Modelling Repeated Measurements of Renal Function during dialysis with cut off due to complete kidney failure

Westhoff, G.G.A.

Gerard G.A. Westhoff

Modelling Repeated Measurements of Renal Function during dialysis with cut off due to complete kidney failure

Master thesis, defended on November 12, 2009

Thesis advisors: Prof. dr. R.D. Gill (Mathematisch Instituut) & Dr. S. le Cessie (LUMC)

Mastertrack: Applied Mathematics

Mathematisch Instituut, Universiteit Leiden

Contents

List of Tables

List of Figures

1 The Necosad Study, introduction

The Necosadstudy (NEderlandse COoperatieve Studie naar de Adequaatheid van Dialyse) [3] is a large observational study in which patients are followed after starting renal dialysis.

Healthy kidneys clean the blood by removing excess fluid, minerals, and wastes. They also make hormones that keep the bones strong and the blood healthy. When kidneys fail, harmful wastes build up in the body, blood pressure may rise, and the body may retain excess fluid and not make enough red blood cells. When this happens, treatment is needed to replace the work of the failed kidneys.

In *hemodialysis* (HD), blood is allowed to flow, about 100 grammes at a time, through a special filter that removes wastes and extra fluids. The clean blood is then returned to the body. Removing the harmful wastes and extra salt and fluids helps control blood pressure and keep the proper balance of chemicals like potassium and sodium in the body.

Patients undergoing hemodialysis treatment follow a strict schedule. Most patients go to a clinic three times a week for 3 to 5 or more hours each visit.

In *peritoneal dialysis* (PD), a soft tube called a catheter is used to fill the abdomen with a cleansing liquid called dialysis solution. The walls of the abdominal cavity are lined with a membrane called the peritoneum, which allows waste products and extra fluid to pass from the blood into the dialysis solution. The solution contains a sugar called dextrose that will pull wastes and extra fluid into the abdominal cavity. These wastes and fluid then leave the body when the dialysis solution is drained. The used solution, containing wastes and extra fluid, is then thrown away. The process of draining and filling takes about 30 to 40 minutes.

Since the patient doesn't have to schedule dialysis sessions at a center, PD gives the patient more control. Treatments are possible at home, at work, or on trips.

The data registered by the Necosad study in patients in the month before the start of dialysis, include demographic data (date of birth, sex, ethnic origin), initial therapy and the reason for choosing this, comorbidity, height, body weight and residual urine volume.

At the start of the therapy the kidney usually exhibits some residual function. At regular intervals (3, 6, 12, 18 etc months) the *Residual Glomerular Filtration Rate* (GFR), or short *filtration rate* is measured. GFR is the best overall index of kidney function and describes the flow rate of filtered fluid through the kidney. The filtration rate is corrected for body surface area and is measured in $\mathrm{mL/min/m^2}$.

The normal filtration rate varies according to age, sex, and body size, and declines with age. When dialysis therapy starts, GFR will also depend on the method of dialysis, hemodialysis or peritoneal dialysis.

As a rule, GFR slowly decreases over time, until renal functioning completely breaks down and GFR reaches the value 0. The patient is from then on *anuric*.

In this thesis we will model repeated measurements of the GFR over time; this is called *longitudinal analysis*. Such models are complicated by the fact that as soon as the patient is anuric (GFR = 0), the kidney remains in this state (*absorbing state*), that is, as a random variable, GFR is not distributed normally.

Several models will be studied:

- Longitudinal analysis of multivariate distributed outcomes. In which we will assume the existence of an underlying *latent* variable, normally distributed.
- Markov models in which we model the expected value of GFR at a point in time, given the value at an earlier point in time. This we will implement with and without zero as a special absorbing state.

The original data base is in SPSS format. SPSS is the standard statistical software package used in the LUMC. In this thesis, SPSS is used to perform standard analysis. R is used for more advanced analysis.

2 The dataset

2.1 Cleaning of the dataset

In behalf of the Necosad study, data were collected for 1780 patients.

As the very first step in validation, for each point in time the variable GFR_COR, which is the GFR discussed in this thesis, was set to zero if the variable DIURES was less than 200 mL/24h.

The original data base is in *wide format* which means that each patient is assigned a row and each value of GFR is a separate variable. We transformed it to *long format* in which each measurement of GFR is a row in the data base.

The following patients were excluded:

- Patients for which urine never was collected (variable DIURES missing on all visits, even though the patient participated in the study),
- Patient is already anuric at the start of the therapy, that is, has GFR missing or equal to 0 at the first two measurement occasions.

After exclusion of these patients the data set (wide format) contained $N = 1428$ patients. This is the data set of patients we will use in this thesis.

Each record in the original data base contains measurements made on 16 occasions, which we call *visits*. However, in this study we will only use

the measurements made in the first two years, corresponding to $n = 6$ visits, including the data gathered at the start of the study (*baseline*). The reason is that over time more and more patients drop out of the study, due to kidney transplant, death of the patient, or the patient leaving the Necosad study. When plotting the mean observed GFR for HD and PD for the 16 visits, it is clear that from about visit 6, the effect of therapy on filtration rate becomes less unambiguous. More advanced methods will then be required to study the group effect on GFR.

The 6 measurements are made at $0, 3, 6, 12, 18$ and 24 months after the start of the therapy. In the sequel these points in time will be coded $j = 1, 2, 3, 4, 5, 6$ where $j = 1$ stands for the baseline data.

