They can be established by a three-part specification of the classical case, as done by McCullagh and Nelder (1989).

- The random component: Instead of having the response $y_i \overset{iid}{\sim} N\left(\mu_i, \sigma^2\right)$ we allow any distribution from the exponential family. A definition of the exponential family can be found in Appendix A.1.
- The systematic component: The vector of covariates $\boldsymbol{x_i} = (x_{i1}, \ldots, x_{ip})^T$ produces a linear predictor η_i given by $\eta_i = \boldsymbol{x_i}^T\boldsymbol{\beta}$, where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is the parameter vector to be estimated.
- The link between the random and systematic components: For the link function $g(\cdot)$, with $g(\mu_i) = \eta_i$ we allow other monotonic differentiable functions than the identity link $g(\cdot) = id(\cdot)$.

Objective for the Generalized Linear Regression is, to describe the relation between the mean response $\mu_i = E[y_i]$ and the covariates. The GLM, we want to fit for n independent observations y_i, is given by

$$g(\mu_i) = \boldsymbol{x_i}^T \boldsymbol{\beta} \qquad i = 1, \dots, n$$

Estimation for GLMs is generally done by minimizing a Goodness of fit measure, that assesses the fit between our observed values y_i and the fitted values $\hat{\mu}_i$, generated by our model, for example the deviance:

$$D(\hat{\boldsymbol{\mu}}; \mathbf{y}) = 2\left(l(\mathbf{y}; \mathbf{y}) - l(\hat{\boldsymbol{\mu}}; \mathbf{y})\right)$$

where $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \dots, \hat{\mu}_n)^T$ and $\mathbf{y} = (y_1, \dots, y_n)^T$. To calculate $l(\mathbf{y}; \mathbf{y})$ we take the observed values as fitted values. This is the saturated model. Thus $l(\mathbf{y}; \mathbf{y})$ is the highest value obtainable for the maximum likelihood function and does not depend on the unknown parameters.

It follows that maximizing the likelihood function is equivalent to minimizing the deviance, since $l(\mathbf{y}; \mathbf{y})$ is a constant. Further $l(\hat{\boldsymbol{\mu}}; \mathbf{y})$ is the log-likelihood function maximized over $\boldsymbol{\beta}$ in terms of the parameter $\hat{\boldsymbol{\mu}}$, which is obtained through the relationship $g(\hat{\mu}_i) = \boldsymbol{x_i}^T \hat{\boldsymbol{\beta}}$.

In the case of GLMs, the Maximum Likelihood Estimator of $\boldsymbol{\beta}$ is denoted by $\hat{\boldsymbol{\beta}}$ and can be obtained by solving the Score equations (6.1), which take the form

$$S(\boldsymbol{\beta}) = \frac{\partial l(\boldsymbol{\beta}; \mathbf{y})}{\partial \boldsymbol{\beta}} = \mathbf{0}$$

With the log-likelihood function $l(\boldsymbol{\theta}; \mathbf{y}) := \ln f(\mathbf{y}; \boldsymbol{\theta})$ as a function of $\boldsymbol{\theta}$ with fixed $\mathbf{y}$, we can derive expressions for the expectation of the first and second order derivative.

$$E\left[\frac{\partial l(\boldsymbol{\theta}; \mathbf{y})}{\partial \boldsymbol{\theta}}\right] = \mathbf{0} \qquad \text{and} \qquad E\left[\frac{\partial^2 l(\boldsymbol{\theta}; \mathbf{y})}{\partial \boldsymbol{\theta}^2}\right] + E\left[\left(\frac{\partial l(\boldsymbol{\theta}; \mathbf{y})}{\partial \boldsymbol{\theta}}\right)^2\right] = \mathbf{0} \tag{6.8}$$

A proof of both of these equations (6.8) can be found in Appendix A.1.

In the next section, where we follow Diggle, Liang, and Zeger (1996) and McCullagh and Nelder (1989), we are going to examine the maximum likelihood estimation in GLMs for observations that follow an exponential distribution. The strong assumptions on the actual form of the distribution function are relaxed in the section thereafter by introducing quasi-likelihood functions. In the quasi-likelihood approach one only specifies the relationship between mean and variance, via the variance function, and between mean and outcome, via the link function.

6.2 Maximum Likelihood Estimation (MLE) in GLMs

6.2.1 The Likelihood Function of a GLM

Given the density function $f(y_i, \theta_i, \phi)$ for one observation y_i from the exponential family, we form the log-likelihood as $l(\theta_i, \phi; y_i) := \ln f(y_i; \theta_i, \phi)$. Whereas the density function is seen as a function of y_i for fixed θ_i, the log-likelihood function is a function of θ_i for a given observation y_i and we express it as a function of the mean $\mu_i = E[y_i]$. Using the independence of the n observations $\mathbf{y} = (y_1, \ldots, y_n)^T$ with expectations $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T$ the log-likelihood is the sum of the individual contributions, that is

$$l(\boldsymbol{\theta}, \phi; \mathbf{y}) = \sum_{i=1}^{n} \left\{ \left(\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} \right) + c(y_i, \phi) \right\}$$

Note that the functions a, b and c are defined in Appendix A.1. Estimation proceeds in setting the derivative of the log-likelihood function zero. This gives us the Score equations that have now to be solved for the unknown parameter $\boldsymbol{\beta}$.

6.2.2 Score Equations for GLMs

In the case of the GLMs the Score equations can be expressed as follows, and the solution can be obtained by iterative weighted least-squares as we will see in Sections 6.2.4:

$$S_j(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial l}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial l}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j} \overset{!}{=} 0 \qquad j = 1, \ldots, p$$

Rearranging the Score equations we obtain for j = 1, ..., p

$$S_j(\boldsymbol{\beta}) = \frac{\partial l(\mathbf{y}; \boldsymbol{\beta}, \phi)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial l(y_i; \boldsymbol{\beta}, \phi)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial l(y_i; \boldsymbol{\beta}, \phi)}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j} \qquad \text{by the chain rule,}$$

where the second equation follows by independence of the y_i's. In the following we derive these derivatives; for the first one we get:

$$\frac{\partial l}{\partial \theta_i} = \frac{\partial l}{\partial \mu_i} \frac{\partial \mu_i}{\partial \theta_i} \qquad \Rightarrow \qquad \frac{\partial l}{\partial \mu_i} = \frac{\partial l}{\partial \theta_i} / \frac{\partial \mu_i}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{a(\phi)} / b''(\theta_i) = \frac{y_i - \mu_i}{v_i}$$

where μ_i and v_i are defined as

$$\mu_i := E[y_i] = b'(\theta_i) \qquad\qquad v_i := Var[y_i] = b''(\theta_i) \cdot a(\phi)$$

For the last derivative we obtain:

$$\frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial \boldsymbol{x_i}^T \boldsymbol{\beta}}{\partial \beta_j} = x_{ij}$$

We define a new variable w_i as

$$w_i := v_i^{-1} \left(\frac{\partial \eta_i}{\partial \mu_i} \right)^{-2} = v_i^{-1} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^{2}$$

Consequently the Score equations can be expressed as

$$S_j(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{y_i - \mu_i}{v_i} \left(\frac{\partial \mu_i}{\partial \eta_i}\right) x_{ij} = \sum_{i=1}^{n} w_i (y_i - \mu_i) \frac{\partial \eta_i}{\partial \mu_i} x_{ij} = 0 \qquad j = 1, \ldots, p \tag{6.9}$$

This is a system of non-linear equations in $\boldsymbol{\beta}$. As mentioned in Section 6.1 we can calculate its solution using the Newton-Raphson and Fisher-Scoring method.

6.2.3 Fisher-Scoring Method for GLMs

Solving the Score equations for $\hat{\boldsymbol{\beta}}$, we have to calculate the Jacobi-Matrix of $S(\boldsymbol{\beta})$, i.e. we need the Hessian Matrix $H(\boldsymbol{\beta})$ of the log-likelihood function. Lets define the matrix $A(\boldsymbol{\beta})$ as minus the information matrix, that is minus the expected value of the Hessian matrix:

$$A(\boldsymbol{\beta}) := -E[H(\boldsymbol{\beta})] = -E\left[\left(\frac{\partial^2 l}{\partial \beta_r \partial \beta_s}\right)_{s,r=1,\ldots,p}\right] \tag{6.10}$$

The matrix $H(\boldsymbol{\beta})$ has the following elements:

$$H(\boldsymbol{\beta}) := \begin{pmatrix} \frac{\partial^2 l}{\partial \beta_1 \partial \beta_1} & \cdots & \frac{\partial^2 l}{\partial \beta_p \partial \beta_1} \\ \cdots & \cdots & \cdots \\ \frac{\partial^2 l}{\partial \beta_1 \partial \beta_p} & \cdots & \frac{\partial^2 l}{\partial \beta_p \partial \beta_p} \end{pmatrix}$$

Therefore we calculate element (r, s) of the matrix $H(\boldsymbol{\beta})$, which is $\partial^2 l / \partial \beta_s \partial \beta_r$, and obtain

$$\begin{aligned} \frac{\partial^2 l}{\partial \beta_s \partial \beta_r} &= \frac{\partial}{\partial \beta_s}\left[\sum_{i=1}^{n} \frac{y_i - \mu_i}{v_i} \frac{\partial \mu_i}{\partial \eta_i} x_{ir}\right] \\ &= \sum_{i=1}^{n} (y_i - \mu_i) \frac{\partial}{\partial \beta_s}\left[v_i^{-1} \frac{\partial \mu_i}{\partial \eta_i} x_{ir}\right] + \sum_{i=1}^{n} \frac{\partial}{\partial \beta_s}(y_i - \mu_i)\left[v_i^{-1} \frac{\partial \mu_i}{\partial \eta_i} x_{ir}\right] \\ &= \sum_{i=1}^{n} (y_i - \mu_i) \frac{\partial}{\partial \beta_s}\left[v_i^{-1} \frac{\partial \mu_i}{\partial \eta_i} x_{ir}\right] - \sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_s}\left[v_i^{-1} \frac{\partial \mu_i}{\partial \eta_i} x_{ir}\right] \\ &= \sum_{i=1}^{n} (y_i - \mu_i) \frac{\partial}{\partial \beta_s}\left[v_i^{-1} \frac{\partial \mu_i}{\partial \eta_i} x_{ir}\right] - \sum_{i=1}^{n} v_i^{-1} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 x_{is} x_{ir} \end{aligned}$$

This expression depends in most cases still on the data. Taking expectations the first term vanishes, because $E[y_i - \mu_i] = 0$, while the second term becomes:

$$-E\left[\frac{\partial^2 l}{\partial \beta_s \partial \beta_r}\right] = E\left[\sum_{i=1}^{n} v_i^{-1} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 x_{is} x_{ir}\right] = E\left[\sum_{i=1}^{n} w_i x_{is} x_{ir}\right] \tag{6.11}$$

Thus inserting (6.11) into (6.10) we can write:

$$A(\boldsymbol{\beta}) = -E[H(\boldsymbol{\beta})] = X^T W X \tag{6.12}$$

where X is the design matrix with elements x_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, p$, and W is the $n \times n$ diagonal matrix with the diagonal elements given by $w_1, \ldots, w_n$.

As seen in Section 6.1.3 maximum likelihood gives us in large samples asymptotically unbiased and efficient estimators. Its variance is, according to (6.7), given by

$$V = -\left[E\frac{\partial^2 l(\boldsymbol{\beta};\mathbf{y})}{\partial\boldsymbol{\beta}^2}\right]^{-1} = \left(\sum_{i=1}^{n}\frac{\partial\mu_i}{\partial\boldsymbol{\beta}}v_i^{-1}\frac{\partial\mu_i}{\partial\boldsymbol{\beta}}\right)^{-1}$$

Consequently $\hat{\boldsymbol{\beta}}$ is asymptotically normally distributed, that is $\hat{\boldsymbol{\beta}} \sim AN(\boldsymbol{\beta}, V)$. Rewriting equation (6.11) we obtain for the elements of V^{-1}:

$$A_{rs} = -E\left[\frac{\partial^2 l}{\partial\beta_s\partial\beta_r}\right] = E\left[\sum_{i=1}^{n}v_i^{-1}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2 x_{is}x_{ir}\right] = E\left[\sum_{i=1}^{n}\frac{\partial\mu_i}{\partial\beta_s}v_i^{-1}\frac{\partial\mu_i}{\partial\beta_r}\right]$$

6.2.4 Iterative Weighted Least-Squares Algorithm for GLMs

The actual estimation using the Newton-Raphson and Fisher-Scoring method is done by an iterative algorithm. Given a current estimate $\mathbf{b}$ of $\boldsymbol{\beta}$ as a solution of the Score equations (6.9) we obtain a new estimate $\mathbf{b}^*$ using the iterative algorithm introduced in (6.6).

$$\mathbf{b}^* = \mathbf{b} + A^{-1}(\mathbf{b})S(\mathbf{b}) \qquad \Rightarrow \qquad A(\mathbf{b})\mathbf{b}^* = A(\mathbf{b})\mathbf{b} + S(\mathbf{b}) \tag{6.13}$$

The left side of the second expression in (6.13) gives:

$$\begin{aligned}\left(A(\mathbf{b})\mathbf{b}^*\right)_j &= \sum_{s=1}^{p}A_{js}(\mathbf{b})b_s^* \overset{(6.12)}{=} \sum_{s=1}^{p}\sum_{i=1}^{n}\left(w_i x_{ij} x_{is}\right)b_s^* \\ &= \sum_{i=1}^{n}w_i x_{ij}\underbrace{\sum_{s=1}^{p}x_{is}b_s^*}_{\eta_i^* := \boldsymbol{x_i}^T\mathbf{b}^*} = \sum_{i=1}^{n}w_i x_{ij}\eta_i^* \end{aligned} \tag{6.14}$$

The right side of the second expression in (6.13) gives:

$$\begin{aligned}\left(A(\mathbf{b})\mathbf{b} + S(\mathbf{b})\right)_j &= \sum_{s=1}^{p}A_{js}(b)b_s + S(b)_j \\ &\overset{(6.9),(6.12)}{=} \sum_{s=1}^{p}\sum_{i=1}^{n}w_i x_{ij}x_{is}b_s + \sum_{i=1}^{n}w_i(y_i-\mu_i)\frac{\partial\eta_i}{\partial\mu_i}x_{ij} \\ &= \sum_{i=1}^{n}w_i x_{ij}\left[\underbrace{\sum_{s=1}^{p}x_{is}b_s}_{\eta_i := \boldsymbol{x_i}^T\mathbf{b}} + (y_i-\mu_i)\frac{\partial\eta_i}{\partial\mu_i}\right] \\ &= \sum_{i=1}^{n}w_i x_{ij}\underbrace{\left[\eta_i + (y_i-\mu_i)\frac{\partial\eta_i}{\partial\mu_i}\right]}_{=:z_i} \\ &= \sum_{i=1}^{n}w_i x_{ij}z_i\end{aligned}$$

Putting the results from (6.14) and (6.15) together we finally obtain:

$$\sum_{i=1}^{n} w_i x_{ij} \eta_i^* = \sum_{i=1}^{n} w_i x_{ij} z_i \tag{6.15}$$

These equations have the form of linear weighted least-squares equations with weights w_i and dependent variate z_i (see Appendix A.3). Using first order Taylor expansion for $g(y_i)$ one sees that the dependent variable z_i is a linearized form of the link function applied to y_i.

$$g(y_i) \approx \underbrace{g(\mu_i)}_{\eta_i} + (y_i - \mu_i)\underbrace{g'(\mu_i)}_{\frac{\partial \eta_i}{\partial \mu_i}} = \eta_i + (y_i - \mu_i)\frac{\partial \eta_i}{\partial \mu_i} = z_i$$

Assuming that η_i and μ_i are fixed we get for the variance of z_i.

$$Var[z_i] = Var[y_i - \mu_i]\left(\frac{\partial \eta_i}{\partial \mu_i}\right)^2 = v_i \left(\frac{\partial \eta_i}{\partial \mu_i}\right)^2 = w_i^{-1}$$

Since the dependent variable z_i and the weights w_i depend both on the fitted values μ_i only current estimates are available, therefore the process is iterative. Usually the data itself can be used as a first estimate to start with, e.g. $\hat{\mu}_i^{(0)} := y_i$. Consequently we have the following algorithm:

- Let $\hat{\boldsymbol{\eta}}^{(j)}$ be the current estimate of the linear predictor vector with corresponding fitted value vector $\hat{\boldsymbol{\mu}}^{(j)}$, which has been defined as

$$\hat{\boldsymbol{\eta}}^{(j)} := g(\hat{\boldsymbol{\mu}}^{(j)}) = \boldsymbol{x_i}^T \hat{\boldsymbol{\beta}}^{(j)}$$

- Calculate with $\hat{\boldsymbol{\eta}}^{(j)}$ the adjusted dependent variate $\boldsymbol{z}^{(j)} = (z_1^{(j)}, \ldots, z_n^{(j)})^T$ and the weights $\boldsymbol{w}^{(j)} = (w_1^{(j)}, \ldots, w_n^{(j)})^T$, evaluating the derivative of the link function $\partial\eta_i/\partial\mu_i$ at $\mu_i = \hat{\mu}_i^{(j)}$, that is calculate the following quantities:

$$\begin{aligned} z_i^{(j)} &:= \hat{\eta}_i^{(j)} + \left(y_i - \hat{\mu}_i^{(j)}\right)\frac{\partial \eta_i}{\partial \mu_i}\Big|_{\mu_i = \hat{\mu}_i^{(j)}} \\ w_i^{(j)} &:= \frac{1}{v_i^{(j)}}\left(\frac{\partial \eta_i}{\partial \mu_i}\right)^{-2}\Big|_{\mu_i = \hat{\mu}_i^{(j)}} \end{aligned}$$

- Regress $\boldsymbol{z}^{(j)}$ on the covariates $x_{i1}, \ldots, x_{ip}$, $i = 1, \ldots, n$, with weights $\boldsymbol{w}^{(j)}$, to obtain a new estimate of the parameter $\boldsymbol{\beta}$, denoted by $\hat{\boldsymbol{\beta}}^{(j+1)}$

- Calculate a new estimate $\hat{\boldsymbol{\eta}}^{(j+1)}$ of the linear predictor, i.e.

$$\hat{\boldsymbol{\eta}}^{(j+1)} := g(\hat{\boldsymbol{\mu}}^{(j+1)}) = \boldsymbol{x_i}^T \hat{\boldsymbol{\beta}}^{(j+1)}$$

- Repeat until changes are sufficiently small, i.e. a satisfied degree of convergence is obatined:

$$\left\|\hat{\boldsymbol{\beta}}^{(j+1)} - \hat{\boldsymbol{\beta}}^{(j)}\right\| < \varepsilon \qquad \varepsilon > 0$$

6.3 Maximum Quasi-Likelihood Estimation (MQLE)

6.3.1 The Quasi-Likelihood Function

In above likelihood analysis we specified the whole distribution function of the y_i's by choosing a distribution from the exponential family. As we have seen, the Score function depends in this case only on the mean and variance of the y_i's (see equation (6.9)). The approach in which we make just assumptions about the link function and the variance function without attempting to specify the entire distribution of the y_i's leads us to the quasi-likelihood methods, as first proposed by Wedderburn (1974). We only specify the relationship between the mean and the variance and we allow any choice of link function. As in the likelihood case, the Score function then only depends on the mean and variance of the y_i's. We will see that the quasi-likelihood function can be used for estimation in the same way as the likelihood function.

Wedderburn (1974) assumed independent observations y_i, $i = 1, \ldots, n$ with expectation μ_i and variance $V(\mu_i)$, where V is a known function of μ_i. Further μ_i is a function of unknown parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$. In the case of GLMs this function is called the mean function $\mu_i = g^{-1}(\boldsymbol{x_i}^T\boldsymbol{\beta})$.

For each observation the quasi-likelihood function, denoted by $K(y_i, \mu_i)$, is defined by the relation

$$\frac{\partial K(y_i, \mu_i)}{\partial \mu_i} = \frac{y_i - \mu_i}{V(\mu_i)}$$

or equivalently

$$K(y_i, \mu_i) = \int_{y_i}^{\mu_i} \frac{y_i - t}{V(t)} dt$$

This function can be interpreted as a log-likelihood function. Thus for n independent observations we obtain the following function:

$$K(\mathbf{y}, \boldsymbol{\mu}) = \sum_{i=1}^{n} \int_{y_i}^{\mu_i} \frac{y_i - t}{V(t)} dt$$

Above defined function K shares many properties with the likelihood function. One can show that K is the log-likelihood function, if y has an one-parameter distribution from the exponential family, which we will proof later in Section 6.3.2. But first we have a look to certain derivatives of the quasi-likelihood function:

$$\begin{aligned}
E\left[\frac{\partial K(y_i, \mu_i)}{\partial \mu_i}\right] &= E\left[\frac{y_i - \mu_i}{V(\mu_i)}\right] \\
&= \frac{1}{V(\mu_i)} E\left[y_i - \mu_i\right] = 0 \qquad i = 1, \ldots, n \\
E\left[\frac{\partial K(y_i, \mu_i)}{\partial \beta_j}\right] &= E\left[\frac{\partial K(y_i, \mu_i)}{\partial \mu_i}\frac{\partial \mu_i}{\partial \beta_j}\right] = E\left[\frac{y_i - \mu_i}{V(\mu_i)}\frac{\partial \mu_i}{\partial \beta_j}\right] \\
&= \frac{\partial \mu_i}{\partial \beta_j} E\left[\frac{y_i - \mu_i}{V(\mu_i)}\right] = 0 \qquad j = 1, \ldots, p \\
E\left[\left(\frac{\partial K(y_i, \mu_i)}{\partial \mu_i}\right)^2\right] &= E\left[\left(\frac{y_i - \mu_i}{V(\mu_i}\right)^2\right] = \frac{1}{V(\mu_i)^2} E\left[(y_i - \mu_i)^2\right] = \frac{1}{V(\mu_i)}
\end{aligned}$$

$$\begin{aligned}
-E\left[\frac{\partial^2 K(y_i,\mu_i)}{\partial\mu_i^2}\right] &= -E\left[\frac{\partial}{\partial\mu_i}\left(\frac{y_i-\mu_i}{V(\mu_i)}\right)\right] = -E\left[-\frac{1}{V(\mu_i)}\right] = \frac{1}{V(\mu_i)} \\
E\left[\frac{\partial K}{\partial\beta_r}\frac{\partial K}{\partial\beta_s}\right] &= E\left[\left(\frac{\partial K(y_i,\mu_i)}{\partial\mu_i}\right)^2\frac{\partial\mu_i}{\partial\beta_r}\frac{\partial\mu_i}{\partial\beta_s}\right] \\
&= E\left[\left(\frac{y_i-\mu_i}{V(\mu_i)}\right)^2\frac{\partial\mu_i}{\partial\beta_r}\frac{\partial\mu_i}{\partial\beta_s}\right] = \frac{1}{V(\mu_i)}\frac{\partial\mu_i}{\partial\beta_r}\frac{\partial\mu_i}{\partial\beta_s} \qquad (6.16) \\
-E\left[\frac{\partial^2 K}{\partial\beta_r\partial\beta_s}\right] &= -E\left[\frac{\partial}{\partial\beta_s}\left(\frac{y_i-\mu_i}{V(\mu_i)}\frac{\partial\mu_i}{\partial\beta_r}\right)\right] \\
&= -E\left[(y_i-\mu_i)\frac{\partial}{\partial\beta_s}\left(\frac{1}{V(\mu_i)}\frac{\partial\mu_i}{\partial\beta_r}\right) - \frac{1}{V(\mu_i)}\frac{\partial\mu_i}{\partial\beta_r}\frac{\partial\mu_i}{\partial\beta_r}\right] \\
&= \frac{1}{V(\mu_i)}\frac{\partial\mu_i}{\partial\beta_i}\frac{\partial\mu}{\partial\beta_j} \qquad (6.17)
\end{aligned}$$

Using these results one easily sees that the same properties hold as in the case of log-likelihood, which we studied in (6.8), namely

$$E\left[\frac{\partial K(y_i,\mu_i)}{\partial\mu_i}\right] = E\left[\frac{y_i-\mu_i}{V(\mu_i)}\right] = 0$$

$$\Rightarrow E\left[y_i\right] = \mu_i$$

$$E\left[\frac{\partial^2 K(y_i,\mu_i)}{\partial\mu_i^2}\right] + E\left[\left(\frac{\partial K(y_i,\mu_i)}{\partial\mu_i}\right)^2\right] = -\frac{1}{V(\mu_i)} + E\left[\left(\frac{y_i-\mu_i}{V(\mu_i)}\right)^2\right] = 0$$

$$\Rightarrow \qquad Var\left[y_i\right] = V(\mu_i)$$

6.3.2 Quasi-Likelihood of Exponential Families

Assuming that Y has an one-parameter family of distributions with μ as parameter we can define a log-likelihood function. This log-likelihood function is, according to Wedderburn (1974), identical to the quasi-likelihood function if and only if this family is an exponential family.