We checked for typos and other errors in the data. Table (1) shows the range of some patient parameters. There were some obvious errors, e.g. a body length of 1.78 cm, which were corrected by multiplying by 100. In some cases values of the Body Mass Index (BMI) are high, but this can occur. If a BMI of greater than 60 is found, a correction is applied.

Values of the GFR of 50 and higher are peculiar, but we decided not to adjust them since we do not expect that our results will be influenced significantly.

We could not find unrealistic data in other patient parameters.

2.2 General patient characteristics at start of therapy

Table (1) displays some general characteristics about the population in the trial, excluding patients who are anuric at baseline.

To check that the differences between HD and PD are not due to chance, we performed statistical tests, the outcomes of which are also shown in the table.

Patients starting on hemodialysis are 10 years older on average than patients in whom peritoneal dialysis was started. HD patients also had more comorbidity. In patients starting on HD the underlying renal disease was more often related to vascular suffering than in PD patients. In the latter group the incidence of glomerulonephritis was relatively greater. The incidence of diabetes mellitus as the cause of the renal insufficiency was equal in both groups. The GFR in PD patients at the start of dialysis was slightly higher on average.

2.3 Sample graphs of the filtration rate over time

To get a feeling for the filtration rate over time, we plotted the filtration rate over the first two years for 20 randomly selected patients, see figure 1. For each sample, the curves for each indivual are plotted and the means for HD and PD.

		All patients	HD (chronic)	PD (chronic)	Test (p-value)
Number of patients		1428	840	588	
Gender (% men)		63	59	67	
Age	av. (SD)	59.0 (15.2)	63.2 (14.0)	53.1 (14.8)	t-test (< 0.001)
	(min, max)	(18.3,91.6)			
Kahn comorbidity score (%)	Low	40	31	54	$\chi^2(< 0.001)$
	Medium	33	37	28	
	High	26	32	18	
Primary Renal Disease (%)	Diabetes	16	16	16	$\chi^2(< 0.001)$
	Glomerulonephritis	14	9	20	
	Renal Vascular Conditions	18	22	12	
	Other	53	53	52	
GFR	av. (SD)	5.42 (3.51)	5.13 (3.69)	5.78 (3.24)	t-test (0.003)
	(min, max)	(0.0, 51.7)			
BMI	av. (SD)	25.0 (4.7)	25.0 (4.4)	25.0 (5.0)	t-test (> 0.50)
	(min, max)	(14.9, 94.5)			
length (cm)	av. (SD)	172 (10)	170 (10)	173 (11)	t-test (< 0.001)
	(min, max)	(77, 207)			
weight (kg)	av. (SD)	73.5 (14.3)	72.6 (14.2)	74.9 (14.5)	t-test (0.004)
	(min, max)	(34.0, 150.0)			

Table 1: Patient characteristics at the start of the treatment, excluding patients who are anuric at baseline

For some patients, GFR can be seen to remain zero from a certain point in time, which means kidney function completely brakes down. As referred to in the introduction, the kidney then enters an absorbing state, illustrating the complication that GFR is continuous, but not normally distributed.

In other cases, GFR curves are interrupted, due to kidney transplant, death of the patient, or the patient leaving the Necosad study. This is also illustrated in figure 2, which shows the decreasing number of patients still participating in the study. The small number of measurements after visit 6, explains why we use only data for the first 2 years (first 6 visits, including baseline). Returning to the complete data set (after cleaning up), figure 3 displays a graph of the observed mean GFR by therapy; patients who are anuric from the start of the treatment, are excluded ($N = 1428$).

2.4 Defining patients as anuric

In the Necosad study, a patient is said to be *anuric*, that is, has kidneys that no longer function, if he or she has GFR $= 0$ at two consecutive visits.

Figure 4 shows the number patients who are for the first time anuric. p_j is the estimated probability of a patient who is not anuric at time t_{j-1} to become anuric at time t_j

$$p_j = \frac{\#\text{ persons for the first time anuric}}{\#\text{ persons for the first time anuric} + \#\text{ persons not yet anuric}}.$$

ρ_j is the estimated probability to be not yet anuric at time t_j

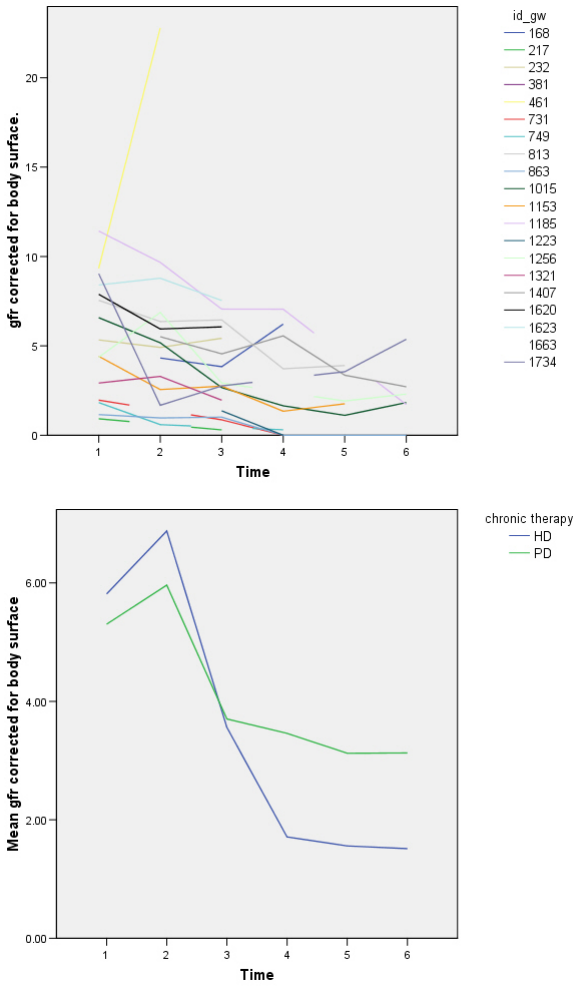

Figure 1: GFR and mean GFR by therapy (bottom)for 20 randomly selected patients. (Visit 1 =baseline)

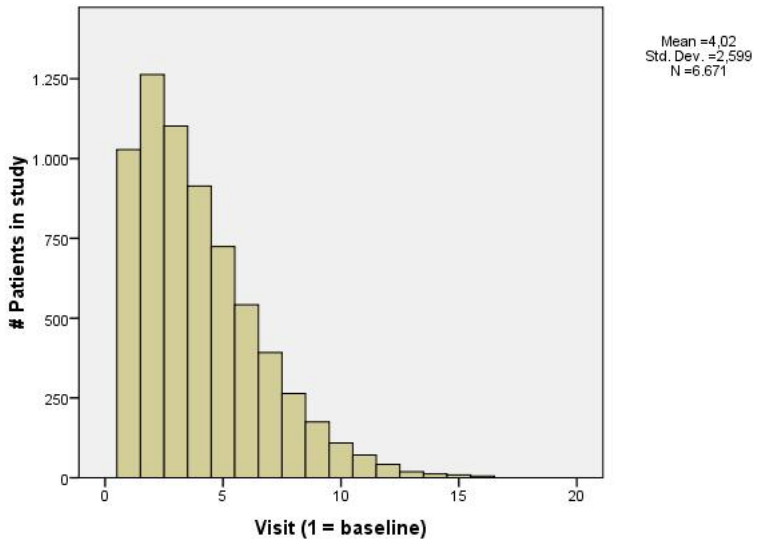

Figure 2: Number of patients with a GFR measurement (GFR ≥ 0), excluding patients who are anuric at baseline.

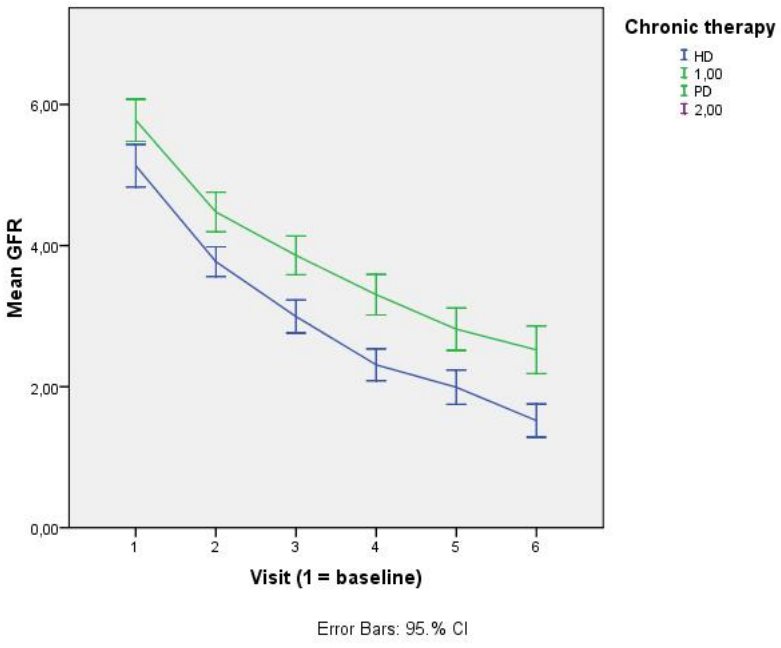

Figure 3: Observed mean GFR, excluding patients who are anuric at baseline.

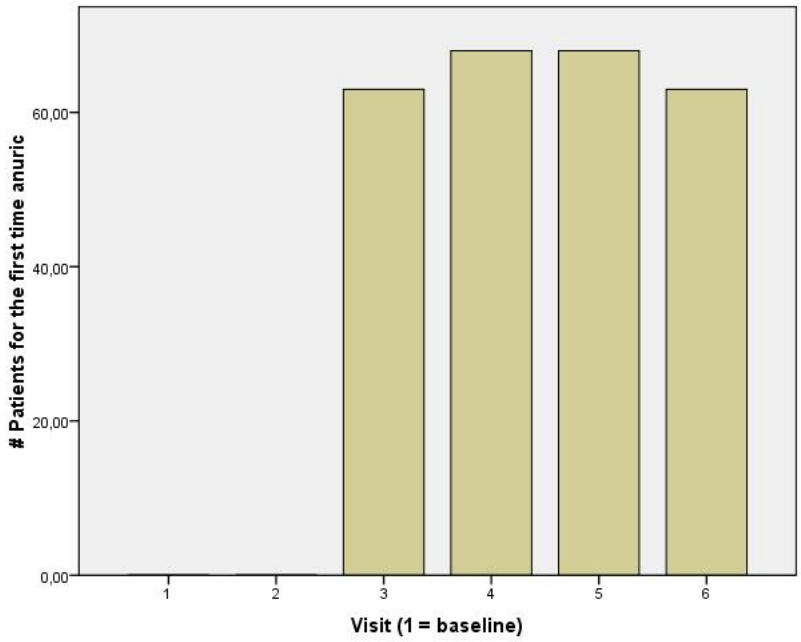

Figure 4: Number of patients who are for the first time anuric.