Theorem 6.1 *For one observation y, the log-likelihood function L has the property*

$$\frac{\partial L}{\partial\mu} = \frac{y-\mu}{V(\mu)}$$

where $\mu := E[Y]$ and $V(\mu) := Var[Y]$ if and only if the density of y with respect to some measure can be written in the form $exp\{\theta y - b(\theta)\}$, where θ is some function of μ.

Proof:

”$\Longrightarrow$”

If $\partial L/\partial\mu = (y-\mu)/V(\mu)$, integrating with respect to μ leads to

$$l(\mu;y) = \int\frac{\partial L}{\partial\mu}d\mu = \int\frac{y-\mu}{V(\mu)}d\mu = \int\frac{y}{V(\mu)}d\mu - \int\frac{\mu}{V(\mu)}d\mu = \theta y - b(\theta)$$

where $\theta := \int d\mu/V(\mu)$ and the function b is defined as a function of θ, which itself is a function of μ, that is $b(\theta) := \int \mu/V(\mu)d\mu$.

"$\Longleftarrow$"
Suppose that for some measure m, the distribution of Y is given by $\exp\{y\theta - b(\theta)\}\, dm(y)$. The definition of a density function gives $\int e^{y\theta}e^{-b(\theta)}dm(y) = 1$. Therefore we obtain $\int e^{y\theta}dm(y) = e^{b(\theta)}$. Thus the moment generating function $\Psi(t)$ of Y is

$$\Psi(t) = E\left[e^{ty}\right] = \int e^{ty}e^{\theta y}e^{-b(\theta)}dm(y) = e^{-b(\theta)}\int e^{y(\theta+t)}dm(y) = e^{b(\theta+t)-b(\theta)}$$

Expectation and variance can be represented with the 1^{st} and 2^{nd} order derivatives of the moment generating function $\Psi(t)$ evaluated at zero. Thus the following relationship holds:

$$\begin{aligned} \Psi'(0) &= E[Y] \\ \Psi''(0) - \left(\Psi'(0)\right)^2 &= Var[Y] \end{aligned}$$

This gives us here

$$\begin{aligned} \Psi(s) &= \exp\{\ b(\theta+s) - b(\theta)\} \\ \Psi'(s) &= b'(\theta+s)\Psi(s) \\ &\Rightarrow \qquad E[Y] = \Psi'(0) = b'(\theta) \\ \Psi''(s) &= b''(\theta+s)\Psi(s) + \left(b'(\theta+s)\right)^2\Psi(s) \\ &\Rightarrow \qquad Var[Y] = \Psi''(0) - (\Psi'(0)^2) = b''(\theta) \end{aligned}$$

We set $\mu := b'(\theta)$, $V(\mu) := b''(\theta)$, then $\partial\mu/\partial\theta = \partial b'(\theta)/\partial\theta = b''(\theta) = V(\mu)$. Inserting this in the log-likelihood function we get:

$$\frac{\partial L}{\partial\mu} = \frac{\partial L}{\partial\theta}\frac{\partial\theta}{\partial\mu} = \frac{\partial}{\partial\theta}\left(y\theta - b(\theta)\right)\frac{\partial\theta}{\partial\mu} = \left(y - b'(\theta)\right)\frac{\partial\theta}{\partial\mu} = \frac{y-\mu}{V(\mu)}$$

□

As seen in Theorem 6.1 the log-likelihood and the quasi-likelihood are identical for an one-parameter exponential family. The information matrix $-E\left[\partial^2 L/\partial\mu^2\right]$ is therefore minimized and its value is equal to $-E\left[\partial^2 K/\partial\mu^2\right]$. Generally speaking we have from the Cramér-Rao inequality (see Appendix A.5) and using $-E\left[\partial^2 K/\partial\mu^2\right] = 1/Var(Y)$ the following relationship:

$$\begin{aligned} Var\left[Y\right] \geq -1/E\left[\frac{\partial^2 L}{\partial\mu^2}\right] \qquad &\Rightarrow\ -1/E\left[\frac{\partial^2 K}{\partial\mu^2}\right] \geq -1/E\left[\frac{\partial^2 L}{\partial\mu^2}\right] \\ &\Rightarrow\ -E\left[\frac{\partial^2 K}{\partial\mu^2}\right] \leq -E\left[\frac{\partial^2 L}{\partial\mu^2}\right] \qquad (6.18) \end{aligned}$$

One can interpret $-E\left[\partial^2 K/\partial\mu^2\right]$ as the information Y gives concerning μ, when only the mean-variance relationship is known.

From above it follows that $E\left[\partial^2(K-L)/\partial\mu^2\right]$ is always non-negative and we can regard it as the additional information knowing the distribution of Y. Therefore assuming an one-parameter exponential family for Y is equivalent to making no assumption other than the mean-variance relation, since in this case $-E\left[\partial^2(K-L)/\partial\mu^2\right] = 0$.

6.3.3 Score-like Equations for Quasi-Likelihood

As already seen the quasi-likelihood function shares many properties with the log-likelihood function. This holds also true for the way how we estimate regression parameters and judge their precision. The estimator $\hat{\boldsymbol{\beta}}_{QL}$ for $\boldsymbol{\beta}$ is defined to be the solution of

$$\begin{aligned}
SL_j(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \frac{\partial K(y_i;\mu_i)}{\partial \beta_j} \qquad j = 1,\dots,p \\
&= \sum_{i=1}^{n} \frac{\partial K(y_i;\mu_i)}{\partial \mu_i}\frac{\partial \mu_i}{\partial \beta_j} \\
&= \sum_{i=1}^{n} \frac{y_i - \mu_i}{V(\mu_i)}\frac{\partial \mu_i}{\partial \beta_j} \\
&= \sum_{i=1}^{n} \frac{r_i v_j}{V(\mu_i)} = 0 \qquad j = 1,\dots,p \qquad (6.19)
\end{aligned}$$

where $r_i := y_i - \mu_i$ and $v_j = \partial\mu_i/\partial\beta_j$. This equation reduces to the Score function if the distribution of the y_i's is from the exponential family. This can be easily understood, comparing expression (6.9) with (6.19).

6.3.4 Fisher-Scoring Method for Quasi-Likelihood

In the maximum likelihood case we estimated the precision of the estimator with the expectation of the matrix of 2^{nd} order derivatives. For the maximum quasi-likelihood function K we do likewise, and denote the matrix of expected 2^{nd} order derivatives by H_{QL}. Further we define the matrix D as minus the expectation of H_{QL}.

$$D := -E[H_{QL}] = -E\left[\left(\frac{\partial^2 K}{\partial\beta_i\partial\beta_j}\right)_{s,r=1,\dots,p}\right]$$

the $(r,s)^{th}$ element of D can be calculated using (6.17) as follows:

$$\begin{aligned}
D_{rs} &= -E\left[\frac{\partial^2 \sum_{i=1}^{n} K(y_i;\mu_i)}{\partial\beta_s\partial\beta_r}\right] = -E\left[\frac{\partial}{\partial\beta_s}\sum_{i=1}^{n}\frac{y_i-\mu_i}{V(\mu_i)}v_r\right] \\
&= -E\left[\sum_{i=1}^{n}\frac{\partial}{\partial\beta_s}(y_i-\mu_i)\frac{v_r}{V(\mu_i)} + \sum_{i=1}^{n}(y_i-\mu_i)\frac{\partial}{\partial\beta_s}\frac{v_r}{V(\mu_i)}\right] \\
&= -E\left[\sum_{i=1}^{n} -\frac{v_r}{V(\mu_i)}\frac{\partial\mu_i}{\partial\beta_s}\right] = \sum_{i=1}^{n}\frac{v_r v_s}{V(\mu_i)} \qquad (6.20)
\end{aligned}$$

Using first order Taylor expansion for $\hat{\boldsymbol{\beta}}_{QL}$ we obtain the following equation with the Newton-Raphson method (see Section 6.1.1):

$$\hat{\boldsymbol{\beta}}_{QL} = \boldsymbol{\beta} - H_{QL}^{-1}\sum_{i=1}^{n}\frac{\partial K(y_i;\mu_i)}{\partial\beta_j}$$

As required by the Fisher-scoring we replace H_{QL} by it's expectation $-D$ and obtain:

$$\hat{\boldsymbol{\beta}}_{QL} = \boldsymbol{\beta} + D^{-1} \sum_{i=1}^{n} \frac{\partial K(y_i; \mu_i)}{\partial \beta_j} \tag{6.21}$$

this is analogue to the maximum likelihood approach. Therefore the following theorem holds:

Theorem 6.2 *Maximum quasi-likelihood estimate $\hat{\boldsymbol{\beta}}_{QL}$ has approximate covariance matrix*

$$D^{-1} = E[H_{QL}^{-1}],$$

where H_{QL} is the matrix of 2^{nd} order derivatives of $\sum_{i=1}^{n} K(y_i; \mu_i)$.

6.3.5 Iterative Weighted Least-Squares Algorithm for Quasi-Likelihood

As in the last section an iterative algorithm can be obtained using the Newton-Raphson and Fisher-Scoring method. Given a current estimate $\mathbf{b}$ of $\hat{\boldsymbol{\beta}}_{QL}$ we obtain a new estimate $\mathbf{b}^*$ using an iterative procedure. Updating (6.21) we get:

$$D(\mathbf{b})\mathbf{b}^* = D(\mathbf{b})\mathbf{b} + \sum_{i=1}^{n} \frac{\partial K(y_i; \mu_i)}{\partial \beta_j} \tag{6.22}$$

The left side of (6.22) gives:

$$\begin{aligned} \left(D(\mathbf{b})\mathbf{b}^*\right)_j &= \sum_{s=1}^{p} D_{js}(\mathbf{b}) b_s^* \overset{(6.20)}{=} \sum_{s=1}^{p} \sum_{i=1}^{n} \frac{v_j v_s}{V(\mu_i)} b_s^* \\ &= \sum_{i=1}^{n} \frac{v_j}{V(\mu_i)} \sum_{s=1}^{p} v_s b_s^* \end{aligned} \tag{6.23}$$

The right side of (6.22) gives:

$$\begin{aligned} \left(D(\mathbf{b})\mathbf{b} + \sum_{i=1}^{n} \frac{\partial K(y_i; \mu_i)}{\partial \beta_j} \right)_j &= \sum_{s=1}^{p} D_{js}(\mathbf{b}) b_s + \sum_{i=1}^{n} \frac{\partial K(y_i, \mu_i)}{\partial \beta_j} \\ &= \sum_{s=1}^{p} \sum_{i=1}^{n} \frac{v_j v_s}{V(\mu_i)} b_s + \sum_{i=1}^{n} \frac{(y_i - \mu_i)}{V(\mu_i)} \frac{\partial \mu_i}{\partial \beta_j} \\ &= \sum_{i=1}^{n} \frac{v_j}{V(\mu_i)} \sum_{s=1}^{p} v_s b_s + \sum_{i=1}^{n} (y_i - \mu_i) \frac{v_j}{V(\mu_i)} \end{aligned} \tag{6.24}$$

Putting (6.23) and (6.24) together we obtain:

$$\sum_{i=1}^{n} \frac{v_j}{V(\mu_i)} \sum_{s=1}^{p} v_s b_s^* = \sum_{i=1}^{n} \frac{v_j}{V(\mu_i)} \sum_{s=1}^{p} v_s b_s + \sum_{i=1}^{n} (y_i - \mu_i) \frac{v_j}{V(\mu_i)}$$

$$\sum_{s=1}^{p} \sum_{i=1}^{n} \frac{v_j v_s}{V(\mu_i)} \underbrace{(b_s^* - b_s)}_{=:\delta b_s} = \sum_{i=1}^{n} (y_i - \mu_i) \frac{v_j}{V(\mu_i)} \tag{6.25}$$

This means that, if we obtain successive approximations to $\hat{\boldsymbol{\beta}}_{QL}$ using the Newton-Raphson method with the 2^{nd} order derivatives of K replaced by their expectations, we obtain corrections $\delta\mathbf{b} = \mathbf{b}^* - \mathbf{b}$ to the estimates. We are summarizing these results in the following theorem:

Theorem 6.3 *Using the Newton-Raphson method with expected second derivatives K to calculate $\hat{\boldsymbol{\beta}}_{QL}$ is equivalent to calculating iteratively a weighted linear regression of the residuals, $r_i := y_i - \mu_i$, on the derivatives of μ_i with respect to the $\boldsymbol{\beta}$'s with weight $1/V(\mu_i)$, and using the regression coefficients as correction to $\hat{\boldsymbol{\beta}}$. Here $V(\mu_i)$ and the derivatives of μ_i are calculated at the current estimate of $\hat{\boldsymbol{\beta}}_{QL}$.*

Therefore we obtain the following algorithm:

- Let $\hat{\boldsymbol{\beta}}_{QL}^{(l)}$ be the current estimate for $\boldsymbol{\beta}$
- Regress the residuals $r_i^{(l)} := y_i - \mu_i^{(l)}$ on the derivatives of μ_i with respect to the $\boldsymbol{\beta}$'s, denoted by v_j, with weight $1/V(\mu_i)$, evaluated at the current estimate $\hat{\boldsymbol{\beta}}_{QL}^{(l)}$.

$$v_j := \frac{\partial \mu_i}{\partial \beta_j}\bigg|_{\beta_j = \hat{\beta}_{QL_j}^{(l)}}$$

- Use the regression coefficients to calculate a corrected estimate for $\boldsymbol{\beta}$, denoted by $\hat{\boldsymbol{\beta}}_{QL}^{(l+1)}$

To see the strong relationship between equation (6.15) and (6.25) we assume that a GLM holds and $\boldsymbol{\eta} = \boldsymbol{x_i}^T\boldsymbol{\beta}$. We replace the derivative

$$v_j = \frac{\partial \mu_i}{\partial \beta_j} \quad \text{by} \quad \frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_i} x_{ij}$$

thus we can rewrite (6.25) as

$$\begin{aligned}
\sum_{s=1}^{p}\sum_{i=1}^{n} \frac{v_j v_s}{V(\mu_i)}(b_s^* - b_s) &= \sum_{i=1}^{n}(y_i - \mu_i)\frac{v_j}{V(\mu_i)} \\
\sum_{s=1}^{p}\sum_{i=1}^{n} \frac{1}{V(\mu_i)}\frac{\partial \mu_i}{\partial \eta_i}x_{ij}\frac{\partial \mu_i}{\partial \eta_i}x_{is}\,(b_s^* - b_s) &= \sum_{i=1}^{n}(y_i - \mu_i)\frac{1}{V(\mu_i)}\frac{\partial \mu_i}{\partial \eta_i}x_{ij} \\
\sum_{i=1}^{n} \frac{1}{V(\mu_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 x_{ij}\sum_{s=1}^{p} x_{is}\,(b_s^* - b_s) &= \sum_{i=1}^{n}\frac{y_i - \mu_i}{V(\mu_i)}\frac{\partial \mu_i}{\partial \eta_i}x_{ij} \\
\text{again using } w_i &:= \frac{1}{V(\mu_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 \\
\sum_{i=1}^{n} w_i x_{ij}\,(\eta_i^* - \eta_i) &= \sum_{i=1}^{n} w_i\,(y_i - \mu_i)\,\frac{\partial \eta_i}{\mu_i}x_{ij} \\
\sum_{i=1}^{n} w_i x_{ij}\eta_i^* &= \sum_{i=1}^{n} w_i x_{ij}\underbrace{\left(\eta_i + (y_i - \mu_i)\,\frac{\partial \eta_i}{\partial \mu_i}\right)}_{=:z_i} \qquad (6.26)
\end{aligned}$$

This is equal to equation (6.15) and the solution here can be obtained in exactly the same way.

6.4 Estimation Methods for Correlated Data

6.4.1 Introduction

So far we had one observation for each subject and assumed these observations to be independent, as they were from different subjects. If we study more than one observation from the same subject or take correlation between subjects into account this assumption can no longer hold. These and similar questions where studied by Diggle, Liang, and Zeger (1996) for longitudinal studies.

In population studies one distinguishes between cross-sectional studies and longitudinal studies. In cross-sectional studies the outcome for each individual is measured once, whereas in longitudinal studies individuals are measured repeatedly through time. Therefore longitudinal studies can distinguish changes over time within individuals, known as age effects (e. g. growth), from differences among people in their baseline levels, the so-called cohort effects. Clearly, independence between subjects can be assumed in both, cross-sectional studies and longitudinal studies, but in the latter one has to account for correlation within the outcomes of a subject to obtain a correct statistical model.

Therefore we need a model for the joint distribution of the repeated observations. When the outcome variable is approximately Gaussian, statistical methods for longitudinal data haven been developed, as mentioned by Liang and Zeger (1986). For non-Gaussian outcomes however, less development has taken place and few models are available. It seems therefore obvious to use quasi-likelihood methods, where only the relation between mean and covariates in addition to the relation between mean and variance has to be specified. This is in contrast to likelihood methods, where we specify the whole joint distribution of the outcome variables.

Zeger and Liang (1986) describe the quasi-likelihood approach as a model where a known transformation of the marginal expectation of the outcome (link function) is assumed to be a linear function of the covariates. Instead of specifying the joint distribution of dependent variables, its variance is assumed a known function of its expectation (variance function). In addition a ”working” correlation matrix for the observations of each subject is specified.

This setup leads us to the so-called Generalized Estimating Equations (GEEs) as proposed by Liang and Zeger (1986) and Zeger and Liang (1986). Given that the regression model is correctly specified these equations give under mild assumptions on the time dependence, even when the time dependence is misspecified - as often is expected, consistent estimates of the regression coefficients and their variances as well as asymptotically normally distributed estimates. Furthermore the GEEs reduce to the Score equations for multivariate Gaussian outcomes and can therefore be used in an analogue way.

To see the strong relation between the equations obtained using GLMs and the GEEs, we start with a working model in which the ”independence working” assumption holds. We assume that the marginal distribution of the dependent variable follows a GLM. Repeated observations are assumed to be independent, that means, the ”working” correlation matrix is identical to the identity matrix. Later we will generalize this independent working model to explicitly account for correlation, giving us GEEs estimates.

6.4.2 Models

Let us assume the following data situation: $(y_{it}; \mathbf{x}_{it})$ are observed at times $t = 1, \ldots, n_i$ for subjects $i = 1, \ldots, n$, hereby

- y_{it} is the response with mean μ_{it} and variance v_{it}
- $\mathbf{x}_{it}$ is the $1 \times p$ vector of covariates $(x_{it1}, \ldots, x_{itp})$
- $\mathbf{y}_i$ is the $n_i \times 1$ vector, that collects the outcomes of subject i: $(y_{i1}, \ldots, y_{in_i})^T$ with mean $\boldsymbol{\mu}_i$ and variance $\mathbf{v}_i$
- $\mathbf{X}_i$ is the $n_i \times p$ matrix, that collects the covariates of subject i: $(\mathbf{x}_{i1}, \ldots, \mathbf{x}_{in_i})^T$

Finally, $\mathbf{y}$ is the vector of length $N = \sum_{i=1}^{n} n_i$, that collects the outcomes of all subjects $i = 1, \ldots, n$. The unknown vector, to be estimated is $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$.

To summarize the above, we have for subject i the following data:

$$\mathbf{y}_i = \begin{pmatrix} y_{i1} \\ \ldots \\ y_{in_i} \end{pmatrix}; \qquad \mathbf{X}_i = \begin{pmatrix} \mathbf{x}_{i1} \\ \ldots \\ \mathbf{x}_{in_i} \end{pmatrix} = \begin{pmatrix} x_{i11} \ldots x_{i1p} \\ \ldots\ldots\ldots \\ x_{in_i1} \ldots x_{in_ip} \end{pmatrix} \quad \in \mathbb{R}^{n_i \times p}$$

When convenient we let $n_i = n$ for all subjects without loss of generality to simplify notation. If $n_i = 1$ for all i, we have nothing else than the situation of quasi-likelihood of Section 6.3. Additionally, if we specify the distribution of y_{it} with a distribution from the exponential family we have the situation of the classical likelihood, as discussed in Section 6.2.

Liang and Zeger (1986) and Zeger and Liang (1986) discussed two different approaches in deriving the GEEs. The first one assumes a distribution from the exponential family for the y_{it}'s and starts with likelihood methods, whereas the second one only specifies the link and variance function and thus, starts with quasi-likelihood methods. Both methods derive the same set of GEEs and differ only in the additional specification of the distribution function in the first case. However, the equations obtained in the second case, reduce to the equations, derived by the first approach, if one takes a distribution from the exponential family.

Let y_{it} have any distribution from the exponential family:

$$f(y_{it}) = \exp\left\{(y_{it}\theta_{it} - b(\theta_{it}) + c(y_{it}))\,\phi\right\} \tag{6.27}$$

where $\theta_{it} := h(\eta_{it})$ and $h(\cdot)$ will be defined below.

As seen in Section 6.2 we have using (6.27)

$$\begin{aligned} \mu_{it} &:= E[y_{it}] = b'(\theta_{it}) \\ v_{it} &:= Var[y_{it}] = b''(\theta_{it})/\phi \end{aligned}$$

where ϕ is the scale parameter and $\boldsymbol{\mu}_i$ a $n_i \times 1$ vector, containing the elements $\mu_{i1}, \ldots, \mu_{in_i}$.

Using the link function and above relationship we obtain:

$$\begin{array}{lcl} g(\mu_{it}) = \eta_{it} & \Longrightarrow & \mu_{it} = g^{-1}(\eta_{it}) \\ \mu_{it} = b'(\theta_{it}) & \Longrightarrow & \theta_{it} = b'^{-1}(\mu_{it}) := m(\mu_{it}) \end{array}$$

If we take both together, this leads to

$$\theta_{it} = m(\mu_{it}) = m(g^{-1}(\eta_{it})) := h(\eta_{it})$$

Further

$$\eta_{it} := \mathbf{x}_{it}\boldsymbol{\beta} = x_{it1}\beta_1 + \ldots + x_{itp}\beta_p$$

Recall that x_{it} is already a $1 \times p$ vector and therefore does not need to be transposed. This is in contrast to the notation used for GLMs and accounts for the additional dimension over time. In an analogue way to $\mathbf{y}_i$ we define $\boldsymbol{\mu}_i$ and $\boldsymbol{\theta}_i$ as the $n_i \times 1$ vectors of the elements $\eta_{i1}, \ldots, \eta_{in_i}$ and $\theta_{i1}, \ldots, \theta_{in_i}$, respectively.

The two quantities μ_{it} and v_{it} are specified in the quasi-likelihood approach as

$$\mu_{it} := h(\mathbf{x}_{it}\boldsymbol{\beta})$$

$$v_{it} := k(\mu_{it})/\phi$$

In the following the actual specification of μ_{it} and v_{it} does not matter. We treat both cases in only referring to μ_{it} and v_{it}, ignoring the actual assumption on the underlying probability density: A distribution from the exponential family or the quasi-likelihood approach. To obtain the first or the second case it is only necessary to substitute the relevant quantities for μ_{it} and v_{it}, respectively. In both cases the following relationship holds:

$$\frac{\partial l(y_{it}, \mu_{it})}{\partial \mu_{it}} = \frac{y_{it} - \mu_{it}}{v_{it}}$$

For the quasi-likelihood approach this is true by definition, in the case of a distribution from the exponential family, we can obtain the same relationship:

$$\frac{\partial l(y_{it}, \mu_{it})}{\partial \theta_{it}} = \frac{\partial l(y_{it}, \mu_{it})}{\partial \mu_{it}} \frac{\partial \mu_{it}}{\partial \theta_{it}} \qquad \Rightarrow \qquad \frac{\partial l(y_{it}, \mu_{it})}{\partial \mu_{it}} = \frac{\partial l(y_{it}, \mu_{it})}{\partial \theta_{it}} / \frac{\partial \mu_{it}}{\partial \theta_{it}}$$

Applying this to the definition of the exponential family (6.27) we obtain:

$$\frac{\partial l(y_{it}, \mu_{it})}{\partial \mu_{it}} = \frac{(y_{it} - \mu_{it})\phi}{b''(\theta_{it})} = \frac{y_{it} - \mu_{it}}{v_{it}}$$

This is of the same form as required by the definition of the quasi-likelihood function. So in the next section we can use μ_{it} and v_{it} in both cases without loss of generality.

6.4.3 Independence Estimating Equations

Now we assume additionally that though observations come from the same subject (repeated observations of one subject), they are nevertheless independent of one another. We call this the "independence working" assumption.