Visit j	patients becoming anuric	patients not yet anuric	p_j	ρ_j
1 (= baseline)	0	1028	0	1.00
2	0	1263	0	1.00
3	63	1039	0.0572	0.943
4	68	808	0.0776	0.870
5	68	588	0.104	0.779
6	63	405	0.135	0.675

Table 2: Probability for a patient to become anuric

$$\rho_1 = (1-p_1), \quad \rho_2 = (1-p_1)(1-p_2), \ldots, \rho_j = (1-p_1)\ldots(1-p_j), \quad j = 1, \ldots, n.$$

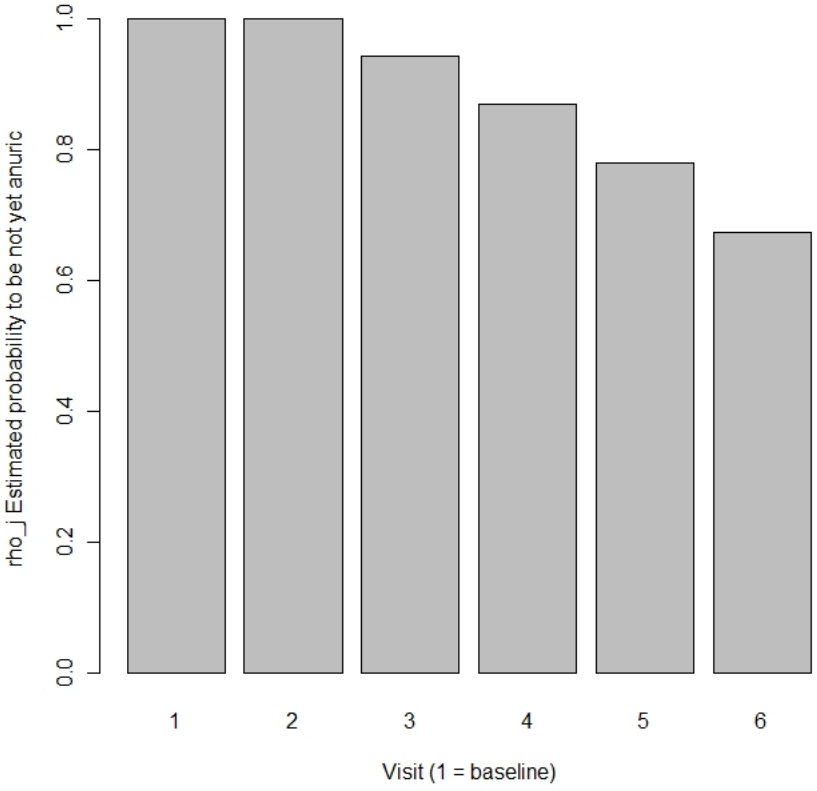

Figure 5: Estimated probability to be not yet anuric.

3 Models for longitudinal data, some general aspects

This outline is based on [1], Part II, in particular chapters 3, 4 and 5. This is the standard approach to the analysis of repeated measurements which also can be executed by standard software packages like SPSS

Assumption: longitunal responses have a multivariate normal distribution. Theory is based on this assumption, but is not required (asymptotic behaviour). In this section we will use notation and terminology employed by Fitzmaurice et al: the participants are referred to as *individuals* or *subjects*. The individuals are measured repeatedly at different *occasions* or *times*.

3.1 Notation

Consider a sample of N subjects, measured repeatedly over time. In this study, all measurements are scheduled to take place on the same n occasions (*balanced* design). Let Y_{ij} denote the random response variable (e.g. kidney filtration rate), for the i^{th} subject ($i = 1, \ldots, N$) on the j^{th} measurement occasion ($j = 1, \ldots, n$).

Due to missing data and drop outs, usually not all subjects are measured at all time points. The term *drop out* refers to the special case where, if Y_{ik} is missing, then $Y_{i,k+1}, \ldots, Y_{in}$ are also missing. In the Necosad study, patients "drop out" when they leave the study, die, or have a kidney transplant. Drop outs will not be discussed, but missing data will be examined in section 3.3.2

We now first consider the model without missing data.

We group the responses for subject i in the vector

$$Y_i = \begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in} \end{pmatrix}, \quad i = 1, \ldots, N$$

We may expect the random variables Y_i to be independent of each other, e.g. it is not to be expected that the set of n observed filtration rates for one patient, will influence those of other patients. But, the repeated measures on the *same* patient are certainly not expected to be independent observations.

Associated with each respons Y_{ij}, there is a p-vector of covariates

$$X_{ij} = \begin{pmatrix} X_{ij1} \\ X_{ij2} \\ \vdots \\ X_{ijp} \end{pmatrix}, \quad i = 1, \ldots, N; \quad j = 1, \ldots, n,$$

in which each element corresponds to a different covariate. Covariates can be grouped into two main types: those whose value do not change for the duration of the study and covariates whose values change over time. To the first group belong parameters like gender and fixed experimental treatments, e.g. mode of dialysis at the start of the study.

Now consider a linear regression model for changes in the mean response over time and for relating the changes to the covariates,

$$(1) \qquad Y_{ij}|X_{ij} = \beta_1 X_{ij1} + \beta_2 X_{ij2} + \cdots + \beta_p X_{ijp} + e_{ij}, \quad j = 1, \ldots, n,$$

where $\beta_1, \ldots, \beta_p$ are unknown regression coefficients, relating the mean of Y_{ij} to its corresponding covariates. The e_{ij} are random errors, with mean zero. Taking the mean:

$$(2) \qquad \mathbb{E}(Y_{ij}|X_{ij}) = \mu_{ij} = \beta_1 X_{ij1} + \beta_2 X_{ij2} + \cdots + \beta_p X_{ijp}$$

If $X_{ij} = 1$, for all subjects i and all time points j, then β_1 is the intercept term in the model.

There are n separate regression equations for the response variables and the regression model can be expressed in the compact form

$$(3) \qquad Y_i = X_i\beta + e_i$$

in which

$$X_i = \begin{pmatrix} X_{i11} & X_{i12} & \ldots & X_{i1p} \\ X_{i21} & X_{i22} & \ldots & X_{i2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{in1} & X_{in2} & \ldots & X_{inp} \end{pmatrix},$$

$\beta = (\beta_1, \beta_2, \cdots \beta_p)'$ and $e_i = (e_i1, e_i2, \ldots, e_{in})'$ (The $'$ denotes the tranpose of a vector or matrix). X_i is often called the *design matrix*. The mean of Y_i is

$$(4) \qquad \mathbb{E}(Y_i|X_i) = \mu_i = X_i\beta,$$

where $\mu_i = (\mu_{i1}, \ldots, \mu_{in})'$ is the vector of means for the i^{th} individual.

3.2 Distributional assumptions

As stated before, we suppose that for each individual i the repeated measurements have a multivariate normal distribution. This is in fact the distribution of the random variables $e_i \sim N(0, \Sigma_i)$, but with mean $\mu_= X_i\beta$.

$$Y_i \sim N(\mu_i, \Sigma_i),$$

in which Σ_i is the covariance matrix for this individual:

(5)

$$\Sigma_i = Cov(Y_i) = Cov\begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in} \end{pmatrix} = \begin{pmatrix} Var(Y_{i1}) & Cov(Y_{i1}, Y_{i2}) & \dots & Cov(Y_{i1}, Y_{in}) \\ Cov(Y_{i2}, Y_{i1}) & Var(Y_{i2}) & \dots & Cov(Y_{i2}, Y_{in}) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(Y_{in}, Y_{i1}) & Cov(Y_{in}, Y_{i2}) & \dots & Var(Y_{in}). \end{pmatrix}.$$

We assume the Y_i to be independent of each other, but we consider the repeated measures on the same subject i to be dependent; in general, Σ_i will not be a diagonal matrix.

In this study, where all participants have the same number n of repeated measures, obtained at a common set of occasions, and where there is no dependence of the covariance matrix on the covariates, we can drop the the index i and simply denote the covariance matrix by Σ:

$$\Sigma = \Sigma_i = Cov(Y_i), \quad i = 1, \dots, N.$$

The assumption of the homogeneity of covariance, is the multi-variate analog of the of the assumption of homogeneity of *variance* in linear regression for a *univariate* response.

3.3 The multi-variate normal distribution

The multivariate normal joint probability density function for Y_i equals

(6)
$$\begin{aligned} f(y_i) &= f(y_{i1}, \dots, y_{ij}, \dots, y_{in}) \\ &= \frac{1}{(2\pi)^{n/2}\sqrt{\det\Sigma}} e^{-\frac{1}{2}(y_i-\mu_i)'\Sigma^{-1}(y_i-\mu_i)}, \end{aligned}$$

where $-\infty < y_{ij} < \infty$.

3.3.1 Estimating regression parameters and the covariance matrix

All of the models for longitudinal data analysis can be expressed in terms of a general linear regression model for the mean response vector

$$\mathbb{E}(Y_i|X_i) = \mu_i = X_i\beta,$$

where the response vector Y_i is assumed to arise from a multivariate normal distribution. We consider a balanced design without missing data $n_i = n, \quad i = 1, \dots, N$ in which all subjects to have the same covariance matrix

(7)
$$Cov(Y_i) = \Sigma_i = \Sigma(\theta),$$

in which θ is a vector of length q. If the covariance matrix is *unstructured* as in section (3.5.1), the elements of q are the n variances and $n(n-1)/2$ pairwise covariances stacked in a single q-vector, where $q = n(n+1)/2$.

In this section we discuss methods for estimating the unknown parameters β and θ (or Σ).

The method of maximum likelihood for correlated observations
The responses Y_i are multivariate normally distributed and entirely specified by the mean vector $X_i\beta$ and covariance matrix $\Sigma(\theta)$. In order to build a model that explains the data these quantities have to be estimated. A very general approach to estimation is the method of *maximum likelihood*. We give an introduction and refer for more details to e.g. [4].

The maximum likelihood estimates of β and θ are those values of β and θ that maximize the joint probability of the response variables evaluated at their observed values. The probability of the response variables evaluated at the fixed set of observed values and regarded as functions of β and θ, is known as the *likelihood function*. Taking the log we obtain the *log-likelihood function* l. Thus β and θ are estimated by maximizing l. The estimates for β and $\Sigma(\theta)$ thus obtained are usually denoted by $\widehat{\beta}$ and $\widehat{\Sigma(\theta)}$.