The estimator $\hat{\boldsymbol{\beta}}_{\boldsymbol{I}} = (\hat{\beta}_1^I, \ldots, \hat{\beta}_p^I)^T$ is defined to be the solution of the Score-like equations from likelihood analysis given below. Besides $\boldsymbol{\beta}$ also the quantity ϕ has to be estimated. As the Score equations are not affected by ϕ, we can proceed by setting ϕ to one. As observations are assumed independent, we derive the Score equations in the case of the exponential family as

$$\begin{aligned} U_I(\boldsymbol{\beta}) &= \sum_{i=1}^{n}\sum_{t=1}^{n_i} \frac{\partial l(y_{it};\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \\ &= \sum_{i=1}^{n}\sum_{t=1}^{n_i} \frac{\partial l}{\partial \theta_{it}} \frac{\partial \theta_{it}}{\partial \mu_{it}} \frac{\partial \mu_{it}}{\partial \eta_{it}} \frac{\partial \eta_{it}}{\partial \boldsymbol{\beta}} \\ &= \sum_{i=1}^{n}\sum_{t=1}^{n_i} \frac{\partial l}{\partial \theta_{it}} \frac{\partial \theta_{it}}{\partial \eta_{it}} \frac{\partial \eta_{it}}{\partial \boldsymbol{\beta}} = \mathbf{0} \end{aligned} \tag{6.28}$$

Defining $\Delta_i := diag(\partial\theta_{it}/\partial\eta_{it})$, that is a $n_i \times n_i$ diagonal matrix with diagonal elements $\partial\theta_{it}/\partial\eta_{it}$, where $t = 1, \ldots, n_i$ and $S_i := \mathbf{y}_i - \boldsymbol{\mu}_i \in \mathbb{R}^{n_i \times 1}$ and using $\eta_{it} := \mathbf{x}_{it}\boldsymbol{\beta}$ we obtain

$$\begin{aligned} U_I(\boldsymbol{\beta}) &= \sum_{i=1}^{n}\sum_{t=1}^{n_i} \frac{\partial l(y_{it};\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \\ &= \sum_{i=1}^{n}\sum_{t=1}^{n_i} (y_{it} - \mu_{it}) \frac{\partial \theta_{it}}{\partial \eta_{it}} \mathbf{x_{it}}^T \\ &= \sum_{i=1}^{n} \mathbf{X}_i^T diag\left(\frac{\partial \theta_{ij}}{\partial \eta_{ij}}\right)_{j=1,\ldots,n_i} (\mathbf{y}_i - \boldsymbol{\mu}_i) \\ &= \sum_{i=1}^{n} \mathbf{X}_i^T \Delta_i S_i = \mathbf{0} \end{aligned} \tag{6.29}$$

Further we define $A_i := diag(\mathbf{v}_i\phi)$ for each i, which is equal to $diag(b''(\theta_{it}))$ in the case of the exponentially family and equal to $diag(k(\mu_{it}))$ in the case of the quasi-likelihood. Nevertheless, in both cases, independent of the parameter ϕ. One can show that under mild regularity conditions the following theorem holds. The proof of this theorem is omitted, as this is a special case of Theorem 6.5 we are going to prove in Section 6.4.5.

In the case of quasi-likelihood the following Score equations are derived:

$$\begin{aligned} U_I(\boldsymbol{\beta}) &= \sum_{i=1}^{n}\sum_{t=1}^{n_i} \frac{\partial l(y_{it},\boldsymbol{\beta})}{\partial \mu_{it}} \frac{\partial \mu_{it}}{\partial \boldsymbol{\beta}} \\ &= \sum_{i=1}^{n}\sum_{t=1}^{n_i} \frac{y_{it} - \mu_{it}}{v_{it}} \frac{\partial h(\mathbf{x}_{it}\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \mathbf{x}_{it}^T \\ &= \sum_{i=1}^{n} \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}\right)^T S_i = \mathbf{0} \end{aligned}$$

Theorem 6.4 *The estimator $\hat{\boldsymbol{\beta}}_I$ of $\boldsymbol{\beta}$ is consistent and $n^{\frac{1}{2}}(\hat{\boldsymbol{\beta}}_I - \boldsymbol{\beta})$ asymptotically multivariate Gaussian as $n \to \infty$ with zero mean and covariance matrix V_I given by*

$$\begin{aligned} V_I &= \lim_{n\to\infty} n \left(\sum_{i=1}^{n} \mathbf{X}_i^T \Delta_i A_i \Delta_i \mathbf{X}_i \right)^{-1} \left(\sum_{i=1}^{n} \mathbf{X}_i^T \Delta_i Cov\left[\mathbf{y}_i\right] \Delta_i \mathbf{X}_i \right) \left(\sum_{i=1}^{n} \mathbf{X}_i^T \Delta_i A_i \Delta_i \mathbf{X}_i \right)^{-1} \\ &= \lim_{n\to\infty} n\ H_1(\boldsymbol{\beta})^{-1} H_2(\boldsymbol{\beta}) H_1(\boldsymbol{\beta})^{-1} \end{aligned}$$

where

$$\begin{aligned} H_1(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \mathbf{X}_i^T \Delta_i A_i \Delta_i \mathbf{X}_i \\ H_2(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \mathbf{X}_i^T \Delta_i Cov\left[\mathbf{y}_i\right] \Delta_i \mathbf{X}_i \end{aligned}$$

The variance of $\hat{\boldsymbol{\beta}}_I$ can consistently be estimated by

$$\hat{V}_I = H_1(\hat{\boldsymbol{\beta}}_I)^{-1} \left(\sum_{i=1}^{n} \mathbf{X}_i^T \Delta_i S_i S_i^T \Delta_i \mathbf{X}_i \Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_I} \right) H_1(\hat{\boldsymbol{\beta}}_I)^{-1}$$

$\hat{V}_I$ can be determined without knowledge of ϕ, since A_i is independent of ϕ and $Cov[\mathbf{y}_i]$ is estimated directly by $S_i S_i^T \Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_I}$

Although we totally ignored the correlation structure of successive observations of the same subject, there are advantages for doing so:

- $\hat{\boldsymbol{\beta}}_I$ is easy to compute with existing software
- $\hat{\boldsymbol{\beta}}_I$ and $Var[\hat{\boldsymbol{\beta}}_I]$ are consistent estimators, when the regression is correctly specified
- $\hat{\boldsymbol{\beta}}_I$ is reasonable efficient for a few simple designs (Liang and Zeger 1986)

Nevertheless, omitting the correlation structure can cause low efficiency especially when the correlation is large. Higher efficiency can be obtained using the GEEs, that explicitly take the correlation into account.

6.4.4 Generalized Estimating Equations (GEEs)

In this section we relax the "independence working" assumption and take the correlation between repeated observations from the same subject into account. The resulting estimator $\hat{\boldsymbol{\beta}}_G$ for $\boldsymbol{\beta}$ remains consistent but efficiency is increased. Under the assumption that a weighted average of the estimated correlation matrices converges to a fixed matrix, we can also derive an consistent variance estimate of $\hat{\boldsymbol{\beta}}_G$.

We define for each $\mathbf{y}_i$ the $n_i \times n_i$ correlation matrix $R_i(\boldsymbol{\alpha})$, that is fully parametrized by the $s \times 1$ correlation parameter vector $\boldsymbol{\alpha}$, which is the same for all subjects, whereas the times of observations and the correlation matrix can differ from subject to subject. We only require $R_i(\boldsymbol{\alpha})$ to be a correlation matrix and call it the "working" correlation matrix, as we do not expect it to be correctly specified, though we want consistent estimates and have consistent variances of these estimates. We define V_i using again the $n_i \times n_i$ diagonal matrix $A_i := diag(\mathbf{v}_i\phi)$ as

$$V_i := A_i^{\frac{1}{2}} R_i(\boldsymbol{\alpha}) A_i^{\frac{1}{2}} / \phi \tag{6.30}$$

If $R_i(\boldsymbol{\alpha})$ is the true correlation matrix for $\mathbf{y}_i$, the matrix V_i is equal to the true covariance matrix for $\mathbf{y}_i$, that is $V_i = Cov[\mathbf{y}_i]$.

$$Cov[\mathbf{y}_i] = Var[\mathbf{y}_i]^{\frac{1}{2}} Corr[\mathbf{y}_i] Var[\mathbf{y}_i]^{\frac{1}{2}} = diag(\mathbf{v}_i\phi)^{\frac{1}{2}} R_i(\boldsymbol{\alpha}) diag(\mathbf{v}_i\phi)^{\frac{1}{2}} / \phi$$

In addition to $\boldsymbol{\beta}$ and ϕ we have now also to estimate $\boldsymbol{\alpha}$. To obtain the GEEs we calculate, analogue to (6.28), with (6.30)

$$U_G(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{\partial l(\mathbf{y}_i; \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \frac{\partial l}{\partial \boldsymbol{\mu}_i} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \right)^T V_i^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i) = 0 \tag{6.31}$$

Note the similarity to the quasi-likelihood approach in the derivatives used. Further we define D_{it} as the derivative of μ_{it} with respect to $\boldsymbol{\beta}$, that is

$$D_{it} := \frac{\partial \mu_{it}}{\partial \boldsymbol{\beta}} = \frac{\partial \mu_{it}}{\partial \theta_{it}} \frac{\partial \theta_{it}}{\partial \eta_{it}} \frac{\partial \eta_{it}}{\partial \boldsymbol{\beta}} = b''(\theta_{it}) \Delta_{it} \mathbf{x}_{it} = v_{it} \phi \Delta_{it} \mathbf{x}_{it} \tag{6.32}$$

which is the specification in the case of a distribution from the exponential family. We define the $n_i \times p$ matrix D_i using the vectors D_{it} by

$$D_i := (D_{i1}, \ldots, D_{in_i})^T = \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} = A_i \Delta_i \mathbf{X}_i \in \mathbb{R}^{n_i \times p}$$

With (6.32) the GEEs from (6.31) have the form

$$U_G(\boldsymbol{\beta}) = \sum_{i=1}^{n} D_i^T V_i^{-1} S_i = \sum_{i=1}^{n} U_i(\boldsymbol{\beta}, \boldsymbol{\alpha}) = 0 \tag{6.33}$$

where $U_i(\boldsymbol{\beta}, \boldsymbol{\alpha}) := D_i^T V_i^{-1} S_i$. If only one observation is available for each subject, that is $n_i = 1$, this equation becomes identical to the Score-like equation obtained for quasi-likelihood (6.19).

One can easily check that the general estimating equations reduce to the independence equation (6.29) if we specify $R_i(\boldsymbol{\alpha})$ as the identity matrix. In contrast to the quasi-likelihood approach from Wedderburn (1974), the matrix V_i in the function $U_i(\boldsymbol{\beta}, \boldsymbol{\alpha})$ depends for each i not only on the parameter $\boldsymbol{\beta}$ but also on the parameter $\boldsymbol{\alpha}$.

A next step is now to replace $\boldsymbol{\alpha}$ and ϕ by any $n^{\frac{1}{2}}$-consistent estimators. Assuming that $\boldsymbol{\beta}$ and ϕ are known we denote the estimator for $\boldsymbol{\alpha}$ by $\hat{\boldsymbol{\alpha}}(\boldsymbol{\beta},\phi) := \hat{\boldsymbol{\alpha}}(Y,\boldsymbol{\beta},\phi)$. Given $\boldsymbol{\beta}$ we take $\hat{\phi}(\boldsymbol{\beta}) := \hat{\phi}(Y,\boldsymbol{\beta})$ as estimator for ϕ. The estimator $\hat{\boldsymbol{\alpha}}$ is called $n^{\frac{1}{2}}$ consistent, if $n^{\frac{1}{2}}(\hat{\boldsymbol{\alpha}}(Y,\boldsymbol{\beta},\phi)-\boldsymbol{\alpha}) = O_p(1)$. We insert these $n^{\frac{1}{2}}$-consistent estimators in (6.33) and obtain:

$$U_G(\boldsymbol{\beta}) \approx \sum_{i=1}^{n} U_i\left(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}(\boldsymbol{\beta}, \hat{\phi}(\boldsymbol{\beta}))\right) = 0$$

We define $\hat{\boldsymbol{\beta}}_G$ as the solution of this equation. We summarize its large sample properties in the following theorem:

6.4.5 Large Sample Properties of $\hat{\boldsymbol{\beta}}_G$

Theorem 6.5 *Under mild regularity conditions and given that:*

(i) $\hat{\boldsymbol{\alpha}}$ *is* $n^{\frac{1}{2}}$*-consistent given* $\boldsymbol{\beta}$ *and* ϕ*;*

(ii) $\hat{\phi}$ *is* $n^{\frac{1}{2}}$*-consistent given* $\boldsymbol{\beta}$*; and*

(iii) $|\partial\hat{\boldsymbol{\alpha}}(\boldsymbol{\beta},\phi)/\partial\phi| \leq H(Y,\boldsymbol{\beta})$ *which is* $O_p(1)$*,*

then $n^{\frac{1}{2}}\left(\hat{\boldsymbol{\beta}}_G - \boldsymbol{\beta}\right)$ *is asymptotically multivariate Gaussian with zero mean and covariance matrix* V_G *given by*

$$V_G = \lim_{n\to\infty} n \left(\sum_{i=1}^{n} D_i^T V_i^{-1} D_i\right)^{-1} \left(\sum_{i=1}^{n} D_i^T V_i^{-1} Cov\left[\mathbf{y}_i\right] V_i^{-1} D_i\right) \left(\sum_{i=1}^{n} D_i^T V_i^{-1} D_i\right)^{-1}$$

Proof:

Under mild regularity conditions we obtain, using first order Taylor approximation and defining $\boldsymbol{\alpha}^*(\boldsymbol{\beta}) := \hat{\boldsymbol{\alpha}}(\boldsymbol{\beta}, \hat{\phi}(\boldsymbol{\beta}))$:

$$\hat{\boldsymbol{\beta}}_G = \boldsymbol{\beta} - \left(\sum_{i=1}^{n} \frac{\partial}{\partial\boldsymbol{\beta}}\left[U_i(\boldsymbol{\beta}, \boldsymbol{\alpha}^*(\boldsymbol{\beta}))\right]\right)^{-1} \left(\sum_{i=1}^{n} U_i(\boldsymbol{\beta}, \boldsymbol{\alpha}^*(\boldsymbol{\beta}))\right) + o_p(n^{-\frac{1}{2}})$$

For details on the O_p, o_p notation see Appendix A.4. Thus we get

$$n^{\frac{1}{2}}\left(\hat{\boldsymbol{\beta}}_G - \boldsymbol{\beta}\right) = \left(\sum_{i=1}^{n} -\frac{\partial}{\partial\boldsymbol{\beta}}\left[U_i(\boldsymbol{\beta}, \boldsymbol{\alpha}^*(\boldsymbol{\beta}))\right]/n\right)^{-1} \left(\sum_{i=1}^{n} U_i(\boldsymbol{\beta}, \boldsymbol{\alpha}^*(\boldsymbol{\beta}))/n^{\frac{1}{2}}\right) + o_p(1)$$

where

$$\frac{\partial}{\partial\boldsymbol{\beta}}\left[U_i(\boldsymbol{\beta}, \boldsymbol{\alpha}^*(\boldsymbol{\beta}))\right] = \underbrace{\frac{\partial U_i(\boldsymbol{\beta}, \boldsymbol{\alpha}^*(\boldsymbol{\beta}))}{\partial\boldsymbol{\beta}}}_{A_i} + \underbrace{\frac{U_i(\boldsymbol{\beta}, \boldsymbol{\alpha}^*(\boldsymbol{\beta}))}{\partial\boldsymbol{\alpha}^*(\boldsymbol{\beta})}}_{B_i}\underbrace{\frac{\partial\boldsymbol{\alpha}^*(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}}_{C}$$

Remembering, that we defined D_i as $\partial\boldsymbol{\mu}_i/\partial\boldsymbol{\beta}$, we know that the derivative of $U_i(\boldsymbol{\beta},\boldsymbol{\alpha})$ with respect to $\boldsymbol{\beta}$ can be derived as follows:

$$\begin{aligned}
\sum_{i=1}^{n}\frac{\partial U_i(\boldsymbol{\beta},\boldsymbol{\alpha})}{\partial\boldsymbol{\beta}} &= \sum_{i=1}^{n}\frac{\partial}{\partial\boldsymbol{\beta}}\left(D_i^T V_i^{-1}(\mathbf{y}_i-\boldsymbol{\mu}_i)\right)\\
&= \sum_{i=1}^{n}\left[\frac{\partial}{\partial\boldsymbol{\beta}}\left(D_i^T V_i^{-1}\right)\right](\mathbf{y}_i-\boldsymbol{\mu}_i)+\left(D_i^T V_i^{-1}\right)\frac{\partial}{\partial\boldsymbol{\beta}}(\mathbf{y}_i-\boldsymbol{\mu}_i)\\
&= \sum_{i=1}^{n}\frac{\partial}{\partial\boldsymbol{\beta}}\left(D_i^T V_i^{-1}\right)(\mathbf{y}_i-\boldsymbol{\mu}_i)-\left(D_i^T V_i^{-1}\right)\frac{\partial\boldsymbol{\mu}_i}{\partial\boldsymbol{\beta}}\\
&= \sum_{i=1}^{n}\frac{\partial}{\partial\boldsymbol{\beta}}\left(D_i^T V_i^{-1}\right)(\mathbf{y}_i-\boldsymbol{\mu}_i)-D_i^T V_i^{-1}D_i^T
\end{aligned}$$

Taking expectation the first term vanishes and we are left with the last one. Further note that $\partial U_i(\boldsymbol{\beta},\boldsymbol{\alpha})/\partial\boldsymbol{\alpha}=\partial D_i^T V_i^{-1}S_i/\partial\boldsymbol{\alpha}$ is a linear function of the S_i's. Consequently this function has mean zero. Taking these results in mind and using the definition of A_i, B_i and C as above we have now the following equations:

$$\begin{aligned}
\sum_{i=1}^{n}A_i/n &= \sum_{i=1}^{n}\frac{\partial U_i(\boldsymbol{\beta},\boldsymbol{\alpha}^*(\boldsymbol{\beta}))}{\partial\boldsymbol{\beta}}/n \approx -\sum_{i=1}^{n}D_i^T V_i^{-1}D_i/n \qquad as \quad n\to\infty\\
\sum_{i=1}^{n}B_i/n &= \sum_{i=1}^{n}\frac{\partial U_i(\boldsymbol{\beta},\boldsymbol{\alpha}^*(\boldsymbol{\beta}))}{\partial\boldsymbol{\alpha}^*(\boldsymbol{\beta})}/n = o_p(1)\\
C &= \frac{\partial\boldsymbol{\alpha}^*(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}=\frac{\partial\hat{\boldsymbol{\alpha}}(\boldsymbol{\beta},\hat{\phi}(\boldsymbol{\beta}))}{\partial\boldsymbol{\beta}}=\frac{\hat{\alpha}(\boldsymbol{\beta},\phi)-\hat{\alpha}(\boldsymbol{\beta},\hat{\phi}(\boldsymbol{\beta}))}{\phi-\hat{\phi}(\boldsymbol{\beta})}\overset{(6.34)}{=}\frac{O_p(1)}{O_p(1)}=O_p(1)
\end{aligned}$$

where the last equation holds, as both, $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\phi}}$ are $n^{\frac{1}{2}}$-consistent estimators. For fixed $\boldsymbol{\beta}$ a first order Taylor expansion of $U_i(\boldsymbol{\beta},\boldsymbol{\alpha}^*(\boldsymbol{\beta}))$ around $\boldsymbol{\alpha}$ gives

$$\frac{\sum_{i=1}^{n}U_i(\boldsymbol{\beta},\boldsymbol{\alpha}^*(\boldsymbol{\beta}))}{n^{\frac{1}{2}}}=\underbrace{\frac{\sum_{i=1}^{n}U_i(\boldsymbol{\beta},\boldsymbol{\alpha})}{n^{\frac{1}{2}}}}_{A^*}+\underbrace{\frac{\sum_{i=1}^{n}\frac{\partial}{\partial\boldsymbol{\alpha}}U_i(\boldsymbol{\beta},\boldsymbol{\alpha})}{n}}_{B^*}\underbrace{n^{\frac{1}{2}}\left(\boldsymbol{\alpha}^*(\boldsymbol{\beta})-\boldsymbol{\alpha}\right)}_{C^*}+o_p(1)$$

Again we use the fact that $\partial U_i(\boldsymbol{\beta},\boldsymbol{\alpha})/\partial\boldsymbol{\alpha}=\partial D_i^T V_i^{-1}S_i/\partial\boldsymbol{\alpha}$ is a linear function of the S_i's and its expectation is therefore zero. Thus we obtain for B^*

$$B^*=\sum_{i=1}^{n}\frac{\partial}{\partial\boldsymbol{\alpha}}U_i(\boldsymbol{\beta},\boldsymbol{\alpha})/n=o_p(1)$$

and for C^* we get using Taylor expansion

$$\begin{aligned}
C^* &= n^{\frac{1}{2}}\left(\boldsymbol{\alpha}^*-\boldsymbol{\alpha}\right)=n^{\frac{1}{2}}\left(\hat{\boldsymbol{\alpha}}(\boldsymbol{\beta},\hat{\phi}(\boldsymbol{\beta}))-\hat{\boldsymbol{\alpha}}(\boldsymbol{\beta},\phi)+\hat{\boldsymbol{\alpha}}(\boldsymbol{\beta},\phi)-\boldsymbol{\alpha}\right)+n^{\frac{1}{2}}\frac{o_p(1)}{n^{\frac{1}{2}}}\\
&\overset{Taylor}{=} n^{\frac{1}{2}}\left(\underbrace{\frac{\partial\hat{\boldsymbol{\alpha}}}{\partial\hat{\phi}(\boldsymbol{\beta})}(\boldsymbol{\beta},\hat{\phi}(\boldsymbol{\beta}))}_{(iii)\Rightarrow O_p(1)}\underbrace{(\hat{\phi}(\boldsymbol{\beta})-\phi)}_{(ii)\Rightarrow O_p(1)}+\underbrace{\hat{\boldsymbol{\alpha}}(\boldsymbol{\beta},\phi)-\boldsymbol{\alpha}}_{(i)\Rightarrow O_p(1)}\right)=O_p(1) \qquad (6.34)
\end{aligned}$$

Consequently, $\sum_{i=1}^{n} U_i(\boldsymbol{\beta}, \boldsymbol{\alpha}^*(\boldsymbol{\beta}))$ is asymptotically equivalent to $\sum_{i=1}^{n} U_i(\boldsymbol{\beta}, \boldsymbol{\alpha})$ whose asymptotic distribution is multivariate Gaussian with mean zero and covariance matrix given by

$$V = \lim_{n\to\infty} \left(\sum_{i=1}^{n} D_i^T V_i^{-1} Cov[\mathbf{y}_i] V_i^{-1} D_i / n \right)$$

This is, because

$$A^* = \frac{\sum_{i=1}^{n} U_i(\boldsymbol{\beta}, \boldsymbol{\alpha})}{n^{\frac{1}{2}}} \overset{(6.33)}{=} \sum_{i=1}^{n} \frac{1}{n^{\frac{1}{2}}} D_i^T V_i^{-1} S_i = \sum_{i=1}^{n} \underbrace{\frac{1}{n^{\frac{1}{2}}} D_i^T V_i^{-1}}_{A^{(i)}} (\mathbf{y}_i - \boldsymbol{\mu}_i)$$

Using the formulas $E[A^{(i)}X] = A^{(i)}E[X]$ and $Var[A^{(i)}X] = A^{(i)}Var[X](A^{(i)})^T$ we get

$$\begin{aligned} E\left[A^*\right] &= \sum_{i=1}^{n} A^{(i)} E\left[\mathbf{y}_i - \boldsymbol{\mu}_i\right] = 0 \\ Var\left[A^*\right] &= \sum_{i=1}^{n} A^{(i)} Var\left[\mathbf{y}_i - \boldsymbol{\mu}_i\right] (A^{(i)})^T = \frac{1}{n} \sum_{i=1}^{n} D_i^T V_i^{-1} Cov\left[\mathbf{y}_i - \boldsymbol{\mu}_i\right] V_i^{-1} D_i \end{aligned}$$

This proves that the asymptotic variance of $\sum_{i=1}^{n} U_i(\boldsymbol{\beta}, \boldsymbol{\alpha})$ is V and leaves us with

$$\begin{aligned} n^{\frac{1}{2}} \left(\hat{\boldsymbol{\beta}}_G - \boldsymbol{\beta} \right) &= \left(\sum_{i=1}^{n} -\frac{\partial}{\partial \boldsymbol{\beta}} U_i(\boldsymbol{\beta}, \boldsymbol{\alpha}^*(\boldsymbol{\beta}))/n \right)^{-1} \left(\sum_{i=1}^{n} U_i(\boldsymbol{\beta}, \boldsymbol{\alpha}^*(\boldsymbol{\beta}))/n^{\frac{1}{2}} \right) + o_p(1) \\ &= \left(\sum_{i=1}^{n} D_i^T V_i^{-1} D_i / n \right)^{-1} \left(\sum_{i=1}^{n} U_i(\boldsymbol{\beta}, \boldsymbol{\alpha})/n^{\frac{1}{2}} + o_p(1) \right) + o_p(1) \end{aligned}$$

As $n \to \infty$ we can now calculate the asymptotic distribution for

$$\underbrace{\left(\sum_{i=1}^{n} D_i^T V_i^{-1} D_i / n \right)^{-1}}_{=:A} \left(\sum_{i=1}^{n} U_i(\boldsymbol{\beta}, \boldsymbol{\alpha})/n^{\frac{1}{2}} \right)$$

The second term has mean zero and variance V. Again using the formulas $E[AX] = AE[X]$ and $Var[AX] = AVar[X]A^T$ we finally obtain the required result:

$$\begin{aligned} V_G &:= \lim_{n\to\infty} Var\left[n^{\frac{1}{2}} \left(\hat{\boldsymbol{\beta}}_G - \boldsymbol{\beta} \right) \right] \\ &= \lim_{n\to\infty} \left(\sum_{i=1}^{n} D_i^T V_i^{-1} D_i / n \right)^{-1} \left(\sum_{i=1}^{n} D_i^T V_i^{-1} Cov\left[\mathbf{y}_i\right] V_i^{-1} D_i / n \right) \left(\sum_{i=1}^{n} D_i^T V_i^{-1} D_i / n \right)^{-1} \\ &= \lim_{n\to\infty} n \left(\sum_{i=1}^{n} D_i^T V_i^{-1} D_i \right)^{-1} \left(\sum_{i=1}^{n} D_i^T V_i^{-1} Cov\left[\mathbf{y}_i\right] V_i^{-1} D_i \right) \left(\sum_{i=1}^{n} D_i^T V_i^{-1} D_i \right)^{-1} \end{aligned}$$

We can estimate V_G by inserting the estimators for $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$ and ϕ in above formula and replacing $Cov[\mathbf{y}_i]$ by the empirical covariance $S_i S_i^T$. The consistency of $\hat{\boldsymbol{\beta}}_G$ and $\hat{V}_G$ depends only on the correct specification of the mean and not on the choice of $R_i(\boldsymbol{\alpha})$. Equation (6.33) converges to zero and consequently has consistent roots, if $E[S_i] = E[\mathbf{y}_i] - \boldsymbol{\mu}_i = \mathbf{0}$. The asymptotic distribution of $\hat{\boldsymbol{\beta}}_G$ does not depend on the specific choice of $\boldsymbol{\alpha}$ and ϕ as long as they are $n^{\frac{1}{2}}$-consistent.

6.4.6 Iterative Weighted Least-Squares Algorithm for GEEs

We apply the same procedure as with the likelihood and quasi-likelihood function to obtain a solution for the Score and Score-like equations, respectively. We use the Newton-Raphson method and replace the matrix of 2^{nd} derivatives of the log-likelihood function by its expectation (Fisher-scoring). The expectation of the Hessian matrix in the GLM case is

$$\begin{aligned}
-E\left[\frac{\partial^2 l(\boldsymbol{\beta};\mathbf{y}_i)}{\partial^2\boldsymbol{\beta}}\right] &= -E\left[\sum_{i=1}^{n}\frac{\partial}{\partial\boldsymbol{\beta}}\left(\frac{\partial l(\boldsymbol{\beta};\mathbf{y}_i)}{\partial\boldsymbol{\beta}}\right)\right] \\
&= -E\left[\sum_{i=1}^{n}\frac{\partial}{\partial\boldsymbol{\beta}}\left((\mathbf{y}_i-\boldsymbol{\mu}_i)\,V_i^{-1}\frac{\partial\boldsymbol{\mu}_i}{\partial\boldsymbol{\beta}}\right)\right] \\
&= E\left[\sum_{i=1}^{n}\left(\frac{\partial\boldsymbol{\mu}_i}{\partial\boldsymbol{\beta}}\right)^T V_i^{-1}\frac{\partial\boldsymbol{\mu}_i}{\partial\boldsymbol{\beta}}\right] \\
&= \sum_{i=1}^{n} D_i^T V_i^{-1} D_i
\end{aligned}$$

But in the case of general estimating equation we have to account for additional parameter estimation for $\boldsymbol{\alpha}$. So we iterate between Fisher-scoring for $\boldsymbol{\beta}$ and moment estimation of $\boldsymbol{\alpha}$ and ϕ. Given current estimates $\hat{\boldsymbol{\alpha}}$ and $\hat{\phi}$ the procedure for $\boldsymbol{\beta}$ is the following:

$$\hat{\boldsymbol{\beta}}_{j+1} = \hat{\boldsymbol{\beta}}_j - \left(\sum_{i=1}^{n} D_i^T(\hat{\boldsymbol{\beta}}_j)\tilde{V}_i(\hat{\boldsymbol{\beta}}_j)^{-1}D_i(\hat{\boldsymbol{\beta}}_j)\right)^{-1}\left(\sum_{i=1}^{n} D_i^T(\hat{\boldsymbol{\beta}}_j)\tilde{V}_i(\hat{\boldsymbol{\beta}}_j)^{-1}S_i(\hat{\boldsymbol{\beta}}_j)\right) \tag{6.35}$$

where $\tilde{V}_i(\boldsymbol{\beta}) = V_i[\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}(\boldsymbol{\beta}, \hat{\phi}(\boldsymbol{\beta}))]$. This procedure is a modification of Fisher-scoring method. We replace the matrix of expectation of 2^{nd} derivatives of the likelihood function by the limiting value of the expectation of the derivative of $\sum_{i=1}^{n} U_i[\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}(\boldsymbol{\beta}, \hat{\phi}(\boldsymbol{\beta}))]$.

Define $D := (D_1^T, \ldots, D_n^T)^T$, $S := (S_1^T, \ldots, S_n^T)$ and $Z := D\boldsymbol{\beta} - S$. Further let $\tilde{V}$ be a $n_i n \times n_i n$ block diagonal matrix with $\tilde{V}_i$'s as the diagonal elements. Then the iterative procedure (6.35) for calculating $\hat{\boldsymbol{\beta}}_G$ is equivalent to performing an iteratively re-weighted linear regression of Z, as dependent variable on D with weight $\tilde{V}^{-1}$. We iteratively solve for the regression coefficients and the correlation and scale parameters $\boldsymbol{\alpha}$ and ϕ.

To understand this better, notice once again the close relation to the derivation of the re-weighted linear regression in the two previous sections.

To simplify notation we abbreviate the quantities from equation (6.35) as follows:

$$\boldsymbol{\beta}_G = \boldsymbol{\beta} - A_G^{-1}U_G \qquad \Rightarrow \qquad A_G\boldsymbol{\beta}_G = A_G\boldsymbol{\beta} - U_G$$

Inserting the results, we derived in this section, we obtain:

$$
\begin{aligned}
\sum_{i=1}^{n} D_i^T V_i^{-1} D_i \boldsymbol{\beta}_G &= \sum_{i=1}^{n} D_i^T V_i^{-1} D_i \boldsymbol{\beta} - \sum_{i=1}^{n} D_i^T V_i^{-1} S_i \\
\sum_{i=1}^{n} D_i^T V_i^{-1} D_i \boldsymbol{\beta}_G &= \sum_{i=1}^{n} D_i^T V_i^{-1} \underbrace{(D_i \boldsymbol{\beta} - S_i)}_{=:Z_i} \\
\sum_{i=1}^{n} \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\eta}_i} \frac{\partial \boldsymbol{\eta}_i}{\partial \boldsymbol{\beta}} \right)^T V_i^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\eta}_i} \frac{\partial \boldsymbol{\eta}_i}{\partial \boldsymbol{\beta}} \boldsymbol{\beta}_G &= \sum_{i=1}^{n} \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\eta}_i} \frac{\partial \boldsymbol{\eta}_i}{\partial \boldsymbol{\beta}} \right)^T V_i^{-1} \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\eta}_i} \frac{\partial \boldsymbol{\eta}_i}{\partial \boldsymbol{\beta}} \boldsymbol{\beta} - S_i \right) \\
W_i &= \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\eta}_i} \right)^T V_i^{-1} \frac{\partial \mu_i}{\partial \boldsymbol{\eta}_i} \\
\sum_{i=1}^{n} \left(\frac{\partial \boldsymbol{\eta}_i}{\partial \boldsymbol{\beta}} \right)^T W_i \frac{\partial \boldsymbol{\eta}_i}{\partial \boldsymbol{\beta}} \boldsymbol{\beta}_G &= \sum_{i=1}^{n} \left(\frac{\partial \boldsymbol{\eta}_i}{\partial \boldsymbol{\beta}} \right)^T W_i \frac{\partial \boldsymbol{\eta}_i}{\partial \boldsymbol{\mu}_i} \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\eta}_i} \frac{\partial \boldsymbol{\eta}_i}{\partial \boldsymbol{\beta}} \boldsymbol{\beta} - S_i \right) \\
\sum_{i=1}^{n} \mathbf{X}_i^T W_i \boldsymbol{\eta}_i^* &= \sum_{i=1}^{n} \mathbf{X}_i^T W_i \left(\boldsymbol{\eta}_i - \frac{\partial \boldsymbol{\eta}_i}{\partial \boldsymbol{\mu}_i} (\mathbf{y}_i - \boldsymbol{\mu}_i) \right)
\end{aligned}
$$

where $\boldsymbol{\eta}_i^* := \mathbf{X}_i \boldsymbol{\beta}_G$. Comparing this with the results derived for the Maximum quasi-likelihood estimation in (6.26), one sees the similarities between both approaches.

6.4.7 Estimation of the Parameters $\boldsymbol{\alpha}$ and ϕ

For estimating the values of $\boldsymbol{\alpha}$ and ϕ we use Pearson residuals (see Liang and Zeger (1986)). They can be calculated in each step of the iteration given the current value for $\boldsymbol{\beta}$ as

$$
\hat{r}_{it} := \frac{y_{it} - b'(\hat{\theta}_{it})}{b''(\hat{\theta}_{it})^{\frac{1}{2}}}
$$

where we calculate $\hat{\theta}_{it}$ from the current value for $\boldsymbol{\beta}$. Let N be again the number of all observations, that is $N = \sum_{i=1}^{n} n_i$, we can estimate ϕ by

$$
\hat{\phi}^{-1} = \sum_{i=1}^{n} \sum_{t=1}^{n_i} \frac{\hat{r}_{it}^2}{N-p}
$$

which is the longitudinal analogue of the familiar Pearson statistic (Zeger and Liang 1986). This can be proved as follows:

$$
E\left[\sum_{i=1}^{n} \sum_{t=1}^{n_i} \frac{(y_{it} - \mu_{it})^2}{V(\mu_{it})/\phi} \right] = E\left[\sum_{i=1}^{n} \sum_{t=1}^{n_i} \frac{(y_{it} - \mu_{it})^2}{Var[y_{it}]} \right] \approx N - p
$$

$$
\Rightarrow \qquad \phi^{-1} \approx \frac{1}{N-p} E\left[\sum_{i=1}^{n} \sum_{t=1}^{n_i} \frac{(y_{it} - \mu_{it})^2}{V(\mu_{it})} \right] \approx \sum_{i=1}^{n} \sum_{t=1}^{n_i} \frac{\hat{r}_{it}^2}{N-p}
$$

The estimation of the parameter $\boldsymbol{\alpha}$ depends on the correlation structure selected for the working correlation matrix. In the case that observation times are the same for all subjects so that the working correlation matrix $R_i(\boldsymbol{\alpha}) = R(\boldsymbol{\alpha})$ for all $i = 1, \ldots, n$, we give in the next section estimators for different types of correlation structures following Liang and Zeger (1986):

Correlation structures for GEEs

Liang and Zeger (1986) presented five different types of correlation structure: An "independent", "exchangeable", "unstructured", "autoregressive (AR-I)" and "one-dependent" working correlation. In the following we are going to define above mentioned correlation structures and give possible estimators for them:

The working correlation matrix $R(\boldsymbol{\alpha})$ can be chosen to be the identity matrix. $R(\boldsymbol{\alpha}) = \mathbf{I}$. Obviously different outcomes from the same observation are then assumed to have zero correlation.

$$Corr[y_{it}, y_{it'}] = 0 \qquad \forall \quad t \neq t'$$

In this case we do not allow any correlation between observations even if we measure the same observations at different points in time. Regardless if outcomes are from different or the same time series, they are assumed to be independent and the GEEs reduce to the independence estimating equations.

Liang and Zeger (1986) investigated the effect of the choice of the correlation matrix on the efficiency of $\boldsymbol{\beta}$ in simulation studies. They found out that in cases with small correlation both estimators, the estimator obtained by solving the independence estimating equations $\boldsymbol{\beta}_I$ and the estimator obtained by solving the generalized estimating equation $\boldsymbol{\beta}_G$, are efficient, but, as correlation increases $\boldsymbol{\beta}_G$ remains efficient, whereas $\boldsymbol{\beta}_I$ does not.

The estimation of the correlation matrix in case of an "independent" working correlation matrix is unnecessary here, since the correlation matrix is fixed to the identity matrix.

If we choose the working correlation matrix as "exchangeable" the correlation between different observations within a time series is the same regardless of the distance in time.

$$Corr[y_{it}, y_{it'}] = \alpha \qquad \forall \quad t \neq t'$$

An estimator for this correlation structure is given by

$$\hat{\alpha} = \frac{\phi}{n} \sum_{i=1}^{n} \sum_{t>t'} \hat{r}_{it}\hat{r}_{it'} / \left(\frac{1}{2} \cdot n_i \cdot (n_i - 1) - p \right)$$

In contrast to the identity matrix, that does not allow for any correlation, we could use a totally unspecified working correlation matrix. In this "unstructured" case we have to estimate all $\frac{1}{2} \cdot n_i \cdot (n_i - 1)$ correlations. This can be done by

$$\frac{\phi}{n} \sum_{i=1}^{n} A_i^{-\frac{1}{2}} S_i S_i^T A_i^{-\frac{1}{2}}$$

where $A_i = diag(\mathbf{v}_i \phi)$ and $S_i = \mathbf{y}_i - \boldsymbol{\mu}_i$ are the quantities as defined in Section 6.4.4 and Section 6.4.3, respectively.

The "autoregressive (AR-I)" working correlation is nothing else than the correlation structure of a continuous first-order autoregressive process (AR-I). This means that observations with the same distance in time have the same correlation, where the correlation decreases polynomially as the distance increases.

$$Corr[y_{it}, y_{it'}] = \alpha^{|t-t'|}$$

Since in the case of an "autoregressive (AR-I)" correlation structure $E[\hat{r}_{it}\hat{r}_{it'}] \approx \alpha^{|t-t'|}$, Liang and Zeger (1986) suggested to estimate the parameter α by the slope obtained from the regression of $\log\{\hat{r}_{it}\hat{r}_{it'}\}$ on $\log\{|t-t'|\}$.

In contrast to the polynomial decrease of the correlations in the "autoregressive (AR-I)" case is the "dependent" correlation structure. Observations with the same distance do still have the same correlation, but for each distance a separate value, not necessarily decreasing, is estimated.

$$Corr[y_{it}, y_{it'}] = \alpha_{|t-t'|} \qquad \text{if } t \neq t'$$

A special case of the "dependent" working correlation matrix is the "one-dependent" structure. This is equivalent to the correlation structure of a stationary Markov process of degree one, i.e.

$$Corr[y_{it}, y_{it'}] = \alpha_{|t-t'|} \qquad \text{if } t \leq 1$$

and zero otherwise. The "one-dependent" correlation structure needs $n_i - 1$ parameters, that can be estimated by

$$\hat{\alpha}_t = \frac{\phi}{n-p}\sum_{i=1}^{n}\hat{r}_{i,t}\hat{r}_{i,t+1}$$

In the case where $\alpha_t = \alpha$ for all $t = 1, \ldots, n_i - 1$ we can estimate the overall α as

$$\hat{\alpha} = \frac{1}{n_i - 1}\sum_{i=1}^{n_i-1}\hat{\alpha}_t$$

Since $\boldsymbol{\beta}_G$ and V_G are robust to the choice of the correlation structure (see Liang and Zeger (1986)) we obtain according to Theorem 6.5 asymptotically correct estimates even if the correlation structure is misspecified. Clearly, if we choose the working correlation matrix close to the true correlation the estimates will be more efficient. For details on simulation studies using different correlation structures and misspecified correlation structures see Liang and Zeger (1986).

The program written by Mark X. Norleans provides all of these five correlation structures, namely the "independent", "exchangeable", which is called "compoundsymmetric" in this program , the "unstructured", "autoregressive (AR-I)" and "dependent" correlation structure. The program can be obtained from http://lib.stat.cmu.edu/ and is designed for Splus.

Another program called "Oswald" is developed by the Statistics Group at the University of Lancaster and can be obtained from http://www.maths.lancs.ac.uk/Software/Oswald/ as a Splus library. Here additional correlation structures are possible: Beside the above mentioned correlation structures one can choose between a "stationary Markov process" or a "non-stationary Markov process" structure of degree "Mv", where "Mv" is a quantity to be specified. Further a fixed user-specified matrix "R" can be used as well as the correlation structure of an autoregressive process of degree "Mv".

Chapter 7

Application to Data

In this section we want to apply the results derived in the previous chapters to LTCI. We use a Markovian multi-state model with states corresponding to the levels and places of care according to the German compulsory LTCI system, the fifth component of the German welfare system, that has been introduced in 1995. The relevant law (§14, BGB 11 1994, see Sozialgesetzbuch (1994)) defines persons eligible for benefits in context of LTCI as follows:

> *"LTC beneficiaries are persons who, on account of a physical, mental or psychic illness or disability, are in considerable or even more serious need of care for usual and regular recurring activities of daily living on a continuing base, presumably for at least six months."*

The German system distinguishes between two places of care, "Care at home" and "Care in a nursing home" and between three levels of care, "Level 1", "Level 2" and "Level 3", where the severity of care increases from "Level 1" to "Level 3". In the definition of the three levels of care four areas, personal hygiene, nutrition, mobility and household activities, are specified, where assistance might be necessary. The first three areas are subsumed as basic care. For each area certain activities are specified. For example for personal hygiene the following activities are outlined: Washing, showering, bathing, dental care, combing, shaving and using the toilet. To qualify for any of the three levels an individual has to require assistance for a certain pre-specified amount of time:

- Level 1 (considerably in need of care): The individual requires at least once a day for more than 90 minutes help, including at least 45 minutes help for at least two activities of one or more areas of basic care.
- Level 2 (seriously in need of care): The individual requires at least three times a day, at different times of the day, for more than 3 hours help, including at least 2 hours help for basic care.
- Level 3 (extremely in need of care): The individual requires at least five times a day, around the clock and also at night time, for more than 5 hours help, including at least 4 hours help for basic care.

Benefits are paid, depending on the place of care and level needed, up to certain ceilings. In the case of "Care at home" one can choose between benefits in kind (up to 1432 EUR in Level 3) or a care allowances (up to 665 EUR in Level 3). For "Care in a nursing home" the benefits range from 1023 EUR (Level 1) to 1432 EUR (Level 3). Additional benefits are also possible.

We are going to analyze the case of a three-state model with states "Care at home", "Care in a nursing home" and "Death". This is equivalent to the Illness-Death model already introduced in Section 3 (Figure 3.1). Specifying the Illness-Death model for our purposes we obtain:

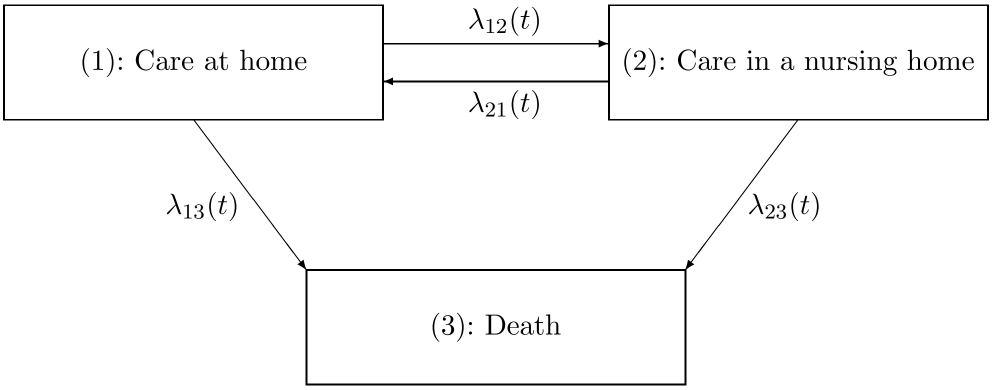

Figure 7.1: Markov model with states "Care at home", "Care in a nursing home" and "Death"

Given a representative sample of claim-records from the private compulsory LTCI, we calculated pseudo-values as explained in Section 5.2 using the Aalen-Johansen estimator from Section 4.3 and performed a regression using GEEs as introduced in Section 6.4. We calculate the Aalen-Johansen estimator and thus the pseudo-values at different points in time using the same observations. In other words, one observation contributes to the Aalen-Johansen estimator at different points in time and thus influences the outcome, the pseudo-values, at these times. Therefore the assumption of independence required for GLMs does no longer hold and we use GEEs, that in contrast to GLMs take correlation between successive outcomes into account. With the parameters obtained from the GEEs the transition matrix can be estimated and premiums calculated, following the approach sketched in Chapter 8.

The necessary calculations were performed using the programs "S-Plus" and "Matlab". An introduction to "Splus" has been written by Venables and Ripley (2000) and to "Matlab" by Hanselman and Littlefield (1997). The program code of the functions used can be found in Appendix B.1, where we used "Matlab" for the computationally intensive calculation of the Aalen-Johansen estimator and "Splus" for the preparation of the data and the final parameter estimation using GEEs.

Finally we want to derive the one-year transition probabilities for each combination of age z and duration l of years spend in care:

$$\mathbf{P}_{gh}(l,z) = P\left(S(l+1) = h \middle| S(l) = g,\ \text{Age } = z\right)$$

where the Markov chain S denotes now the claim-history of an individual as introduced in Chapter 3, that had been a LTC claimant for l years. The estimated transition probabilities $\hat{\mathbf{P}}_{gh}(l,z)$ are then used in Section 8.6, to calculate the necessary actuarial values and to derive the premiums in an insurance model. But before doing this, lets have a look to the data first:

7.1 Data Description

The data were collected between April 1^{st}, 1995 and December 31^{st}, 1998. Subject to observation were 5593 individuals, 3505 female and 2088 male. Their claim-history is given with corresponding levels of care, "Level 1", "Level 2" or "Level 3" and places of care, "Care at home" (ah) or "Care in a nursing home" (nh) at the times a transition occurred. The definition of level and place of care is due to German legislation as explained above. Out of the 5593 individuals, 3264 were censored during the survey due to various reasons. In total 7348 transitions could be observed; their relative frequency is displayed in the following table:

	cens.	dead	recov.	Level 1 (ah)	Level 2 (ah)	Level 3 (ah)	Level 1 (nh)	Level 2 (nh)	Level 3 (nh)
Level 1 (ah)	1011	279	28	0	444	75	118	68	34
Level 2 (ah)	873	597	1	46	0	296	9	208	58
Level 3 (ah)	307	631	2	2	20	0	0	4	87
Level 1 (nh)	248	85	3	9	1	0	0	108	26
Level 2 (nh)	449	263	2	1	2	1	7	0	116
Level 3 (nh)	376	437	0	1	0	4	2	9	0

It becomes clear from this table that total recovery, that is the individual is no longer LTC patient, as well as an improvement regarding level or place of care, is a rare event. Therefore we will not allow for recoveries and improvements in our model, and treat these observations like censored observations from the time on the individual recovers or improves.

Further investigation tells us that 5198 transitions occurred from "Care at home", 3028 female, 2170 male, and 2150 transitions from "Care in a nursing home", 1646 female and 504 male:

	female transition	male transition	total transitions
at home	3028	2170	5198
nursing home	1646	504	2150

On the other side 2537 transitions occurred from "Level 1", 1684 female and 853 male, 2929 transitions from "Level 2", 1859 female and 1070 male, and 1882 transitions from "Level 3", 1131 female and 751 male. This result is summarized in the following table:

	female transition	male transition	total transitions
Level 1	1684	853	2537
Level 2	1859	1070	2929
Level 3	1131	751	1882

Note that in above tables we recored transitions. If a female observation transfered from state "Level 1" to "Level 2" and died then, it contributed to the number for "Level 1" and "Level 2" in the first column of above table.

7.2 The Three-State Model

In the case of the three-state model we considered, as mentioned above, the states "Care at home", "Care in a nursing home" and "Death".

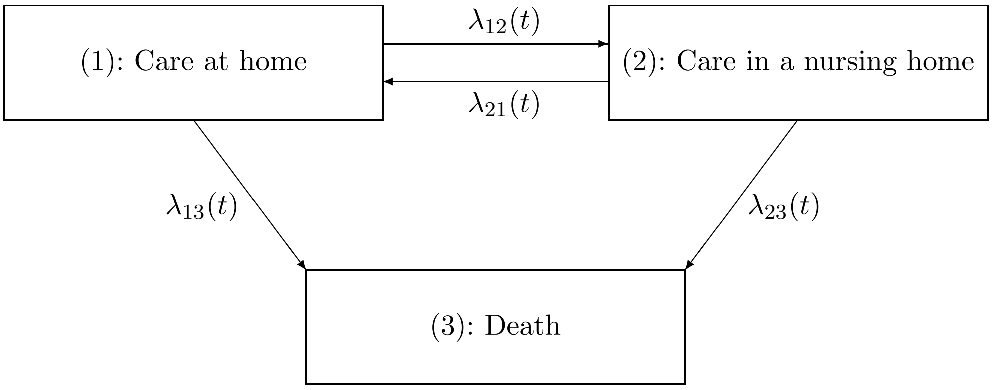

Figure 7.2: Markov model with states "Care at home", "Care in a nursing home" and "Death"

where $\lambda_{ij}(t)$ denotes the hazard-rate function at time t for a transition from state i to state j. We excluded transitions from state 2 to 1, consequently $\lambda_{21}(t) := 0$ for all t, and consider state 3 as an absorbing state for natural reasons. Thus $\lambda_{31}(t)$ and $\lambda_{32}(t)$ are zero, as well. In our dataset we observed the following transitions:

to from	censored	Care at home	Care in a nursing home	Death
Care at home	2222	883	586	1507
Care in a nursing home	1097	0	268	785

where we treated the observations that recovered, that is a transition from "Care in a nursing home" to "Care at home" as censored observations.