When there are n repeated measurements on the same individual, it cannot be assumed that these are independent. As a result we need to consider joint probability density function for the vector of repeated measurements, as expressed by (6). We assumed that the vectors of repeated measures for different subjects are independent of one another. Thus, the log-likelihood function l can be expressed as a sum of the individual multivariate normal probability density functions for Y_i.

To find the maximum likelihood estimate (MLE) of β in the repeated measurements setting, we first suppose that θ is known. and therefore does not need to be estimated. Later we will relax this assumption. To determine the estimate of β, we must maximize the following log-likelihood function:
(8)

$$l = -\frac{nN}{2}log(2\pi) - \frac{1}{2}\sum_{i=1}^{N} log(\det\Sigma(\theta)) - \frac{1}{2}\left\{\sum_{i=1}^{N}(y_i-\mu_i)'\Sigma(\theta)^{-1}(y_i-\mu_i)\right\}.$$

We see that maximizing l with respect to β is equivalent to minimizing

$$\left\{(y_i-\mu_i)'\Sigma(\theta)^{-1}(y_i-\mu_i)\right\}$$

The estimator of β that minimizes this expression is known as the *generalized least squares* (GLS) estimator of β and can be expressed as

$$\widehat{\beta} = \left\{\sum_{i=1}^{N}(X_i'\Sigma(\theta)^{-1}X_i)\right\}^{-1}\sum_{i=1}^{N}(X_i'\Sigma(\theta)^{-1}y_i). \tag{9}$$

Usually however, we do not know θ. Instead we typically must estimate $\Sigma(\theta)$ from the data at hand. Maximum likelihood estimation of θ proceeds in the same way as with estimation of β and is obtained by maximizing the log-likelihood with respect to θ. To this end we take the derivative with respect to θ and the result, called the *score function*, and equate the result to zero. Unfortunately, this equation is non-linear and it is generally not

possible to write the estimator of θ in closed form. Instead we have to rely on iterative techniques. Computer algorithms have been developed to find the solution.

Note, that with *unstructured* covariance matrices, when responses are measured at the same n occasions, that the elements of the covariance matrix are simply the empirical covariances!

Once the estimate of θ has been obtained, we then simply substitute of $\Sigma(\theta)$, which we write as $\widehat{\Sigma} = \Sigma(\widehat{\theta})$, into the estimator of β given by (9) to obtain the following MLE of β:

$$\widehat{\beta} = \left\{ \sum_{i=1}^{N} (X_i' \widehat{\Sigma}^{-1} X_i) \right\}^{-1} \sum_{i=1}^{N} (X_i' \widehat{\Sigma}^{-1} y_i). \tag{10}$$

The MLE $\widehat{\beta}$ of β has some interesting and important large sample properties:

- The MLE is a *consistent estimator* of β, that is, it converges in probability to the true value of β, voor all β (blz 91). If the distribution of the errors e_i is normal or even just symmetric, then $\widehat{\beta}$ is also an *unbiased* estimator of β.
- The MLE is asymptotically *unbiased*: its bias tends to zero as the sample size increases to infinity.
- The MLE is asymptotically *efficient*. This means that no asymptotically unbiased estimator has lower asymptotic mean squared error than the MLE.
- The MLE is asymptotically normal. The distribution for large samples is multivariate normal, with mean β and covariance

  $$\mathrm{Cov}(\widehat{\beta}) = \left\{ \sum_{i=1}^{N} (X_i' \Sigma(\theta)^{-1} X_i) \right\}^{-1},$$

  in which Σ is estimated from the data.

Asymptotic normality extends to the incomplete data setting, when certain assumptions about missingnes hold.

Restricted Maximum Likelihood Estimate When estimating regression parameters with SPSS, choosing Linear Mixed Models, we employ an alternative MLE, namely *Restricted Maximum Likelihood Estimate* (REML). In essence, the REML method deals with linear combinations of the observed values whose expectations are zero. These "error contrasts" are free of any fixed effects in the model. In contrast to maximum likelihood estimates, REML estimates of variances and covariances are known to be *unbiased*—of particular importance when dealing with small samples.

REML estimates in balanced designs are identical to Analysis of Variance estimates.

REML estimates of variance components are known to be unbiased in balanced designs and have the same asymptotic distributional properties as maximum likelihood estimates. These properties can be used to test hypotheses about population variances and covariances, provided sample sizes are sufficiently large.

3.3.2 The problem of missing values in longitudinal data

This discussion is in large part based on [2], page 208 ff.

The key question for analyses with missing data is, under what circumstances, if any, do the analyses we would perform if the data set were fully observed lead to valid answers?

Let Y^* denote the complete set of measurements, e.g. GFR over time, which would be have been obtained were there no missing values and partition this set into $Y^* = (Y^{(o)}, Y^{(m)})$ with $Y^{(o)}$ denoting the measurements actually obtained and $Y^{(m)}$ the measurements which would have been available had they not been missing, for whatever cause. Finally, let R denote the set of indicator random variables, denoting which elements of Y^* fall into $Y^{(o)}$ and which into $Y^{(m)}$.