We calculated now the Aalen-Johansen estimator as described in Section 4.3 as well as the "leave-one-out" estimator as introduced in Section 5.2. From these quantities, according to Section 5.2, pseudo-values were calculated and a regression model using GEEs is fitted as already indicated in Section 5.1. A detailed description of the functions we used and the functions themselves can be found in Appendix B.1.

We used for each element of the pseudo-transition matrix a normal error distribution. As link-function we chose the logit and assume the variance function, to be constant. This leads to the quasi-likelihood approach using GEEs with possible covariates "Age", "Sex", "Level of Care 2", "Level of Care 3" and "Duration of care", that we collect in the vector $\mathbf{Z}_i$. Therefore our model for a transition probability from state g to state h is

$$\tilde{\theta}_{il}^{(gh)} = \frac{\exp\{\alpha_l^{(gh)} + Z_{i1}\beta_1^{(gh)} + Z_{i2}\beta_2^{(gh)} + Z_{i3}\beta_3^{(gh)} + Z_{i4}\beta_4^{(gh)} + Z_{i5}\beta_5^{(gh)}\}}{1 + \exp\{\alpha_l^{(gh)} + Z_{i1}\beta_1^{(gh)} + Z_{i2}\beta_2^{(gh)} + Z_{i3}\beta_3^{(gh)} + Z_{i4}\beta_4^{(gh)} + Z_{i5}\beta_5^{(gh)}\}} + \varepsilon_{il}^{(gh)}$$

for $i = 1, \ldots, n$ and $l = 1, \ldots, k$, where α_l indicates the duration of care, Z_{i1} an intercept term, Z_{i2} the age at the time of the transition, Z_{i3} the sex Z_{i4} and Z_{i5} the level of care.

To see the similarity of above equation and the notation used for the GEEs in Section 6.4.2 note that above $\tilde{\theta}_{il}^{(gh)}$ corresponds to the y_{it} in Section 6.4.2. In both cases $i = 1, \ldots, n$ indicates the different subjects whereas $l = 1, \ldots, k$ corresponds to $t = 1, \ldots, n_i$ indicating the different times of observation. In our model, the times of observations are equal for all subjects; in other words, k is the same for all subjects in contrast to n_i which was allowed to vary for different subjects.

Note that the covariates "Duration of care", "Sex", "Level of care 2" and "Level of care 3" are factors, which will be dummy coded, i.e.

$$\begin{aligned}
\alpha_l^{(gh)} &:= \begin{cases} 1 & \text{if the duration is in the interval } (l, l+1] \\ 0 & \text{otherwise} \end{cases} \\
Z_{i3} &:= \begin{cases} 0 & \text{if the individual is female} \\ 1 & \text{if the individual is male} \end{cases} \\
Z_{i4} &:= \begin{cases} 0 & \text{if the individual is in level of care 1 or 3} \\ 1 & \text{if the individual is in level of care 2} \end{cases} \\
Z_{i5} &:= \begin{cases} 0 & \text{if the individual is in level of care 1 or 2} \\ 1 & \text{if the individual is in level of care 3} \end{cases}
\end{aligned}$$

For above specified model we performed now a regression analysis using GEEs. First we had to choose the working correlation matrix $R(\alpha)$. In a first approach we decided to allow all possible correlation and thus use the "unstructured" working correlation matrix.

Based on the correlation matrix estimated in this approach we decided then which correlation structure would represent this estimated correlation matrix best and performed another regression using this new correlation structure as working correlation matrix. Further we compared the resulting premiums of both approaches (see Figure 8.2 in Section 8.6).

For p_{12}, the transition probability from state "Care at home" to "Care in a nursing home" we estimated the following "unstructured" correlation matrix:

$$\hat{R}(\alpha) = \begin{pmatrix}
1.00 & 0.05 & -0.04 & 0.03 & 0.02 & 0.01 & 0.01 & 0.00 & 0.02 & -0.01 & 0.07 \\
0.05 & 1.00 & 0.14 & 0.07 & -0.01 & 0.01 & 0.05 & -0.01 & -0.01 & 0.00 & 0.00 \\
-0.04 & 0.14 & 1.00 & -0.01 & 0.00 & 0.03 & 0.14 & -0.02 & -0.01 & 0.00 & -0.01 \\
0.03 & 0.07 & -0.01 & 1.00 & 0.05 & 0.00 & -0.02 & -0.01 & 0.05 & -0.01 & 0.00 \\
0.02 & -0.01 & 0.00 & 0.05 & 1.00 & -0.01 & 0.07 & -0.03 & 0.09 & 0.00 & 0.00 \\
0.01 & 0.01 & 0.03 & 0.00 & -0.01 & 1.00 & 0.00 & -0.02 & -0.01 & -0.01 & -0.01 \\
0.01 & 0.05 & 0.14 & -0.02 & 0.07 & 0.00 & 1.00 & 0.06 & -0.01 & 0.00 & 0.10 \\
0.00 & -0.01 & -0.02 & -0.01 & -0.03 & -0.02 & 0.06 & 1.00 & -0.01 & 0.12 & -0.02 \\
0.02 & -0.01 & -0.01 & 0.05 & 0.09 & -0.01 & -0.01 & -0.01 & 1.00 & 0.01 & -0.01 \\
-0.01 & 0.00 & 0.00 & -0.01 & 0.00 & -0.01 & 0.00 & 0.12 & 0.01 & 1.00 & -0.04 \\
0.07 & 0.00 & -0.01 & 0.00 & 0.00 & -0.01 & 0.10 & -0.02 & -0.01 & -0.04 & 1.00
\end{pmatrix}$$

Since the values in this correlation matrix are negligible we choose for the transition probability p_{12} the "independence" working correlation matrix and obtained after 37 iterations the following estimates:

Call:

gee(formula = p12 ∼ Age + C(factor(Sex), treatment) + C(factor(ZLevel), treatment) + C(factor(time), treatment), family = quasi(link = logit, variance = constant), data = geedata11, subject = id, repeated = time, wc = "ind", QR = T)

Coefficients:

	Values	Stderr	t-values	$Pr(\lvert t\rvert >)$
Intercept	-5.31	0.98	-5.43	0.00
Age	0.02	0.01	1.95	0.05
Sex	0.66	0.23	2.90	0.00
Level of care 2	-1.37	0.37	-3.69	0.00
Level of care 3	-1.38	0.46	-3.04	0.00
Duration of care 1	0.35	0.34	1.04	0.30
Duration of care 2	0.40	0.35	1.15	0.25
Duration of care 3	0.90	0.32	2.77	0.01
Duration of care 4	1.24	0.33	3.79	0.00
Duration of care 5	0.10	0.53	0.19	0.85
Duration of care 6	1.53	0.35	4.43	0.00
Duration of care 7	-0.11	0.72	-0.15	0.88
Duration of care 8	0.25	0.48	0.52	0.60
Duration of care 9	0.80	0.50	1.58	0.11
Duration of care 10	1.26	0.43	2.93	0.00

Degrees of Freedom: 80828 Total; 80813 Residual

Since we assumed in this case the "independence" working correlation matrix, the correlation matrix is fixed to the identity matrix and does not need to estimated.

In contrast to p_{12} we obtained for p_{13} larger correlations using the "unstructured" working correlation matrix:

$$\hat{R}(\alpha) = \begin{pmatrix} 1.00 & 0.42 & 0.37 & 0.36 & 0.36 & 0.24 & 0.16 & 0.12 & 0.26 & 0.04 & 0.02 \\ 0.42 & 1.00 & 0.41 & 0.27 & 0.24 & 0.14 & 0.13 & 0.07 & 0.13 & 0.01 & 0.00 \\ 0.37 & 0.41 & 1.00 & 0.44 & 0.27 & 0.21 & 0.17 & 0.12 & 0.49 & 0.14 & 0.02 \\ 0.36 & 0.27 & 0.44 & 1.00 & 0.34 & 0.22 & 0.20 & 0.17 & 0.49 & 0.18 & 0.12 \\ 0.36 & 0.24 & 0.27 & 0.34 & 1.00 & 0.34 & 0.16 & 0.07 & 0.31 & 0.10 & 0.00 \\ 0.24 & 0.14 & 0.21 & 0.22 & 0.34 & 1.00 & 0.34 & 0.19 & 0.68 & 0.18 & 0.01 \\ 0.16 & 0.13 & 0.17 & 0.20 & 0.16 & 0.34 & 1.00 & 0.52 & 1.20 & 0.44 & 0.14 \\ 0.12 & 0.07 & 0.12 & 0.17 & 0.07 & 0.19 & 0.52 & 1.00 & 1.26 & 0.38 & 0.13 \\ 0.26 & 0.13 & 0.49 & 0.49 & 0.31 & 0.68 & 1.20 & 1.26 & 1.00 & 1.28 & 0.24 \\ 0.04 & -0.01 & 0.14 & 0.18 & 0.10 & 0.18 & 0.44 & 0.38 & 0.13 & 1.00 & 0.33 \\ 0.02 & 0.00 & 0.02 & 0.12 & 0.00 & 0.01 & 0.14 & 0.13 & 0.24 & 0.33 & 1.00 \end{pmatrix}$$

The correlations tend to decrease when the distance in time increases, therefore we considered an "autoregressive (AR-I)" correlation structure sufficient for this transition probability giving the following estimates:

Call:

gee(formula = p13 ∼ Age + C(factor(Sex), treatment) + C(factor(ZLevel), treatment) + C(factor(time), treatment), family = quasi(link = logit, variance = constant), data = geedata11, subject = id, repeated = time, wc = "aut", QR = T)

Coefficients:

	Values	Stderr	t-values	$Pr(\|t\| >)$
Intercept	-2.16	0.65	-3.32	0.00
Age	0.02	0.01	2.42	0.02
Sex	0.36	0.20	1.81	0.07
Level of care 2	-0.69	0.21	-3.22	0.00
Level of care 3	-0.53	0.23	-2.25	0.02
Duration of care 1	0.02	0.06	0.37	0.71
Duration of care 2	0.18	0.07	2.70	0.01
Duration of care 3	0.88	0.09	1.03	0.00
Duration of care 4	0.99	0.12	8.55	0.00
Duration of care 5	0.74	0.14	5.38	0.00
Duration of care 6	0.41	0.16	2.65	0.01
Duration of care 7	0.24	0.18	1.34	0.18
Duration of care 8	0.82	0.24	3.49	0.00
Duration of care 9	1.19	0.33	3.65	0.00
Duration of care 10	-0.48	0.31	-1.55	0.12

Degrees of Freedom: 80828 Total; 80813 Residual

The estimate for the correlation matrix in this case is as follows:

$$\hat{R}(\alpha) = \begin{pmatrix} 1.00 & 0.39 & 0.15 & 0.06 & 0.02 & 0.01 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.39 & 1.00 & 0.39 & 0.15 & 0.06 & 0.02 & 0.01 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.15 & 0.39 & 1.00 & 0.39 & 0.15 & 0.06 & 0.02 & 0.01 & 0.00 & 0.00 & 0.00 \\ 0.06 & 0.15 & 0.39 & 1.00 & 0.39 & 0.15 & 0.06 & 0.02 & 0.01 & 0.00 & 0.00 \\ 0.02 & 0.06 & 0.15 & 0.39 & 1.00 & 0.04 & 0.15 & 0.06 & 0.02 & 0.01 & 0.00 \\ 0.01 & 0.02 & 0.06 & 0.15 & 0.39 & 1.00 & 0.39 & 0.15 & 0.06 & 0.02 & 0.01 \\ 0.00 & 0.01 & 0.02 & 0.06 & 0.15 & 0.39 & 1.00 & 0.39 & 0.15 & 0.06 & 0.02 \\ 0.00 & 0.00 & 0.01 & 0.02 & 0.06 & 0.15 & 0.39 & 1.00 & 0.39 & 0.15 & 0.06 \\ 0.00 & 0.00 & 0.00 & 0.01 & 0.02 & 0.06 & 0.15 & 0.39 & 1.00 & 0.39 & 0.15 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.01 & 0.02 & 0.06 & 0.15 & 0.39 & 1.00 & 0.39 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.01 & 0.02 & 0.06 & 0.15 & 0.39 & 1.00 \end{pmatrix}$$

In the case of p_{23} we also decided to use the "autoregressive (AR-I)" working correlation matrix, since a similar behavior in the correlation estimate of the "unstructured" working correlation matrix can be observed:

$$\hat{R}(\alpha) = \begin{pmatrix} 1.00 & 0.29 & 0.18 & 0.19 & 0.09 & 0.08 & -0.01 & 0.05 & 0.03 & 0.01 & 0.09 \\ 0.29 & 1.00 & 0.31 & 0.24 & 0.29 & 0.05 & 0.04 & -0.01 & -0.01 & 0.00 & -0.03 \\ 0.18 & 0.31 & 1.00 & 0.37 & 0.18 & 0.16 & 0.04 & 0.07 & 0.08 & 0.00 & -0.02 \\ 0.19 & 0.24 & 0.37 & 1.00 & 0.12 & 0.11 & 0.15 & 0.10 & 0.00 & 0.00 & -0.02 \\ 0.09 & 0.29 & 0.18 & 0.12 & 1.00 & 0.24 & 0.13 & 0.12 & 0.00 & 0.00 & -0.01 \\ 0.08 & 0.05 & 0.16 & 0.11 & 0.24 & 1.00 & 0.12 & 0.29 & 0.13 & 0.08 & 0.04 \\ -0.01 & 0.04 & 0.04 & 0.15 & 0.13 & 0.12 & 1.00 & 0.35 & 0.21 & 0.14 & 0.30 \\ 0.05 & -0.01 & 0.07 & 0.10 & 0.12 & 0.29 & 0.35 & 1.00 & 0.42 & 0.13 & 0.26 \\ 0.03 & -0.01 & 0.08 & 0.00 & 0.00 & 0.13 & 0.21 & 0.42 & 1.00 & 0.28 & 0.83 \\ 0.01 & 0.00 & 0.00 & 0.00 & 0.00 & 0.08 & 0.14 & 1.33 & 0.28 & 1.00 & 1.22 \\ 0.09 & -0.03 & -0.02 & -0.02 & -0.01 & 0.04 & 0.30 & 0.26 & 0.83 & 1.22 & 1.00 \end{pmatrix}$$

As estimates for p_{23} using the "autoregressive (AR-I)" working correlation matrix we obtained

Call:

gee(formula = p23 ~ Age + C(factor(Sex), treatment) + C(factor(ZLevel), treatment) + C(factor(time), treatment), family = quasi(link = logit, variance = constant), data = geedata11, subject = id, repeated = time, wc = "aut", QR = T)

Coefficients:

	Values	Stderr	t-values	$Pr(\|t\| >)$
Intercept	7.40	1.88	3.93	0.00
Age	-0.06	0.02	-3.11	0.00
Sex	1.51	0.43	3.52	0.00
Level of care 2	-1.96	0.47	-4.13	0.00
Level of care 3	-2.27	0.55	-4.16	0.00
Duration of care 1	0.34	0.26	1.30	0.19
Duration of care 2	-1.08	0.28	-3.90	0.00
Duration of care 3	-0.37	0.28	-1.30	0.19
Duration of care 4	-1.68	0.36	-4.67	0.00
Duration of care 5	-1.96	0.39	-5.03	0.00
Duration of care 6	-1.06	0.41	-2.57	0.01
Duration of care 7	-0.65	0.46	-1.41	0.16
Duration of care 8	-0.09	0.62	-0.15	0.88
Duration of care 9	-2.28	0.63	-3.64	0.00
Duration of care 10	-0.03	0.82	-0.03	0.97

Degrees of Freedom: 80828 Total; 80813 Residual

Giving a estimated correlation matrix of

$$\hat{R}(\alpha) = \begin{pmatrix}
1.00 & 0.40 & 0.16 & 0.06 & 0.03 & 0.01 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\
0.40 & 1.00 & 0.40 & 0.16 & 0.06 & 0.03 & 0.01 & 0.00 & 0.00 & 0.00 & 0.00 \\
0.16 & 0.40 & 1.00 & 0.40 & 0.16 & 0.06 & 0.03 & 0.01 & 0.00 & 0.00 & 0.00 \\
0.06 & 0.16 & 0.40 & 1.00 & 0.40 & 0.16 & 0.06 & 0.03 & 0.01 & 0.00 & 0.00 \\
0.03 & 0.06 & 0.16 & 0.40 & 1.00 & 0.40 & 0.16 & 0.06 & 0.03 & 0.01 & 0.00 \\
0.01 & 0.03 & 0.06 & 0.16 & 0.40 & 1.00 & 0.40 & 0.16 & 0.06 & 0.03 & 0.01 \\
0.00 & 0.01 & 0.03 & 0.06 & 0.16 & 0.40 & 1.00 & 0.40 & 0.16 & 0.06 & 0.03 \\
0.00 & 0.00 & 0.01 & 0.03 & 0.06 & 0.16 & 0.40 & 1.00 & 0.40 & 0.16 & 0.06 \\
0.00 & 0.00 & 0.00 & 0.01 & 0.03 & 0.06 & 0.16 & 0.40 & 1.00 & 0.40 & 0.16 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.01 & 0.03 & 0.06 & 0.16 & 0.40 & 1.00 & 0.40 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.01 & 0.03 & 0.06 & 0.16 & 0.40 & 1.00
\end{pmatrix}$$

7.3 Results

In the following table we summarized the estimates obtained from the GEEs for all possible transition probabilities of the model shown in Figure 7.2:

	p_{12}		p_{13}		p_{23}	
	Values	Pr($\|t\|$ >)	Values	Pr($\|t\|$ >)	Values	Pr($\|t\|$ >)
Intercept	-5.31	0.00	-2.16	0.00	7.40	0.00
Age	0.02	0.05	0.02	0.02	-0.06	0.00
Sex	0.66	0.00	0.36	0.07	1.51	0.00
Level of care 2	-1.37	0.00	-0.69	0.00	-1.96	0.00
Level of care 3	-1.38	0.00	-0.53	0.02	-2.27	0.00
Duration of care 1	0.35	0.30	0.02	0.71	0.34	0.19
Duration of care 2	0.40	0.25	0.18	0.01	-1.08	0.00
Duration of care 3	0.90	0.01	0.88	0.00	-0.37	0.19
Duration of care 4	1.24	0.00	0.99	0.00	-1.68	0.00
Duration of care 5	0.10	0.85	0.74	0.00	-1.96	0.00
Duration of care 6	1.53	0.00	0.41	0.01	-1.06	0.01
Duration of care 7	-0.11	0.88	0.24	0.18	-0.65	0.16
Duration of care 8	0.25	0.60	0.82	0.00	-0.09	0.88
Duration of care 9	0.80	0.11	1.19	0.00	-2.28	0.00
Duration of care 10	1.26	0.00	-0.48	0.12	-0.03	0.97

The transition probabilities can now be derived from these estimates for any given set of covariates simply by calculating the mean function, the inverse of the logit, the link function we have chosen. Note that this is a non-linear function.

Generally the transition probabilities p_{12} and p_{13} increase with age, whereas p_{23} decreases but very slightly. The values of all three transition probabilities are higher for males than the values for females as indicated by the high values of the covariate "Sex".

Transitions out of state "Care at home" are more likely to happen to individuals in "Level 1", whereas the values for individuals in "Level 2" and "Level 3" are nearly the same. This can be understood looking at the values for "Level of care 2" and "Level of care 3" that are nearly the same for these transition probabilities. The probability of dying in state "Care at home" decreases from "Level 1" to "Level 2" before it increases to "Level 3" but does not reach the value from "Level 1" again. In contrast, for p_{23} one observes decreasing transition probabilities for an increase in the severeness of care needed.

The "Duration of Care" causes rising transition probabilities for p_{12} and p_{13} until a duration of four years, then a decrease can be observed for three years until they increase again. For the transition from state "Care in a nursing home" to "Death" the values are very close together for all durations and only low changes occur.

Chapter 8

The Multiple-State Insurance Model

In the last chapter we showed how one-year transition probabilities can be derived in a multi-state model. These transition probabilities are necessary to calculate LTC premiums in an insurance model. To calculate an insurance premium we have to define, when premiums and benefits have to be paid, and what kind of benefits are under which circumstances due.

In this chapter we define firstly possible inflows and outflows in a multiple-state insurance model. Then we introduce the concept of random present values that is used to compare random cash-flows that occur at different points in time. A discount factor also has to be assigned. Calculating the expectation of these random present values leads us to actuarial values.

The principle of equivalence is then used to determine the premium to be charged for a given LTC-plan, that is we choose the premium such that the actuarial values of the inflows is equal to the actuarial values of the outflows at the time the policy is issued. Further we explain how reserves are calculated that mark a further constraint on the premiums: The so-called "funding condition" states, that reserves should be greater or equal zero at any time while the insurance contract is in force.

In the last section of this chapter we apply these tools to our three-state model. To do so, we have to add the state "Active" to our model: Lives in the state "Active" pay premiums, lives in the states "Care at home" and "Care in a nursing home" receive benefits until they transfer eventually to state "Death".

Since our data only provided information on LTC claimants we had to use additional sources to calculate transition probabilities from state "Active" to all other three states. We used incidence rates from "Custodial Insurance, Japan" (Appendix B.3) together with the relative frequency of a person becoming LTC claimant and needing care at home or in a nursing home that were observed in our data over a two-year period (Section 8.6) and the mortality rates from "Bavarian life tables 1986-1988" (Appendix B.2).

We discretized time-continuous quantities in our model and used an interest rate of 3.5% per annum. Annual premiums were paid at the start of the year. If a live became LTC claimant it received a certain allowance depending on the place and level of care needed: At home the insurance company paid 25% in "Level 1", 50% in "Level 2" and 75% in "Level 3" of the allowance and in a nursing home 100% of the allowance for all three levels. The actuarial values were calculated using a C-program and the necessary premiums derived for a 10 EUR daily allowance, seperately for female and males for ages 20 to 70.

8.1 Benefits and Premiums

In Chapter 3 we modeled the life-history of an individual as a time-continuous Markovian process $S(t)$, using a multi-state model with finite state space $\mathcal{S} = \{1, \dots, K\}$. At each point in time the individual is in exactly one state and potentially transfers, as time proceeds, from one state to the next one until eventually the state "Death" is reached.

Modeling an insurance contract cash-flows occur: Premiums are paid and benefits received by the insured corresponding to the states visited. As the sample path $s(t)$ of the time-continuous Markovian process $S(t)$ is different for each individual, so are the cash-flows. From the insurance company point of view we distinguish between inflows (e.g the premiums paid by the insured) and outflows (e.g. annuity benefits or lump sums paid by the insurer).

Generally the following types of premiums or benefits are possible. For further details and examples see Haberman and Pitacco (1999):

Inflows:

- a continuous premium at a rate $p_i(t)$ at time t, if $S(t) = i$;
- a premium $\pi_i(t)$ at some fixed time t, if $S(t) = i$;

Outflows:

- a continuous annuity benefit at rate $b_j(t)$ at time t, if $S(t) = j$;
- a lump sum $c_{ij}(t)$, if at time t a transition occurs from state i to state j;
- a lump sum $d_j(t)$ at some fixed time t, if $S(t) = j$, a so-called pure endowment;

We denote by $p_i(t)dt$ the premium amount and by $b_j(t)dt$ the benefit amount paid out in the infinitesimal interval $[t, t + dt)$, respectively. If we assume that $S(0) = 1$, for example in the Illness-Death model all individuals start from the "Disease-Free" state, $\pi_1(0)$ might represent an initial single premium and all other premium functions remain zero.

Further the cumulative premium function and cumulative annuity benefit function are denoted by $\Pi_i(t)$ and $B_j(t)$, respectively. Both are non-negative and non-decreasing functions and will be precisely defined in Section 8.4.

If the individual is in state i in the interval $[t, u)$ a premium of $\Pi_i(u) - \Pi_i(t)$ is due. If the individual is, in contrast, in state j in the interval $[t, u)$ an annuity benefit of $B_j(u) - B_j(t)$ is due. Pure endowments cause jumps in the function $B_j(t)$.

8.2 Random Present Values

Actuarial values, that we are going to define in Section 8.3 more explicitly, are expected present values. Assume that the future premiums paid by the insured were known at the present time as well as the future benefits that the insured receives. These cash-flows generally happen at different points in time. In order to compare the value of the inflows and outflows for the insurer, one uses the concept of present values, that is outflows and inflows are deflated to the present time. The resulting quantities are then comparable given that the deflation factor is correct specified. In insurance this assumption does not hold: Inflows and outflows are random quantities. Therefore random present values have to be determined. A definition of a random present values is necessary. Consider the compound interest model. The force of interest δ is assumed to be deterministic and constant. Thus the annual discount factor, denoted by v, is

$$v = \exp\{\ -\delta\}$$

Consider a continuous premium at rate $p_j(u)$ at time u, if $S(u) = j$. As already mentioned, $p_j(u)du$ is the premium amount paid out in the infinitesimal interval $[u, u + du)$. The random present value of this premium at time t is given by

$$Y_t^{p_j}(u, u + du) := v^{u-t} I_{\{S(u)=j\}} p_j(u) du$$

The same continuous premium paid over the time interval $[u_1, u_2)$, with $t \le u_1 < u_2$, has the following random present value at time t:

$$Y_t^{p_j}(u_1, u_2) := \int_{u_1}^{u_2} v^{u-t} I_{\{S(u)=j\}} p_j(u) du$$

Consider a premium $\pi_j(u)$ at some fixed time u, if $S(u) = j$. The random present value of this benefit at time t is given by

$$Y_t^{\pi_j}(u) := v^{u-t} I_{\{S(u)=i\}} \pi_j(u)$$

Consider a continuous annuity benefit at a rate $b_j(u)$ at time u, if $S(u) = j$. Again, $b_j(u)du$ is the benefit amount paid out over the infinitesimal interval $[u, u+du)$. The random present value of this benefit at time t is given by

$$Y_t^{b_j}(u, u + du) := v^{u-t} I_{\{S(u)=j\}} b_j(u) du$$

The same continuous annuity benefit on the time interval $[u_1, u_2)$, with $t \le u_1 < u_2$, has the following random present value at time t:

$$Y_t^{b_j}(u_1, u_2) := \int_{u_1}^{u_2} v^{u-t} I_{\{S(u)=j\}} b_j(u) du$$

Consider a lump sum $c_{jk}(u)$, paid just after time u, if a transition from state j to k occurs at time u. The random present value of this lump sum at time t is given by

$$Y_t^{c_{jk}}(u) := v^{u-t} I_{\{S(u-)=j,\, S(u)=k\}} c_{jk}(u)$$

The random present value at time t of this lump sum, paid for each transition from state j to state k in the interval $[u_1, u_2)$ is given by

$$Y_t^{c_{jk}}(u_1, u_2) := \int_{u_1}^{u_2} v^{u-t} c_{jk}(u) dN_{jk}(u)$$

where $N_{jk}(u)$ is the number of transitions from state j to k in the interval $[0, u)$. Consequently $dN_{jk}(u)$ is the number of transitions from state j to k in the interval $[u_1, u_2)$.

Consider the pure endowment, a lump sum benefit $d_j(u)$ at some fixed time u, if $S(u) = j$. The random present value at time t is given by

$$Y_t^{d_j}(u) := v^{u-t} I_{\{S(u)=j\}} d_j(u)$$

In the following sections these random present values will be used to calculate the actuarial values, which are the basic tool to determine premiums and reserves.

8.3 Actuarial Values

As mentioned before actuarial values are expected present values. In addition to the financial structure of random present values we need now a probabilistic structure, as well. This is where the assumption of the life-history or claim-history as a time-continuous Markov chain comes in. Further we suppose that the risk is in state i at time t, that is $S(t) = i$, and define the actuarial values as a conditioning event. Actuarial values are therefore conditional expected present values. Following Czado and Rudolph (2002) we define:

Definition 8.1 (Actuarial values) *Actuarial values are expected present values. Assuming that the insured risk is in state i at time t, then the actuarial values are given as conditional expectations of the random present values, that is*

- $E[Y_t(u)|\, S(t) = i]$ *for lump sum payments*
- $E[Y_t(u, u + du)|\, S(t) = i]$ *for annuities*

In the following we are going to specify the actuarial values for the random present values introduced in Section 8.2:

The actuarial value of the continuous premium at rate $p_j(u)$ at time u, if $S(u) = j$, is

$$E\left[Y_t^{p_j}(u, u + du)\middle|\, S(t) = i\right] = v^{u-t} p_{ij}(t, u) p_j(u) du$$

$$E\left[Y_t^{p_j}(u_1, u_2)\middle|\, S(t) = i\right] = \int_{u_1}^{u_2} v^{u-t} p_{ij}(t, u) p_j(u) du$$

The actuarial value of a lump sum $\pi_j(u)$ paid at some fixed time u, if $S(u) = j$, is

$$E\left[Y_t^{\pi_j}(u)\middle|\, S(t) = i\right] = v^{u-t} p_{ij}(t, u) \pi_j(u)$$

The actuarial value of the continuous annuity benefit at rate $b_j(u)$ at time u, if $S(u) = j$, is

$$E\left[Y_t^{b_j}(u, u + du)\middle|\, S(t) = i\right] = v^{u-t} p_{ij}(t, u) b_j(u) du$$

$$E\left[Y_t^{b_j}(u_1, u_2)\middle|\, S(t) = i\right] = \int_{u_1}^{u_2} v^{u-t} p_{ij}(t, u) b_j(u) du$$

The actuarial value of a lump sum c_{jk} paid just after time u, if a transition from state j to k occurs at time u, is

$$E\left[Y_t^{c_{jk}}(u)\middle|\, S(t) = i\right] = v^{u-t} p_{ij}(t, u) \mu_{jk}(u) c_{jk}(u)$$

$$E\left[Y_t^{c_{jk}}(u_1, u_2)\middle|\, S(t) = i\right] = \int_{u_1}^{u_2} v^{u-t} p_{ij}(t, u) \mu_{jk} c_{jk}(u) du$$

The actuarial value of a lump sum $d_j(u)$ paid at some fixed time u, if $S(u) = j$, is

$$E\left[Y_t^{d_j}(u)\middle|\, S(t) = i\right] = v^{u-t} p_{ij}(t, u) d_j(u)$$

The actuarial notation considers usually unit-level premiums or unit annuities. Using these we derive the following quantities:

For a continuous unit premium at a rate $p_j(u)$ or a continuous unit annuity benefit $b_j(u)$ paid in state j during the period $[t, n)$:

$$\bar{a}_{ij}(t,n) = \int_t^n v^{u-t} p_{ij}(t,u) du$$

For an unit premium $\pi_j(u)$ at some fixed time t:

$$\bar{E}_{ij}(t,u) = v^{u-t} p_{ij}(t,u)$$

For an unit lump sum $c_{jk}(u)$ at time t, if a transition occurs from state j to k:

$$\begin{aligned} \bar{A}_{ijk}(t,n) &= \int_t^n v^{u-t} p_{ij}(t,u) \mu_{jk}(u) du \\ \bar{A}_{i.k}(t,n) &= \sum_{j:j \neq k} \bar{A}_{ijk}(t,n) \\ \bar{A}_{ij.}(t,n) &= \sum_{k:k \neq j} \bar{A}_{ijk}(t,n) \end{aligned}$$

For an unit lump sum $d_j(u)$ at some fixed time t:

$$\bar{E}_{ij}(t,u) = v^{u-t} p_{ij}(t,u)$$

8.4 The Principle of Equivalence

If we take all above defined premiums together, we obtain the cumulative premium function. The actuarial value of the continuous-time premium paid at rate $p_j(u)$ at time t, if $S(t) = j$, is given by the following quantity:

$$\mathcal{P}_i(t,n) = \int_t^n v^{u-t} \sum_{j \in S} p_{ij}(t,u) p_j(u) du$$

where n is the policy term. Further we have the actuarial value of the discrete-time premiums $\pi_j(t)$, that is

$$\Pi_i(t) = \sum_{u:u \geq t} v^{u-t} \sum_{j \in S} p_{ij}(t,u) \pi_j(u)$$

Taking all the benefits mentioned above together we obtain the cumulative benefit function $\mathcal{B}_i(t,n)$, that is

$$\begin{aligned} \mathcal{B}_i(t,n) &= \int_t^n v^{u-t} \sum_{j \in S} p_{ij}(t,u) b_j(u) du \\ &+ \int_t^n v^{u-t} \sum_{j \in S} \sum_{k:k \neq j} p_{ij}(t,u) \mu_{jk}(u) c_{jk}(u) du \\ &+ \sum_{u:u \geq t} v^{u-t} \sum_{j \in S} p_{ij}(t,u) d_j(u) \end{aligned}$$

The principle of equivalence states, as mentioned by Czado and Rudolph (2002), that the expected amount of premiums has to be equal to the expected amount of benefits. At the time when the policy is issued, the actuarial value of the benefits, that are paid under this contract, has to be the same as the actuarial value of the premiums, that are received by the insurer. This can formally be described as follows:

Definition 8.2 (The Principle of Equivalence) *For an insured risk with policy end at n and initial state $S(0) = 1$, the equivalence principle is satisfied if and only if*

$$\mathcal{P}_1(0,n) = \mathcal{B}_1(0,n)$$

or equivalently in the case of a discrete-time premium at time t

$$\Pi_1(0,n) = \mathcal{B}_1(0,n)$$

Clearly this relationship might be fulfilled by an infinity of premium functions. The so-called "funding condition" is a further constraint. At any time during the insurance contract is in force we require that

$$\mathcal{B}_{S(t)}(t,n) \geq \mathcal{P}_{S(t)}(t,n)$$

As the principle of equivalence only has to be fulfilled at policy begin, this also enables us to construct insurance contracts with increasing, decreasing or level premiums according to given laws or customers needs.

8.5 Calculation of Reserves

For every insurance contract reserves have to be built up in order to ensure that the insurance company is able to fulfill the promises given in their contracts with the insured. The prospective reserve at time t is defined as the actuarial value of future benefits less the actuarial value of future premiums. Generally speaking, the reserve is equal to the expected amount the insurance company has to pay in the future reduced by the expected amount the insured pays in the future to the insurance company. Given $S(t) = i$, that is

$$\bar{V}_i(t) = \mathcal{B}_i(t,n) - \mathcal{P}_i(t,n)$$

or equivalently in the case of a discrete-time premium at time t:

$$\bar{V}_i(t) = \mathcal{B}_i(t,n) - \Pi_i(t,n)$$

Note that for each state i, that possibly is occupied at time t, a reserve has to be calculated. Consider now an insurance contract, offering benefits $b_j(u)$ and $c_{jk}(u)$ with continuous premium $p_j(u)$, then the prospective reserve is given by

$$\begin{aligned}
\bar{V}_i(t) &= \mathcal{B}_i(t,n) - \mathcal{P}_i(t,n) \\
&= \int_t^n v^{u-t} \sum_{j \in S} p_{ij}(t,u) b_j(u) du + \sum_{j \in S} \sum_{k:k \neq j} p_{ij}(t,u) \mu_{jk}(u) c_{jk}(u) du \\
&- \int_t^n v^{u-t} \sum_{j \in S} p_{ij}(t,u) p_j(u) du \\
&= \int_t^n v^{u-t} \sum_{j \in S} \left(p_{ij}(t,u) b_j(u) + \sum_{k:k \neq j} p_{ij}(t,u) \mu_{jk}(u) c_{jk}(u) - p_{ij}(t,u) p_j(u) \right) du
\end{aligned}$$

Clearly this reserve changes over time driven by underlying quantities such as premium and benefits payment as well as changes on interest and mortality assumptions.

8.6 Calculation of Premiums

So far we modeled only transitions for individuals that already qualified for one of the three levels of care (see Section 7). If we extend our model to the situation of an insurance company, we have to add the state "Active" to our model. In Figure 8.1 we added this state. For the area within the dotted line we were able to calculate the necessary probabilities using our data. For transition probabilities from outside this area additional information is necessary.

Since our data does not provide this information, we used published tables to account for the incidence of care and the mortality of individuals in the "Active" state. Namely we used incidence rates from "Custodial Insurance, Japan" (Appendix B.3) and mortality rates from the "Bavarian life tables 1986-1988" (Appendix B.2).

The incidence rate is the probability of an z-year old individual in the "Active" state to transfer to any state where LTC is needed within the next year.

$$i_z := P\left(S(z+1) = j, \quad j \in \{2,3\} \middle| \, S(z) = 1 \right)$$

Since the incidence rates from "Custodial Insurance, Japan" only distinguish between age and sex, we use the relative frequencies of the place of care, that occurred in our data between January 1^{st}, 1997 and December 31^{st}, 1998, to make a further distinction with respect to the place of care. The relative frequency were the following:

Place of care	Female (in %)	Male (in %)
at home	81.76	85.38
in a nursing home	18.24	14.62

The mortality rates of the so-called "active-life" are usually denoted by q_z^a. This is the probability of an z-year individual in the "Active" state to die within the next year given, that this individual has survived up to age z.

$$q_z^a := P\left(z < T \leq z+1 \middle| \, T > z, \text{ "Active" } \right)$$

In the three-state model transitions only between places of care are possible; transitions between different levels of care are not accounted for. To provide still for this we calculate the necessary actuarial values for each level and use then a weighted average, where weighting was performed with the average duration in the corresponding level:

	Female (in %)	Male (in %)
Level 1	36.44	33.28
Level 2	39.75	40.83
Level 3	23.81	25.89

Lets have a look to the new model:

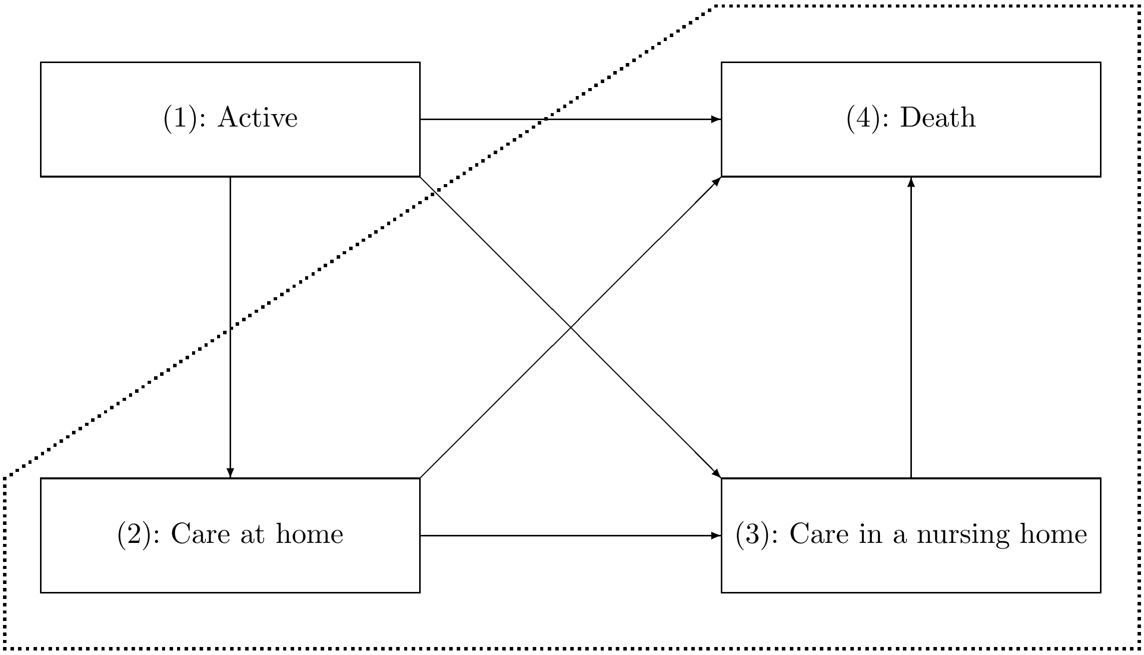

Figure 8.1: Markov four-state Insurance Model

We need for every combination of care duration l and age z the one-year transition probabilities, denoted by $\mathbf{P}_{gh}(l, z)$. This is the probability of an z-year old individual with a care duration of l years to transfer from state g to state h within a year.

$$\mathbf{P}_{gh}(l, z) := P\left(S(l+1) = h \middle| S(l) = g, Age = z\right)$$

After having estimated these probabilities we are able to determine the necessary actuarial values based on these estimates. We consider an insurance contract that pays a lump sum c_{1j}, if a transition from state 1 (”Active”) to state j occurs, where j might be state 2 (”Care at home”) or state 3 (”Care in a nursing home”). We denote the actuarial value of this contract at policy begin by $B_{1,c_{1j}}(0)$.

Further we have to specify a limiting age ω, such that the probability that any individual survives beyond age ω is equal to zero.

From Section 8.3 we know that the actuarial value of such a lump sum c_{1j} can be calculated as

$$\mathcal{B}_{1,c_{1j}} = \sum_{i=0}^{w-z-1} P_{11}(0,i) p_{1j}(i) v^i c_{1j}$$

where $P_{11}(0, i) = P\left(S(i) = 1 \middle| S(0) = 1\right)$ for $0 \leq i$ is the probability of not becoming claimant in the interval $[0, i]$ and $p_{1j}(i) = P(S(i+1) = j| S(i) = 1)$ for $j \neq 1$ is the probability to transfer to state j within the next year after having been in state 1 at time i. $S(t)$ is the Markov chain representing the claim-history of an individual.

Further on this insurance contract, an annuity b_j is paid, while the individuals is claimant in state $j \in \{2,3\}$. The actuarial value of this annuity is given by

$$\mathcal{B}_{1,b_j} = \sum_{i=0}^{w-z-1} P_{1j}(0,i)v^i b_j$$

The insured has to pay an single-annual premium π for this insurance cover, which has an actuarial value of

$$\mathcal{P}_{1,\pi} = \sum_{i=0}^{w-z-1} P_{11}(0,i)v^i \pi$$

In accordance with the principle of equivalence the actuarial value of the benefits paid and the premiums received should be zero at time 0. Thus we have

$$\sum_{j=2}^{3} \mathcal{B}_{1,c_{1j}} + \sum_{j=2}^{3} \mathcal{B}_{1,b_j} = \mathcal{P}_{1,\pi}$$

$$\sum_{j=2}^{3} \sum_{i=0}^{w-z-1} P_{11}(0,i)p_{1j}(i,z)v^i c_{1j} + \sum_{j=2}^{3} \sum_{i=0}^{w-z-1} P_{1j}(0,i)v^i b_j = \sum_{i=0}^{w-z-1} P_{11}(0,i)v^i \pi$$

Solving this equation for π gives the necessary premium for any given set of annuities b_j's and lump sums c_{1j}'s.

In the following we calculate the premiums for the LTC-plan "PET" sold by a German insurer. According to this LTC-plan the insured receives a certain allowance depending on the level of care needed. This is for "Care at home" 25% in "Level 1", 50% in "Level 2" and 75% in "Level 3" and for "Care in a nursing home" 100% of the allowance. Thus, the c_{1j}'s are zero and in the case of "Care at home" $b_j = 1 - 0.25 * (4 - j)$, where $j \in \{1,2,3\}$, for an unit allowance and in the case of "Care in a nursing home" $b_j = 1$, $j \in \{1,2,3\}$.

For the calculation of the premiums we use a modified version of a C-program, which needs the benefits, interest rate and transition probabilities as input. For details see Rudolph (2000). We obtained for a 10 EUR daily allowance the following premiums:

Age	Female	Male	Age	Female	Male
20	04.94	03.85	50	21.01	16.95
25	06.14	04.80	55	27.58	22.40
30	07.70	06.05	60	36.48	29.88
35	09.76	07.72	65	48.53	40.06
40	12.51	09.97	70	64.55	53.62
45	16.14	12.95			

Table 8.1: Premiums fro a 10 EUR daily allowance

In the following table we calculated the premiums for the same LTC-plan with estimates for the transition probabilities using "unstructured" working correlation matrices only, to compare them with the premiums from Table 8.1 in Figure 8.2:

Age	Female	Male	Age	Female	Male
20	04.98	03.94	50	21.16	17.35
25	06.19	04.91	55	27.74	22.93
30	07.76	06.19	60	36.63	30.58
35	09.84	07.90	65	48.58	40.97
40	12.61	10.21	70	64.28	54.78
45	16.27	13.26			

Table 8.2: Premiums fro a 10 EUR daily allowance

Figure 8.2 compares the premiums from Table 8.1 calculated with the transition probabilities from our model with a "independence" working correlation matrix for p_{12} and a "autoregressive (AR-I)" working correlation matrix for both transition probabilities, p_{13} and p_{23} (Model I), with the premiums 8.2 calculated with the transition probabilities based on the "unstructured" working correlation matrix (Model II).

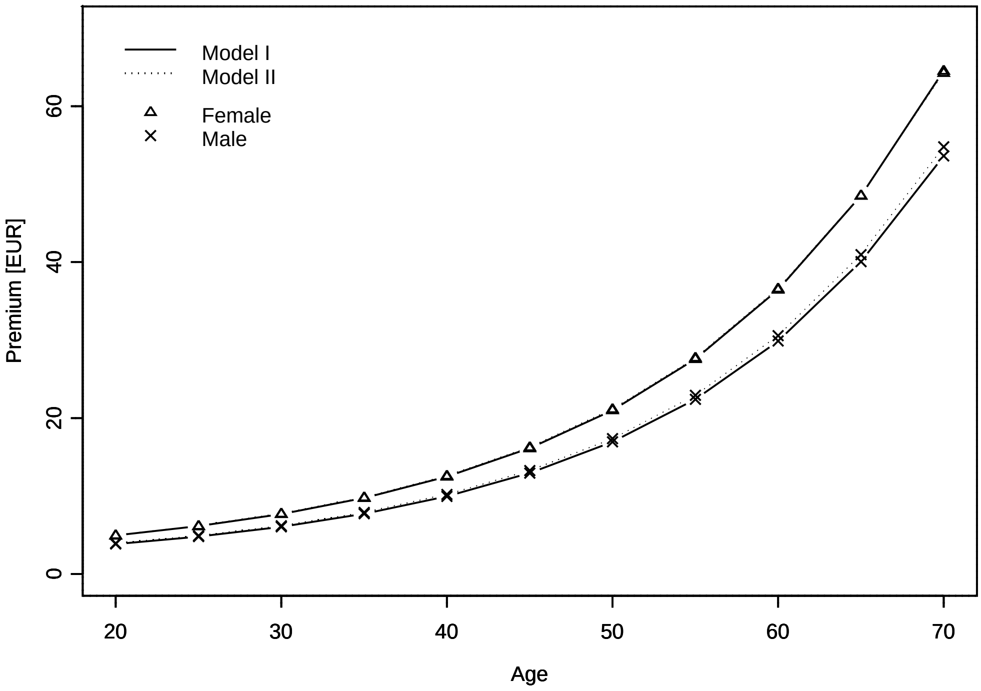

Figure 8.2: Comparison of Premiums using different Correlation Structures

Chapter 9

Summary

We have presented three non-parametric estimators for basic quantities in survival analysis, namely the Nelson-Aalen estimator for the cumulative hazard-rate function, the Kaplan-Meier estimator for the survival distribution function and finally the Aalen-Johansen estimator for the transition matrix of a Markovian multi-state model. All three estimators can be shown to be at least almost unbiased. This property enabled us to define pseudo-values and construct a relationship between the transition probabilities and the covariates of a single observation and thus generate the data required for a regression analysis.

Consequently we could derive the transition probabilities in a Markovian three-state model, with states "Care at home", "Care in a nursing home" and "Death", using GEEs, that in contrast to GLMs take correlation between observation into account. To extend this model to an insurance model, the state "Active" had to be added. Since our data did only provide information on LTC claimants, additional sources had to be used, such as incidence rates for LTC and mortality rates for active individuals. The transition probabilities obtained for this model could be used to calculate the actuarial values and derive premiums for a given LTC-plan.

In the following table we are going to compare the premiums obtained with premiums offered by a German health insurer. As already done in Section 8.6, we calculated premiums for a 10 EUR daily allowance based on the LTC-plan "PET":

Age	Premium based on GEEs		Premium offered by German health insurer	
	Female	Male	Female	Male
20	04.94	03.85	02.12	01.70
25	06.14	04.80	02.92	02.33
30	07.70	06.05	03.90	03.10
35	09.76	07.72	05.05	04.01
40	12.51	09.97	06.44	05.13
45	16.14	12.95	08.16	06.52
50	21.01	16.95	10.39	08.36
55	27.58	22.40	13.32	10.86
60	36.48	29.88	17.31	14.40
65	48.53	40.06	22.01	18.84
70	64.55	53.62	29.04	25.71

We observe higher premiums using GEEs although the data have only slightly been modified. But we can see that the behavior with respect to age is similar as well as the proportion between males and females. It also should be noted that the incidence rates and mortality rates for "Active" individuals include administrative costs, whereas the transition probabilities do not. Therefore a comparison between the calculated premiums using GEEs and the commercial premiums is not very reasonable. Further the LTC definition in different countries, such as Japan and Germany, varies and therefore country-specific incidence rates might be necessary.