A probability model for the missing value mechanism is a specification of the probability distribution of R conditional on $Y^* = (Y^{(o)}, Y^{(m)})$. The missing value mechanism can be classified as:

- *completely random* if R is independent of both $Y^{(o)}$ and $Y^{(m)}$. The abbreviation used is MCAR (*Missing Completely At Random*). If data are MCAR, then consistent results with missing data can be obtained by performing the analyses we would have used had their been no missing data, although there will generally be some loss of information. In practice this means that, under MCAR, the analysis of only those units with complete data gives valid inferences.

  Example: In the Necosad setting, when staff measures the GFR, but then lose the data, these data are MCAR: the reason for missing data is not related to the outcome of the measurement.

- *random* (*Missing At Random*, MAR)if R is independent of $Y^{(m)}$. This is equivalent to saying that the behavior of two subjects who share measured values have the same statistical behavior on the other observations, whether observed or not. Under MAR, the probability of a value being missing will generally depend on measured values, so it does not correspond to the intuitive notion of "random".

  Example: It is decided that the patient receives a kidney tranplantation, based on the observed value of GFR.

- *informative* (or *Missing Not At Random*, MNAR)if R is dependent of $Y^{(m)}$. Even accounting for all the available observed information, the reason for measurements being missing still depends on the unseen measurements themselves.

  Example: Replacing the zero value of GFR in the anuric phase of a patient by the code for missing value.

Drop out refers to a special case of missing data where if Y_{ik} is missing, then $Y_{i,k+1}, \ldots, Y_{in}$ are also missing. A patient in the Necosad study "drops out", e.g. due to kidney transplant, death of the patient, or the patient leaving the study.

We now show that for likelihood-based inference, the crucial distinction is between random and informative missing values. To see this, let $f(y^{(o)}, y^{(m)}, r)$ be the joint probability density function of $(Y^{(o)}, Y^{(m)}, R)$ and use the standard factorization to express this as

$$f(y^{(o)}, y^{(m)}, r) = f(y^{(o)}, y^{(m)}) f(r|y^{(o)}, y^{(m)}) \tag{11}$$

For a likelihood based analysis, we need the joint pdf of the observable random variables $(Y^{(o)}, R)$, which we obtain by integrating (11):

$$f(y^{(o)}, r) = \int f(y^{(o)}, y^{(m)}) f(r|y^{(o)}, y^{(m)}) dy^{(m)} \tag{12}$$

Now if the missing value mechanism is random, $f(r|y^{(o)}, y^{(m)})$ does not depend on $y^{(m)}$ and (12) becomes

$$\begin{aligned} f(y^{(o)}, r) &= f(r|y^{(o)}) \int f(y^{(o)}, y^{(m)}) dy^{(m)} \\ &= f(r|y^{(o)}) f(y^{(o)}) \end{aligned} \tag{13}$$

If we now take logarithms in the last display, the log-likelihood function is

$$l = log f(r|y^{(o)}) + log f(y^{(o)}), \tag{14}$$

which is maximized by separate maximization of the two terms on the right hand side. Since the first term contains no information about the distribution of $Y^{(o)}$, we can ignore it for the purpose of making inferences about $Y^{(o)}$. This explains why the maximum likelihood estimator of β given by (10), extends to the incomplete data setting when the missing value mechanism is *random*; see section 3.3.1. However, in general non-likelihood methods (e.g. based on individuals with fully observed data, moments, estimating equations & including generalized estimating equations) are not valid under MAR, although some can be "fixed up". In particular, ordinary means, and other simple summary statistics from measured data, will be biased.

Because of the above result, both completely random and random missing value mechanisms are sometimes referred to without distinction as *ignorable*.

3.4 Modeling the mean response over time

Much of the focus in the analysis of longitudinal data is on the mean response μ_i. There are two approaches for modeling the mean response over time: the *analysis of response profiles* and *parametric or semi-parametric curves*, e.g, piecewise linear curves. The analysis of response profiles is an example of a non-parametric approach and offers a model free calculation of the the mean response.

3.4.1 Non-parametric curves: the analysis of response profiles

Methods for analyzing response profiles are appealing when there is a single categorical covariate (e.g. denoting a different therapy) and when no specific *a priori* pattern for the differences in the response profiles between groups can be specified. When repeated measures are obtained at the same sequence of occasions, the data can be summarized by the estimated mean response at each occasion, stratified by levels of the group factor. At any given level of the group factor, the sequence of means over time is referred to as the mean *response profile*. The analysis of response profiles can also handle incompleteness due to missing data.

3.4.2 Parametric curves

Here we presume a parametric curve, e.g. a linear or quadratic trend, for the mean response over time. This allows for a dramatically reduced number of model parameters; by their very nature, parametric curves provide a very parsimonious description of trends in the mean response over time, and of covariate effects on then mean response in time. For example, a linear trend in the mean response can be characterized by a single regression parameter that has an interpretation in terms of the constant rate of change in the mean response.

In addition, parametric curves describe the mean as an explicit function of time. As a result, and in contrast to profile analysis, there is no necessity to require that all cases in the study have been measured at the same time points. However, parametric curves impose an explicit structure on the mean responses.

Linear trend over time is the simplest parametric curve that can be used to describe changes in the mean response over time.

3.5 Modeling the covariance

The defining feature of longitudinal data, is that repeated responses are obtained on the same individuals over time and the responses on the same individual are correlated. Accounting for the correlation among repeated measures completes the specification of any regression model for longitudinal

data and usually increases efficiency with which the regression parameters can be estimated. In addition, when there are missing data, correct modeling of the covariance is often a requirement for obtaining valid estimates of the regression parameters.