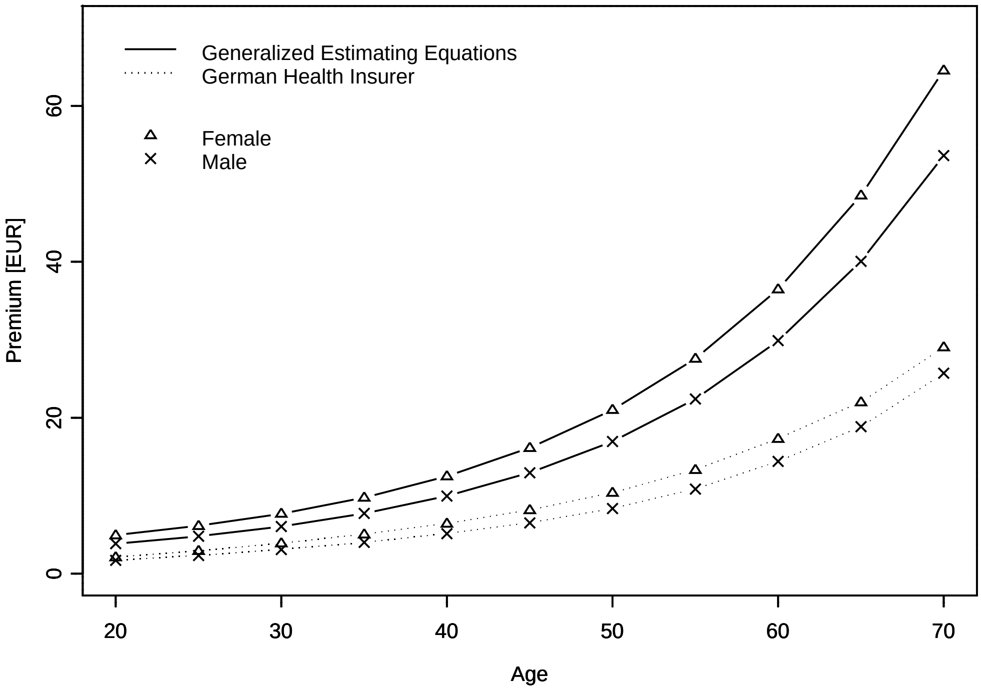

Figure 9.1: Comparison of Premiums

The method we used models the transition probabilities directly. The transition probabilities can be calculated simply using the mean function, i.e. the inverse of the link function, of our model for any given set of covariates, and therefore are simple functions of the regression coefficients.

However, in our case there are no methods available at the moment to examine the goodness-of-fit or confirm the choice of link function. The choice of time-points might also influence the results, but in our case the time-points are given, as we need one-year transition probabilities for the calculation of actuarial values. More precise estimates might be obtained if the correlation matrix is chosen close to the true one.

Appendix A

Parameter Estimation

A.1 The Exponential Family

In the first of the specifications for a GLM (6.1.4) we required that the components y_i have a distribution from the exponential family, which contains the Normal-, Poisson-, Binomial-, Gamma- and Inverse Gaussian distribution. A distribution is called to belong to the exponential family, if we can specify functions $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$ and the canonical parameter θ, such that its density function can be written in the following form:

$$f_Y(y;\theta,\phi) = \exp\left\{\left(\frac{y\theta - b(\theta)}{a(\phi)}\right) + c(y,\phi)\right\} \tag{A.1}$$

With this expression the log-likelihood function $l(\theta,\phi;y) := \ln f_Y(y;\theta,\phi)$, as a function of θ and ϕ with fixed y, can be easily derived. Further expressions for the mean and variance of Y can be derived using the following equations:

$$E\left[\frac{\partial l(\theta,\phi;y)}{\partial\theta}\right] = 0$$

$$E\left[\frac{\partial^2 l(\theta,\phi;y)}{\partial\theta^2}\right] + E\left[\left(\frac{\partial l(\theta,\phi;y)}{\partial\theta}\right)^2\right] = 0$$

Proof:

$$\begin{aligned}
E\left[\frac{\partial l(\theta;y)}{\partial\theta}\right] &= \int \frac{\partial \ln f(y;\theta)}{\partial\theta} f(y;\theta)dy \\
&= \int \frac{1}{f(y;\theta)}\left(\frac{\partial}{\partial\theta} f(y;\theta)\right) f(y;\theta)dy \\
&= \int \frac{\partial}{\partial\theta} f(y;\theta)dy \\
&= \frac{\partial}{\partial\theta} \underbrace{\int f(y;\theta)dy}_{1} = 0
\end{aligned}$$

To simplify notation we omit now the arguments of the log-likelihood and density functions and write instead $l := l(\theta; y)$ and $f := f(y; \theta)$. Further for the derivatives with respect to θ we write $f' := \partial f/\partial\theta$ and $f'' := \partial^2 f/\partial\theta^2$, respectively

$$\begin{aligned}
E\left[\frac{\partial^2 l}{\partial\theta^2}\right] + E\left[\left(\frac{\partial l)}{\partial\theta}\right)^2\right] &= \int \frac{\partial}{\partial\theta}\left(\frac{\partial \ln f}{\partial\theta}\right) f dy + \int \left(\frac{\partial \ln f}{\partial\theta}\right)^2 f dy \\
&= \int \frac{\partial}{\partial\theta}\left(\frac{1}{f}\left(\frac{\partial}{\partial\theta} f\right)\right) f dy + \int \left(\frac{1}{f}\left(\frac{\partial}{\partial\theta} f\right)\right)^2 f dy \\
&= \int \frac{f''f - f'f'}{f^2} f dy + \int \frac{f'^2}{f} dy \\
&= \int f'' dy - \int \frac{f'^2}{f} dy + \int \frac{f'^2}{f} dy \\
&= \int \frac{\partial^2}{\partial\theta^2} f dy \\
&= \frac{\partial^2}{\partial\theta^2} \underbrace{\int f dy}_{1} = 0
\end{aligned}$$

□

Using the representation of the density function (A.1) we get for the log-likelihood function

$$l(\theta, \phi; y) = \left\{ \left(\frac{y\theta - b(\theta)}{a(\phi)}\right) + c(y, \phi) \right\}$$

Therefore we obtain

$$E\left[\frac{\partial l(\theta, \phi; y)}{\partial\theta}\right] = E\left[\frac{y - b'(\theta)}{a(\phi)}\right] = 0$$

$$\Rightarrow E[Y] = b'(\theta)$$

$$E\left[\frac{\partial^2 l(\theta, \phi; y)}{\partial\theta^2}\right] + E\left[\left(\frac{\partial l(\theta, \phi; y)}{\partial\theta}\right)^2\right] = E\left[\frac{-b''(\theta)}{a(\phi)}\right] + E\left[\left(\frac{y - b'(\theta)}{a(\phi)}\right)^2\right] = 0$$

$$\Rightarrow Var[Y] = b''(\theta)a(\phi) \tag{A.2}$$

As we can see in (A.2), the variance of Y is the product of two functions. The first factor is the function $b''(\theta)$, which is referred to as the variance function $V(Y)$. This function depends through the canonical parameter θ on the mean. Thus it will be considered as a function of μ and we will write $V(\mu)$. The second factor is the function $a(\phi)$, which is independent of θ and depends on ϕ only. Each distribution function of the exponential family has a special link functions, such that $\theta = \eta$. This link function is called the canonical link function.

A.2 Ordinary Least-Squares Regression

Aim in ordinary least-squares regressions is to explain the linear relationship between the vectors of covariates $\mathbf{x_{1j}}, \ldots, \mathbf{x_{nj}}$ and the observed random variables $\mathbf{y} = (y_1, \ldots, y_n)^T$. The standard model for multivariate linear Regression is the following:

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i \qquad i = 1, \ldots, n \tag{A.3}$$

where $y_1, \ldots, y_n$ are the independent and normally distributed random variables and $\mathbf{x_{1j}}, \ldots, \mathbf{x_{nj}}$ the $1 \times (p+1)$ vectors of covariates. The vector of unknown parameters is $\boldsymbol{\beta} = (\beta_0, \ldots, \beta p)^T$ and has to be estimated from the data. Finally the quantities $\varepsilon_1, \ldots, \varepsilon_n$ are the error terms, which are assumed to be independent and identically normally distributed with mean zero and variance σ^2. The covariates are collected in the so-called design matrix, that is

$$X = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1p} \\ \ldots & \ldots & \ldots & \ldots \\ 1 & x_{n1} & \ldots & x_{np} \end{pmatrix}$$

The design matrix can be used to write (A.3) as a matrix equation given by

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

The most important assumption is that the error terms are independent and identically normally distributed:

$$\boldsymbol{\varepsilon} \overset{iid}{\sim} N(\mathbf{0}, \sigma^2 \mathbf{I_n})$$

From this assumption it follows that the vector $\mathbf{y}$ has expectation $X\boldsymbol{\beta}$ and variance $\sigma^2\mathbf{I_n}$ using the independence of observations.

$$\begin{aligned} E[\mathbf{y}] &= E[X\boldsymbol{\beta}] + E[\boldsymbol{\varepsilon}] = X\boldsymbol{\beta} \\ Var[\mathbf{y}] &= Var[\boldsymbol{\varepsilon}] = \sigma^2\mathbf{I_n} \end{aligned}$$

The estimator $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ is obtained by least-squares, that is we minimizes $\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon}$ over $\boldsymbol{\beta}$.

$$(\mathbf{y} - X\boldsymbol{\beta})^T (\mathbf{y} - X\boldsymbol{\beta}) \tag{A.4}$$

To obtain an unique solution of this minimizing problem we require $n \geq p+1$ and the vectors of covariates $\mathbf{x_{1j}}, \ldots, \mathbf{x_{nj}}$ to be linearly independent. If these two requirements are fulfilled, we get, taking first derivative of (A.4) with respect to $\boldsymbol{\beta}$ and setting the result to zero, the estimator $\hat{\boldsymbol{\beta}}$.

$$X^T(\mathbf{y} - X\boldsymbol{\beta}) = 0 \qquad \Longleftrightarrow \qquad X^TX\boldsymbol{\beta} = X^T\mathbf{y}$$

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

Using the property that $E[AY] = AE[Y]$ and $Var[AY] = AVar[Y]A^T$ we obtain:

$$\begin{aligned} E[\hat{\boldsymbol{\beta}}] &= E[\left(X^TX\right)^{-1} X^T\mathbf{y}] = \left(X^TX\right)^{-1} X^T E[\mathbf{y}] = \left(X^TX\right)^{-1} X^T\boldsymbol{\beta} = \boldsymbol{\beta} \\ Var[\hat{\boldsymbol{\beta}}] &= Var[\left(X^TX\right)^{-1} X^T\mathbf{y}] = \left(X^TX\right)^{-1} X^T Var[\mathbf{y}] X \left(X^TX\right)^{-T} \\ &= \left(X^TX\right)^{-1} X^T \sigma^2 \mathbf{I_n} X \left(X^TX\right)^{-T} = \sigma^2 \left(X^TX\right)^{-1} \end{aligned}$$

This estimator is known as the best linear unbiased estimator (BLUE) for the above model. For further details on ordinary least-squares regression see the book by Fahrmeir, Künstler, Piegeot, and Tutz (2003).

A.3 Weighted Least-Squares Regression

For ordinary least-squares we assumed that the error term $\boldsymbol{\varepsilon}$ is normally distributed with expectation zero and covariance matrix $\sigma^2 I_n$. For a weighted least-squares we assume that the covariance matrix of the error term is equal to $\sigma^2 V$, where V is a diagonal matrix with elements $v_i > 0$. This leads us to

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad \boldsymbol{\varepsilon} \overset{iid}{\sim} N(\mathbf{0}; \sigma^2 V) \tag{A.5}$$

We try to track this problem back to ordinary least-squares and multiply above equation by the matrix $V^{-\frac{1}{2}}$:

$$V^{-\frac{1}{2}}\mathbf{y} = V^{-\frac{1}{2}}X\boldsymbol{\beta} + V^{-\frac{1}{2}}\boldsymbol{\varepsilon} \qquad \boldsymbol{\varepsilon} \overset{iid}{\sim} N(\mathbf{0}; \sigma^2 V)$$

Defining $\mathbf{y}^* := V^{-\frac{1}{2}}\mathbf{y}$, $X^* := V^{-\frac{1}{2}}X$ and $\boldsymbol{\varepsilon}^* := V^{-\frac{1}{2}}\ \boldsymbol{\varepsilon}$ we get the following setup, similar to the ordinary least-squares regression model:

$$\begin{aligned} E[\boldsymbol{\varepsilon}^*] &= E[V^{-\frac{1}{2}}\boldsymbol{\varepsilon}] = V^{-\frac{1}{2}}E[\boldsymbol{\varepsilon}] = 0 \\ Var[\boldsymbol{\varepsilon}^*] &= Var[V^{-\frac{1}{2}}\boldsymbol{\varepsilon}] = V^{-\frac{1}{2}}Var[\boldsymbol{\varepsilon}]\left(V^{-\frac{1}{2}}\right)^T = V^{-\frac{1}{2}}\sigma^2 I_n\left(V^{-\frac{1}{2}}\right)^T = \sigma_2 I_n \end{aligned}$$

Therefore

$$\mathbf{y}^* = X^*\boldsymbol{\beta} + \boldsymbol{\varepsilon}^* \qquad \boldsymbol{\varepsilon}^* \overset{iid}{\sim} N(\mathbf{0}; \sigma^2 I_n)$$

From ordinary least-squares we know that the estimator for $\boldsymbol{\beta}$ is defined as

$$\begin{aligned} \hat{\boldsymbol{\beta}}^* &= (X^{*T}X^*)^{-1}X^*\mathbf{y}^* \\ &= (X^T(V^{-\frac{1}{2}})^T V^{-\frac{1}{2}}X)^{-1}X^T(V^{-\frac{1}{2}})^T V^{-\frac{1}{2}}\mathbf{y} \\ &= (X^T V^{-1}X)^{-1}X^T V^{-1}\mathbf{y} \end{aligned}$$

Further we can calculate the expectation and covariance matrix of the estimator $\hat{\boldsymbol{\beta}}^*$:

$$\begin{aligned} E[\hat{\boldsymbol{\beta}}^*] &= (X^T V^{-1}X)^{-1}X^T V^{-1}E[\mathbf{y}] = (X^T V^{-1}X)^{-1}X^T V^{-1}X\boldsymbol{\beta} = \boldsymbol{\beta} \\ Var[\hat{\boldsymbol{\beta}}^*] &= (X^T V^{-1}X)^{-1}X^T V^{-1}Var[\mathbf{y}]\left((X^T V^{-1}X)^{-1}X^T V^{-1}\right)^T \\ &= (X^T V^{-1}X)^{-1}X^T V^{-1}\sigma^2 V V^{-T}X(X^T V^{-1}X)^{-T} \\ &= \sigma^2(X^T V^{-1}X)^{-1}X^T V^{-1}X(X^T V^{-1}X)^{-T} \\ &= \sigma^2(X^T V^{-1}X)^{-1} \end{aligned}$$

Thus the estimator $\hat{\boldsymbol{\beta}}^*$ has to fulfill the following equations:

$$(X^T V^{-1}X)\hat{\boldsymbol{\beta}}^* = X^T V^{-1}\mathbf{y}$$

$\hat{\boldsymbol{\beta}}^*$ is called the weighted least-squares estimator of $\boldsymbol{\beta}$ in Model A.5.

A.4 The O_p and o_p Notation for Stochastic Sequences

The idea of the O, o notation for non-stochastic sequences as known from Taylor expansion can be extended to stochastic sequences, as well. The information contained in the $O_p(1)$, $o_p(1)$ notation is often referred to as the stochastic order of X_n. The following section refers to Bishop, Fienberg, and Holland (1980) and further information can be found there. They define $o_p(1)$ for a sequence of random variables as follows:

Definition A.1 *The stochastic sequence $X_n = o_p(1)$, if for every $\varepsilon > 0$*

$$\lim_{n\to\infty} P(|X_n| \leq \varepsilon) = 1$$

- *If X_n is a vector, we write $X_n = o_p(1)$, if $||X_n|| = o_p(1)$;*
- *If $X_n/b_n = o_p(1)$, we write $X_n = o_p(b_n)$;*
- *If X_n is a vector, we write $X_n = o_p(b_n)$, if $||X_n|| = o_p(b_n)$;*

To show the differences between the $o_p(1)$ and $O_p(1)$ definition, we rewrite the definition for $o_p(1)$. It is equivalent for $X_n = o_p(1)$ to say, that the absolute value of X_n is $o_p(1)$ with an arbitrarily high probability.

Definition A.2 *If for every $\varepsilon, \eta > 0$ there exists an integer $n(\varepsilon, \eta)$ such that for $n \geq n(\varepsilon, \eta)$*

$$P\left(|X_n| \leq \varepsilon\right) \geq 1 - \eta$$

then the stochastic sequence $X_n = o_p(1)$.

From this definition of $o_p(1)$ we define $O_p(1)$ in an analogue way. We suppose that $X_n = O_p(1)$ means with arbitrarily high probability $|X_n| = O_p(1)$.

Definition A.3 *If for every $\eta > 0$ there exists a constant $K(\eta)$ and an integer $n(\eta)$ such that for $n \geq n(\eta)$*

$$P\left(|X_n| \leq K(\eta)\right) \geq 1 - \eta$$

then the stochastic sequence $X_n = O_p(1)$.

As in the case for $o_p(1)$ similar definitions apply for the case that X_n is a vector. To summarize both definitions, one can refer to $o_p(1)$ as X_n converging to zero in probability, and $= O_p(1)$ as X_n being bounded in probability.

There is a relationship between Tchebychev's inequality and the $O_p(1)$, $o_p(1)$ notation. This allows us to connect the stochastic order of magnitude from the $O_p(1)$, $o_p(1)$ notation with the standard deviation from Tchebychev's inequality. The Tchebychev's inequality is stated in the following theorem:

Theorem A.1 (Tchebychev's inequality) *If X_n is a random variable with mean μ and variance $\sigma^2 < \infty$, then for any positive number h*

$$P(|X_n - \mu| \leq h\sigma) \geq 1 - \frac{1}{h^2}$$

Proof of Theorem A.1:

To prove Tchebychev's inequality we use Markov's inequality, that holds for non-negative random variables $Y_n \geq 0$:

$$aI_{\{Y_n \geq a\}} \leq aI_{\{Y_n \geq a\}} + \underbrace{a}_{\geq 0} \underbrace{I_{\{Y_n \leq a\}}}_{\geq 0} = E\left[Y_n\right]$$

Taking expectations on both sides of Markov's inequality we obtain:

$$aI_{\{Y_n \geq a\}} \leq E\left[Y_n\right] \quad \Rightarrow \quad E\left[aI_{\{Y_n \geq a\}}\right] = aP(Y_n \geq a) \leq E\left[Y_n\right] \quad \Rightarrow \quad P(Y_n \geq a) \leq \frac{E[Y_n]}{a}$$

Setting $Y_n := (X_n - \mu)^2$ and $a = \varepsilon^2$ this leads to

$$\begin{aligned} P\left((X_n - \mu)^2 \geq \varepsilon^2\right) &\leq \frac{E\left[(X_n - \mu)^2\right]}{\varepsilon^2} = \frac{Var[X_n]}{\varepsilon^2} \\ P\left(|X_n - \mu| \leq \varepsilon\right) \geq 1 - \frac{\sigma^2}{\varepsilon^2} &\overset{\varepsilon := h\sigma}{\Longrightarrow} P\left(|X_n - \mu| \leq h\sigma\right) \geq 1 - \frac{\sigma^2}{h^2\sigma^2} = 1 - \frac{1}{h^2} \end{aligned}$$

which gives us exactly the inequality we called Tchebychev's inequality above.

□

The following theorem gives us above mentioned relationship between Tchebychev's inequality and the $O_p(1)$, $o_p(1)$ notation:

Theorem A.2 *If X_n is a stochastic sequence with $\mu_n = E[X_n]$ and $\sigma_n^2 = Var[X_n] < \infty$, then*

$$X_n - \mu_n = O_p(\sigma_n)$$

Proof:

We define $h := \eta^{-\frac{1}{2}}$ for $0 < \eta < 1$. Applying Tchebychev's inequality to X_n, μ_n and σ_n we have:

$$P\left(\frac{|X_n - \mu_n|}{\sigma_n} \leq \eta^{-\frac{1}{2}}\right) \geq 1 - \eta$$

This holds for $n = 1, 2, \ldots$. Now we set $K(\eta) = \eta^{-\frac{1}{2}}$ and apply the definition of $= O_p(1)$ to conclude that

$$P\left(\frac{|X_n - \mu_n|}{\sigma_n} \leq K(\eta)\right) \geq 1 - \eta$$

This corresponds with the definition of $O_p(1)$; therefore

$$\frac{X_n - \mu_n}{\sigma_n} = O_p(1)$$

□

The next theorem gives us a tool to prove that a sequence is $o_p(1)$, if we have already shown that the sequence is $O_p(n^{-\frac{1}{2}})$.

Theorem A.3

$$X_n = O_p(n^{-\frac{1}{2}}) \qquad \Longrightarrow \qquad X_n = o_p(1)$$

Proof:

$$X_n = O_p(n^{-\frac{1}{2}}) \qquad \Longleftrightarrow \qquad n^{\frac{1}{2}} X_n = O_p(1)$$

This is equivalent, that for every $\eta > 0$ there exists a constant $K(\eta)$ and an integer $n(\eta)$ such that for $n \geq n(\eta)$

$$P(n^{\frac{1}{2}}|X_n| \leq K(\eta)) \geq 1 - \eta$$

This can also be written as

$$P(|X_n| \leq n^{-\frac{1}{2}} K(\eta)) \geq 1 - \eta$$

It follows for every $\varepsilon > 0$ that

$$\lim_{n \to \infty} P(|X_n| < \varepsilon) = 1$$

which is nothing else than the definition for $o_p(1)$.

In the following we want to use these notations in the context of Taylor expansion. A first order Taylor expansion for a function f is, as $x \to a$, given by

$$f(x) = f(a) + f'(a)(x - a) + o(|x - a|) \tag{A.6}$$

where we require the function f to be continuously differentiable in a. Assume that we are given a random variable X_n such that $X_n - a = O_p(n^{-\frac{1}{2}})$, then

$$f(X_n) = f(a) + f'(a)(X_n - a) + o_p(n^{-\frac{1}{2}})$$

Proof:

We define a new function $h(x)$ such that

$$h(x) := \begin{cases} \frac{f(x) - f(a) - f'(a)(x-a)}{|x-a|} & x \neq a \\ 0 & x = a \end{cases}$$

Since f is continuous differentiable in a, so is h and $h(X_n) \xrightarrow{P} h(a) = 0$, this is nothing else than $h(X_n) = o_p(1)$ and it follows that $h(X_n)|X_n - a| = o_p(1) \cdot O_p(n^{-\frac{1}{2}}) = o_p(n^{-\frac{1}{2}})$ □

A.5 Cramér-Rao Inequality

The Cramér-Rao inequality, states (see Gart (1959)) that under regularity conditions for a consistent estimator $\hat{\boldsymbol{\theta}}$ the following inequality holds:

$$Cov\left[\hat{\boldsymbol{\theta}}\right] \geq I(\boldsymbol{\theta})^{-1}$$

where $I(\boldsymbol{\theta})$ is the information matrix, which is as defined in (6.2), minus the expectation of the second derivative of the log-likelihood function with respect to $\boldsymbol{\theta}$.

Proof:

If $L(\boldsymbol{\theta};\mathbf{y}) := f(\mathbf{y};\boldsymbol{\theta})$ denotes the likelihood function, it follows by this definition that

$$\int \!\! \ldots \!\! \int L\, dy_1 \ldots dy_n = 1$$

Taking the derivative with respect to $\boldsymbol{\theta}$ we obtain interchanging differentiation and integration:

$$\int \!\! \ldots \!\! \int \frac{\partial L}{\partial \boldsymbol{\theta}}\, dy_1 \ldots dy_n = 0 \qquad \Rightarrow \qquad \int \!\! \ldots \!\! \int \frac{\partial \ln L}{\partial \boldsymbol{\theta}} L\, dy_1 \ldots dy_n = 0$$

Taking a second time the derivative we get:

$$\int \!\! \ldots \!\! \int \left(\frac{\partial^2 \ln L}{\partial \boldsymbol{\theta}^2} + \left(\frac{\partial lnL}{\partial \boldsymbol{\theta}} \right)^2 \right) L\, dy_1 \ldots dy_n = 0$$

This is equal to

$$E\left[\frac{\partial^2 \ln L}{\partial \boldsymbol{\theta}^2}\right] + E\left[\left(\frac{\partial lnL}{\partial \boldsymbol{\theta}}\right)^2\right] = 0 \qquad \Rightarrow \qquad E\left[\left(\frac{\partial lnL}{\partial \boldsymbol{\theta}}\right)^2\right] = -E\left[\frac{\partial^2 \ln L}{\partial \boldsymbol{\theta}^2}\right] = I(\boldsymbol{\theta}) \quad \text{(A.7)}$$

Since $\hat{\boldsymbol{\theta}}$ is an unbiased estimator we have:

$$E[\hat{\boldsymbol{\theta}}] = \int \!\! \ldots \!\! \int \hat{\boldsymbol{\theta}} L\, dy_1 \ldots dy_n = \boldsymbol{\theta}$$

If we take the derivative with respect to $\boldsymbol{\theta}$ on the last equation it follows that

$$\int \!\! \ldots \!\! \int \hat{\boldsymbol{\theta}} \frac{\partial \ln L}{\boldsymbol{\theta}} L\, dy_1 \ldots dy_n = 1 \qquad \text{(A.8)}$$

As seen in (6.8) the expectation of the log-likelihood function with respect to $\boldsymbol{\theta}$ is equal to zero. Therefore we can write:

$$E\left[\frac{\partial \ln L}{\partial \boldsymbol{\theta}}\right] = 0 \qquad \Rightarrow \qquad E\left[\boldsymbol{\theta}\frac{\partial \ln L}{\partial \boldsymbol{\theta}}\right] = \boldsymbol{\theta} E\left[\frac{\partial \ln L}{\partial \boldsymbol{\theta}}\right] = 0 \qquad \text{(A.9)}$$

Using the Cauchy-Schwarz inequality, that is $E[XY]^2 \leq E[X^2]E[Y^2]$, we finally obtain the required result:

$$1 \overset{(A.8)+(A.9)}{=} E\left[\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)\frac{\partial \ln L}{\partial \boldsymbol{\theta}}\right]^2 \leq \underbrace{E\left[\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)^2\right]}_{Cov\left[\hat{\boldsymbol{\theta}}\right]} E\left[\left(\frac{\partial \ln L}{\partial \boldsymbol{\theta}}\right)^2\right] \tag{A.10}$$

Now it is easy to see, using (A.10) and (A.7), that this is nothing else than the Cramér-Rao inequality:

$$Cov[\hat{\boldsymbol{\theta}}] \geq \left(E\left[\left(\frac{\partial \ln L}{\partial \boldsymbol{\theta}}\right)^2\right]\right)^{-1} \overset{(A.7)}{=} \left(-E\left[\frac{\partial^2 \ln L}{\partial \boldsymbol{\theta}^2}\right]\right)^{-1} = I(\boldsymbol{\theta})^{-1}$$

□

Appendix B

Program and Tables

B.1 Program for Calculating the Aalen-Johansen Estimator

The parameter estimation using GEEs was performed using Mark X. Norleans' program "The Generalized Estimating Equations Procedure, Version 2.6" written for "Splus" with a syntax similar to the existing Splus-glm() program. For details see http://lib.stat.cmu.edu/.

To calculate the Aalen-Johansen Estimator as well as the pseudo-values we used an own program, which will be explained in this Section. Our dataset had one row for each transition containing information on each individual's sex, age at the time, the transition to the current state occurred, begin and end of the time spent in this state, place of care, level of care, the state, to which the next transition was observed as well as additional information, which were not of interest for the problem at hand.

In a first step the dataset had to be restricted to rows one to thirteen (only to get rid of unnecessary information). Then we used the function "prep.aj.three.state" or "prep.aj.four.state" to prepare for the next step. Note that we designed one program for a three-state model and a four-state model. Therefore these two preparations functions were necessary as kind of a translator to determine the state from which and to which the corresponding transition took place. Thus we refer to "place of care" or "level of care" as "state" depending on the model used.

After having transferred the so prepared dataset to "Matlab" we used the function "riskset" to calculate at each transition time the corresponding so-called "riskset" at that time, i.e. all individuals that were at this time in the same state. As we need for calculating the actuarial values one-year transition probabilities, we had to account for that with the function "transition". Thus we summarized the l to $l+1$-year transitions in a new dataset "transition.l", where l ranges in $l = 0, \dots, k$. Based on the dataset "transition.l" and the "riskset" we could calculate the Aalen-Johansen estimator. The function "aj" returned already the "leave-one-out" Aalen-Johansen Estimator with the estimator based on the entire sample being in the first row.

Next, based on this matrix the pseudo-values were calculated using the function "pseudovalues" and prepared for the parameter estimation with the function "gee". An additional column identifying the i^{th} observation and a column indicating the "time" of the transition, meaning l to $l+1$-year transition, had to be included. The dataset, returned by the function "gee", had to be transfered back to "Splus" and was after some adjustments then ready for further calculation with above mentioned program for GEEs written by Mark X. Norleans.

The Splus Function ”prep.aj.three.state”

```
prep.aj.three.state_function(data)
{
   working.data_data
   attach(working.data)
   data.length_nrow(working.data)
   j_length(working.data)

   G_rep(0,data.length)
   H_rep(0,data.length)
   Time_rep(0,data.length)
   Sex_rep(0, data.length)
   ZArt_rep(0, data.length)
   ZLevel_rep(0, data.length)

   for (i in 1:data.length){

      G[i]_data$Art[i]

      if (data$Transitto[i]==1){ H[i]_3 }
      if ((data$Transitto[i]==11) || (data$Transitto[i]==12) || (data$Transitto[i]==13)){H[i]_1}
      if ((data$Transitto[i]==21) || (data$Transitto[i]==22) || (data$Transitto[i]==23)){H[i]_2}


      Time[i]_data$LTCend[i]

      if (data$Zsex[i]==”m”){ Sex[i]_1 }

      if (data$Art[i]==”stationaer”){ ZArt[i]_1 }

      if (data$Level[i]==”Stufe1”){ ZLevel[i]_0 }
      if (data$Level[i]==”Stufe2”){ ZLevel[i]_1 }
      if (data$Level[i]==”Stufe3”){ ZLevel[i]_2 }

      if ( (G[i]==2) && (H[i]==1) ){ H[i]_0 }
   }

   add_cbind(G, H, Time, Sex, ZArt, ZLevel)
   working.data_cbind(working.data, add)

   datdata_working.data[order(working.data[,5], working.data[,j+1], working.data[,j+2]), na.last=T]


return(datdata)
}
```

The Matlab Function ”riskset”

```
function A = riskset(A);
n=size(A,1);
m=size(A,2);

for j=1:n
    A(j,m+1)=0;
    A(j,m+2)=0;
end

for i = 1:n
    r=0;

    for j =1:n

        if (A(i,13) == A(j,13)) & (A(i,11)==A(j,11)) & (A(i,12) == A(j,12)) & (A(i,12) ~= 0)
            r=r+1;
            A(i,m+1)=r;
        end

        if (A(j,11)==A(i,11)) & (A(j,3) < A(i,4)) & (A(i,4) <= A(j,4))
            A(i,m+2)=A(i,m+2)+1;
        end

    end

end

A;
```

The Matlab Function ”transition”

```
function C=transition(A, d)
n=size(A,1);
m=size(A,2);
j=1;
B=A;

for i=1:n

    if d*360 ¡= A(i,4) & A(i,4) ¡ (d+1)*360
        B(j,:)=A(i,:);
        j=j+1;
    end

end

C=B(1:j-1,1:m);
```

Calculation of the "leave-one-out" Aalen-Johansen Estimators in Matlab

```
load three_state_model

threeaalen0=aj(threetransition0, riskset);

save threeaalen0
. . .

threeaalenK=aj(threetransitonK, riskset);

save threeaalenK
```

The Matlab Function "aj"

```
function AJ = aj(A, B)
n=size(A,1); m=size(B,1); AJ=zeros(m+1,16);

for i=0:m
  delete=i; p=n-1; DD=A(:,17); RR=A(:,18);

  for t=1:p
    if ( A(t,11)==A(t+1,11) ) & ( A(t,12)==A(t+1,12) ) & ( A(t,13)==A(t+1,13) )
      DD(t+1)=0;
    end
  end

  if ( delete ~= 0 )
    j=delete;
    if (j <= n)
      while ( (DD(j) == 0) & (j > 1) )
        j=j-1;
      end
      DD(j)=DD(j)-1;
    end
  end

  if (delete ~= 0)
    for j=1:n
      if ( A(j, 11) ~= B(i,11) )
        RR(j)=RR(j);
      end
      if ( A(j, 11) == B(i,11) ) & ( A(j, 4) > B(i,3) ) & ( A(j, 4) <= B(i,4) )
        RR(j)=RR(j)-1;
      end
    end
  end
```

```
   mm=0;

   for j=1:n
      if ( DD(j) > 0 ) & ( A(j, 12) ~= A(j,11) )
         mm=mm+1;
      end
   end

   l=1; d=zeros(m,1);
   r=zeros(m,1);
   g=zeros(m,1);
   h=zeros(m,1);
   t=zeros(m,1);

   for j=1:n
      if ( DD(j) > 0 ) & ( A(j, 12) ~= A(j,11) )
         d(l)=DD(j);
         r(l)=RR(j);
         g(l)=A(j,11);
         h(l)=A(j,12);
         t(l)=A(j,13);
         l=l+1;
      end
   end

   workingAJ=aalenjohn(d, r, g, h, t);
   AJ(i+1,1)=workingAJ(1,1);
   AJ(i+1,2)=workingAJ(1,2);
   AJ(i+1,3)=workingAJ(1,3);
   AJ(i+1,4)=workingAJ(1,4);
   AJ(i+1,5)=workingAJ(2,1);
   AJ(i+1,6)=workingAJ(2,2);
   AJ(i+1,7)=workingAJ(2,3);
   AJ(i+1,8)=workingAJ(2,4);
   AJ(i+1,9)=workingAJ(3,1);
   AJ(i+1,10)=workingAJ(3,2);
   AJ(i+1,11)=workingAJ(3,3);
   AJ(i+1,12)=workingAJ(3,4);
   AJ(i+1,13)=workingAJ(4,1);
   AJ(i+1,14)=workingAJ(4,2);
   AJ(i+1,15)=workingAJ(4,3);
   AJ(i+1,16)=workingAJ(4,4);
end

AJ;
```

The Matlab Function ”aalenjohn”

```
function mat = aalenjohn(d, r, g, h, t)
n=length(d); AJ=zeros(n, 16);

for i=1:n
   AJ(i,1)=1; AJ(i,6)=1; AJ(i,11)=1; AJ(i,16)=1;
end

if (n >= 1)

   for j=1:n
      for i=1:n
         if ( h(i) ~ =0 ) & ( t(i)==t(j) ) & ( r(i) > 1)
            if ( g(i)==1 )
               AJ(j,1)=AJ(j,1)-d(i)/r(i);
               AJ(j,h(i))=AJ(j,h(i))+d(i)/r(i);
            end
            if ( g(i)==2 )
               AJ(j,6)=AJ(j,6)-d(i)/r(i);
               AJ(j,h(i)+4)=AJ(j,h(i)+4)+d(i)/r(i);
            end
            if ( g(i)==3 )
               AJ(j,11)=AJ(j,11)-d(i)/r(i);
               AJ(j,h(i)+8)=AJ(j,h(i)+8)+d(i)/r(i);
            end
         end
      end
   end

end

l=1; m=n-1;
if (m < 1)
   mat=eye(4);
end

if (m >= 1)

   for k=1:m
      if (t(k) ~= t(k+1))
         l=l+1;
      end
   end

   k=1; aj=zeros(l, 16);

   for i=1:n
      aj(i,1)=1; aj(i,6)=1; aj(i,11)=1; aj(i,16)=1;
   end
   for j=1:n
      s=1;

      for s=1:16
         aj(k,s)=AJ(j,s);
      end

      if (j < n)
         if (t(j) =t(j+1))
            k=k+1;
         end
      end

   end

   mat=eye(4); test=eye(4);

   for j=1:l
      test(1,1)=aj(j,1); test(1,2)=aj(j,2); test(1,3)=aj(j,3); test(1,4)=aj(j,4);
      test(2,1)=aj(j,5); test(2,2)=aj(j,6); test(2,3)=aj(j,7); test(2,4)=aj(j,8);
      test(3,1)=aj(j,9); test(3,2)=aj(j,10); test(3,3)=aj(j,11); test(3,4)=aj(j,12);
      test(4,1)=aj(j,13); test(4,2)=aj(j,14); test(4,3)=aj(j,15); test(4,4)=aj(j,16);
      mat=mat*test;
   end

end
```

Calculation of the dataset for the GEEs in Matlab

```
load threeaalen0

...

load threeaalenK

threepseudo0=pseudovalues(threeaalen0);

...

threepseudoK=pseudovalues(threeaalenK);

geedata=gee(threepseudo0, ...threepseudoK, threetest);

save checkgee.m geedata
```

The Matlab Function ”pseudovalues”

```
function B = pseudovalues(A)
n=size(A,1)-1;
B=zeros(n,16);

for i=1:n
    B(i,:)=n*A(1,:) - (n-1)*A(i+1,:);
end
```

The Matlab Function ”gee”

```
function A = gee(A1, . . . AK, B)
n=size(B,1);
m=size(B,2);

B1=B;
for i=1:18
    B1(1,m+i)=0;
end
B1(:,m+1)=ones(n,1);
for i=1:16
    B1(:,m+1+i)=A1(:,i);
end
B1(:,m+18)=(1:1:n)’;

...

BK=B;
for i=1:18
    BK(1,m+i)=0;
end
BK(:,m+1)=K*ones(n,1);
for i=1:16
    BK(:,m+1+i)=AK(:,i);
end
BK(:,m+18)=(1:1:n)’;

A=[B1; . . . ; BK];
```

The Splus Function ”geeimport”

```
geedata_importData(”geedata.m”, type=”MATLAB”)

variablen.alt_c(”Indnumber”, ”Zlevel”, ”LTCbegin”, ”LTCend”, ”Transitto”, ”Death”, ”Recovery”,
”DeltaAmb”, ”DeltaStat”, ”Age”, ”G”, ”H”, ”Time”, ”Sex”, ”ZArt”, ”ZLevel”, ”D”, ”R”)


variablen.neu_c(”time”, ”p11”,, ”p12”, ”p13”, ”p14”, ”p21”, ”p22”, ”p23”, ”p24”, ”p31”, ”p32”,
”p33”, ”p34”, ”p41”, ”p42”, ”p43”, ”p44”, ”id”)

variablen_c(variablen.alt, variablen.neu)

dimnames(geedata)[[2]]_variablen

geedata_geedata[order(geedata[,36]), na.last=T]

attach(geedata10)

uns.out.12_gee(p12 ~ Age+C(factor(Sex), treatment)+C(factor(ZLevel), treatment)+C(factor(time),
treatment), quasi(link=logit, variance=constant), geedata, id, time, wc=”uns”, QR=T)
uns.out.12

uns.out.13_gee(p13 ~ Age+C(factor(Sex), treatment)+C(factor(ZLevel), treatment)+C(factor(time),
treatment), quasi(link=logit, variance=constant), geedata, id, time, wc=”uns”, QR=T)
uns.out.13

uns.out.23_gee(p23 ~ Age+C(factor(Sex), treatment)+C(factor(ZLevel), treatment)+C(factor(time),
treatment), quasi(link=logit, variance=constant), geedata, id, time, wc=”uns”, QR=T)
uns.out.23
```

B.2 Bavarian One-Year Mortality Rates (1986-1988)

Age	q_x (Male)	q_y (Female)	Age	q_x (Male)	q_y (Female)
20	0.00131	0.00041	61	0.01658	0.00767
21	0.00126	0.00041	62	0.01806	0.00845
22	0.00121	0.00039	63	0.01970	0.00933
23	0.00117	0.00038	64	0.02153	0.01029
24	0.00112	0.00038	65	0.02357	0.01136
25	0.00109	0.00037	66	0.02583	0.01256
26	0.00106	0.00037	67	0.02835	0.01392
27	0.00104	0.00039	68	0.03114	0.01547
28	0.00104	0.00042	69	0.03426	0.01728
29	0.00106	0.00045	70	0.03774	0.01940
30	0.00111	0.00048	71	0.04164	0.02188
31	0.00119	0.00052	72	0.04597	0.02476
32	0.00126	0.00056	73	0.05077	0.02809
33	0.00131	0.00059	74	0.05607	0.03193
34	0.00138	0.00063	75	0.06190	0.03632
35	0.00148	0.00070	76	0.06830	0.04131
36	0.00159	0.00079	77	0.07530	0.04694
37	0.00174	0.00088	78	0.08293	0.05327
38	0.00191	0.00098	79	0.09120	0.06033
39	0.00210	0.00107	80	0.10015	0.06816
40	0.00228	0.00119	81	0.10978	0.07681
41	0.00247	0.00131	82	0.12011	0.08630
42	0.00270	0.00145	83	0.13116	0.09668
43	0.00296	0.00159	84	0.14295	0.10796
44	0.00327	0.00173	85	0.15550	0.12017
45	0.00360	0.00188	86	0.16881	0.13332
46	0.00395	0.00203	87	0.18290	0.14743
47	0.00435	0.00220	88	0.19778	0.16248
48	0.00480	0.00239	89	0.21345	0.17846
49	0.00531	0.00259	90	0.22991	0.19536
50	0.00590	0.00280	91	0.24715	0.21313
51	0.00654	0.00304	92	0.26515	0.23174
52	0.00724	0.00331	93	0.28391	0.25112
53	0.00800	0.00362	94	0.30338	0.27121
54	0.00882	0.00396	95	0.32353	0.29193
55	0.00972	0.00435	96	0.34431	0.31318
56	0.01068	0.00477	97	0.36569	0.33488
57	0.01171	0.00523	98	0.38759	0.35690
58	0.01281	0.00574	99	0.40995	0.37914
59	0.01398	0.00631	100	0.43271	0.40147
60	0.01523	0.00695	101	1.00000	1.00000

B.3 LTC Incidence Rates (Custodial Insurance, Japan)

Age	i_x (Male)	i_y (Female)	Age	i_x (Male)	i_y (Female)
20	0.00010	0.00010	61	0.00374	0.00374
21	0.00010	0.00010	62	0.00425	0.00425
22	0.00010	0.00010	63	0.00483	0.00483
23	0.00010	0.00010	64	0.00549	0.00549
24	0.00010	0.00010	65	0.00625	0.00625
25	0.00011	0.00011	66	0.00711	0.00710
26	0.00011	0.00011	67	0.00808	0.00808
27	0.00011	0.00011	68	0.00919	0.00919
28	0.00011	0.00011	69	0.01046	0.01046
29	0.00011	0.00011	70	0.01190	0.01190
30	0.00012	0.00012	71	0.01355	0.01354
31	0.00012	0.00012	72	0.01542	0.01542
32	0.00012	0.00012	73	0.01755	0.01755
33	0.00013	0.00013	74	0.01999	0.01998
34	0.00014	0.00013	75	0.02276	0.02276
35	0.00014	0.00014	76	0.02591	0.02592
36	0.00015	0.00015	77	0.02951	0.02953
37	0.00016	0.00016	78	0.03361	0.03364
38	0.00017	0.00017	79	0.03828	0.03834
39	0.00018	0.00018	80	0.04360	0.04369
40	0.00027	0.00027	81	0.04967	0.04980
41	0.00030	0.00030	82	0.05657	0.05678
42	0.00034	0.00034	83	0.06445	0.06475
43	0.00039	0.00039	84	0.07343	0.07386
44	0.00044	0.00044	85	0.08366	0.08427
45	0.00049	0.00050	86	0.09262	0.09336
46	0.00056	0.00056	87	0.10039	0.10120
47	0.00064	0.00064	88	0.10841	0.10933
48	0.00073	0.00072	89	0.11671	0.11777
49	0.00082	0.00082	90	0.12529	0.12657
50	0.00093	0.00093	91	0.13421	0.13579
51	0.00105	0.00105	92	0.14351	0.14542
52	0.00119	0.00119	93	0.15323	0.15550
53	0.00135	0.00135	94	0.16345	0.16612
54	0.00154	0.00154	95	0.17346	0.17651
55	0.00174	0.00174	96	0.18055	0.18373
56	0.00198	0.00198	97	0.18696	0.19014
57	0.00225	0.00225	98	0.19302	0.19604
58	0.00255	0.00255	99	0.19870	0.20136
59	0.00290	0.00290	100	0.20405	0.20605

Bibliography

Aalen, O. O. and S. Johansen (1978). An Empirical Transition Matrix for Non-homogeneous Markov Chains based on Censored Observations. *Scand J Statist 5*, 141–150.

Andersen, P. K., Ø. Borgan, R. D. Gill, and N. Keiding (1993). *Statistical Models Based on Counting Processes.* New York: Springer-Verlag.

Andersen, P. K., J. P. Klein, and S. Rosthøj (2001). From Summary Statistics to Generalized Linear Models for Pseudo-Observations; with Applications to Multi-State Models. *Research Reports 2001 http://www.pubhealth.ku.dk/bsa/publ-e.htm.*

Andersen, P. K., J. P. Klein, and S. Rosthøj (2003). Generalised Linear Models for correlated Pseudo-Observations, with Applications to Multi-State Models. *Biometrika 90, 1*, pp. 15–27.

Bishop, Y. M. M., S. E. Fienberg, and P. W. Holland (1980). *Discrete Multivariate Analysis: Theory and Practice.* Cambridge: The MIT Press.

BMGS (2003). *Pflege Versicherung, Schutz für die ganze Familie.* www.bmgs.bund.de: Bundesministerium füer Gesundheit und Soziale Sicherung.

Borgan, Ø. (1997). Three contributions to the Encyclopedia of Biostatistics: The Nelson-Aalen, Kaplan-Meier, and Aalen-Johansen Estimators. *Encyclopedia of Biostatistics.*

Czado, C. and F. Rudolph (2002). Application of Survival Analysis Methods to Long Term Care Insurance. *Insurance: Mathematics and Economics 31*, 395–413.

Diggle, P. J., K.-Y. Liang, and S. L. Zeger (1996). *Analysis of Longitudinal Data.* Oxford: Oxford Science Publications.

Efron, B. and R. J. Tibshirani (1993). *An Introduction to the Bootstrap.* New York: Chapman & Hall.

Fahrmeir, L., R. Künstler, I. Piegeot, and G. Tutz (2003). *Statistik. Der Weg zur Datenananlyse, 4. verb. Auflage.* Berlin: Springer.

Gart, J. J. (1959). An Extension of the Cramér-Rao Inequality. *The Annals of Mathematical Statistics 30, 2*, 367–380.

Gill, R. D. (2001). Product Integration. *www.math.uu.nl/people/gill.*

Gompertz, B. (1825). On the Nature of the Function Expressive of the Law of Human Mortality, and on a New Mode of Determining Life Contingencies. *Philosophical Transactions of the Royal Society 115*, 513–585.

Haberman, S. and E. Pitacco (1999). *Actuarial Models for Disability Insurance.* London: Chapman & Hall.

Hanselman, D. and B. Littlefield (1997). *Matlab, The Language of Technical Computing.* United States of America: Prentice Hall.

Hougaard, P. (2000). *Analysis of Multivariate Survival Data.* New York: Springer.

Klein, J. P. (1996). Survival Distributions and their Characteristics. *Tech. Report 22 http://www.biostat.mcw.edu/Tech/tr022.pdf.*

Klein, J. P. and M. L. Moeschberger (1997). *Survival Analysis.* New York: Springer.

Liang, K.-Y. and S. L. Zeger (1986). Longitudinal Data Analysis using Generalized Linear Models. *Biometrika 73, 1*, 13–221.

MacDonald, A. S. (1996). An Actuarial Survey of Statistical Models for Decrement and Transition Data. *British Actuarial Journal 2*, 129–155, 429–448, 703–726.

Makeham, W. M. (1860). On the Law of Mortality and the Construction of Annuity Tables. *Journal of the Institute of Actuaries 13*, 325–358.

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models 2nd ed.* London: Chapman and Hall.

McCulloch, C. E. and S. R. Searle (2001). *Generalized, Linear and Mixed Models.* New York: Wiley Series in Probability and Statistics.

Råde, L. and B. Westergren (1997). *Springers Mathematische Formeln, 2. Auflage.* Berlin: Springer.

Rudolph, F. (2000). Anwendungen der Überlebenzeitanalyse in der Pflegeversicherung. *Technische Universität München Diplomarbeit.*

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistic.* New York: Wiley.

Sozialgesetzbuch (1994). Elftes Buch (XI) - Soziale Pflegeversicherung. *Artikel 1 des Gesetzes vom 26. Mai 1994 BGBl. I S. 1014.*

Stoer, J. (1999). *Numerische Mathematik I.* Berlin: Springer.

Trachtenberg, H. L. (1924). The wider Application of the Gompertz Law of Mortality. *Journal of the Royal Statistical Society 87, 2*, 278–290.

Venables, W. N. and B. D. Ripley (2000). *Modern Applied Statistics with S-Plus, Third Edition.* New York: Springer.

Wedderburn, R. W. M. (1974). Quasi-Likelihood Functions, Generalized Linear Models, and the Gauss-Newton Method. *Biometrika 61, 3*, 439–447.

Zeger, S. L. and K.-Y. Liang (1986). Longitudinal Data Analysis for Discrete and Continuous Outcomes. *Biometrics 42, 1*, 121–130.