Longitudinal data present us with two aspects of the data that require modeling: the mean response over time and the covariance among repeated measures on the same individuals. These two aspects of the data are interrelated and the choice of models for the mean response and the covariance are interdependent. This interdependence arises because the vector of residuals depends upon the specification of the model for the mean.

The covariance among repeated measures can be modeled in three different ways: *unstructured covariance*, *covariance pattern models* and *random effects covariance structures*, which we will not discuss.

3.5.1 Unstructured covariance

This allows for any arbitrary pattern of covariance among the repeated measures. This is referred to as *unstructured covariance*: no explicit structure is assumed, other than the homogeneity of covariance.

Thus, when there are n repeated measurements, n variances and $n \times (n-1)/2$ pairwise covariances (or correlations) are estimated.

There are two potential drawbacks with this technique. First, the number of covariance parameters can be quite large; with n measurements, the $n \times n$ covariance matrix has $n \times (n+1)/2$ unique parameters. Thus in a longitudinal study with 10 measurement occasions, an unstructured covariance has 55 parameters (10 variances and 45 covariances). And, when the number of covariance parameters to be estimated is large relative to the sample size, estimates are likely to be unstable, due to errors adding up. In the second place, such an approach only makes sense when all subjects are measured at the same occasions.

In section 4 this model is applied to the GFR data set.

3.5.2 Covariance pattern models

This approach borrows ideas from the statistical literature on time series analysis. Time series analysis follows a single subject and tries to summarize, in as few parameters as possible, the trend in time of the quantity of interest. In longitudinal analysis a (large) group of subjects is followed, but now *group* properties like mean response for a given subgroup are important. Both analyses share a common feature: the repeated measures are (positively) correlated. Also, if the measurements are taken closer in time, usually they are more highly correlated than repeated measurements; thus correlations decay as the time separation increases.

When attempting to impose some structure on the covariance a balance needs to be struck. If too little structure is imposed there will be too many parameters to estimate with the limited amount of data at hand and this will adversely affect the precision the precision with which the main parameter of interest, the vector β of covariates, can be estimated. This is one of the drawbacks of the unstructured covariance considered in the previous section.

When structure is imposed on the covariance, it is possible to improve the precision with which β can be estimated. However, if *too* much structure is imposed, there is a risk of model misspecification, that could ultimately result in misleading inferences concerning β; this is the classic trade off between bias and precision.

Quite often, the correlation among repeated measures is expressed as an explicit function of the distance in time, in which case these models can be used with unequally spaced observations. Many of the models assume *stationarity*, in which variance does not change as a function of time. This way, we can model the covariance structure with only a few parameters.

Here are some examples:

Compound symmetry covariance matrix With compound symmetry covariance, it is assumed that the variance σ^2 is constant across occasions and subjects i and $Corr(Y_{ij}, Y_{ik}) = \rho$ for all i, j and k and $j \neq k$. For $j = k$, the correlation is equal to 1. That is,

$$\Sigma = Cov(Y_i) = \sigma^2 \begin{pmatrix} 1 & \rho & \rho & \dots & \rho \\ \rho & 1 & \rho & \dots & \rho \\ \rho & \rho & 1 & \dots & \rho \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \rho & \dots & 1. \end{pmatrix},$$

with the constraint that $\rho \geq 0$. The compound symmetry covariance is very parsimonious, with only two parameters, regardless of the number of observations. However, it makes the assumption that the correlation between any pair of measurements is the same regardless of the time interval between the measurements, which is rather unrealistic, given the empirical fact that for most longitudinal data, the correlations are expected to decay with time. Also, the assumption of constant variance across time is unrealistic in many settings.

Toeplitz covariance This pattern makes the assumption that any pair of responses that are equally separated in time have the same correlation. It is assumed that the variance σ^2 is constant across measurement occasions

and subjects i and that $Corr(Y_{ij}, Y_{i,j+k}) = \rho_k$ for all i, j and k, $j \neq k$:

$$\Sigma = Cov(Y_i) = \sigma^2 \begin{pmatrix} 1 & \rho_1 & \rho_2 & \dots & \rho_{n-1} \\ \rho_1 & 1 & \rho_1 & \dots & \rho_{n-2} \\ \rho_2 & \rho & 1 & \dots & \rho_{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{n-1} & \rho_{n-2} & \rho_{n-3} & \dots & 1. \end{pmatrix},$$

with $\rho_k \geq 0$. This structure is only appropriate when the observations are made at equal (or approximately) intervals of time. Note that the Toeplitz covariance has n parameters. A special case of Toeplitz covariance is the (first-order) autoregressive covariance.

Autoregressive covariance The variance σ^2 is constant across time and subjects i and $Corr(Y_{ij}, Y_{i,j+k}) = \rho^k$ for all i, j and k, $j \neq k$, and $\rho \geq 0$:

$$\Sigma = Cov(Y_i) = \sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{n-1} \\ \rho & 1 & \rho & \dots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \dots & \rho^{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \dots & 1 \end{pmatrix}.$$

There are only two parameters. The correlations decline over time as the separation between pairs of repeated measures increases, but in many settings the correlations rarely decay that quickly.

The autoregressive process is said to be *first order* (AR(1)) because the error e_{ij} in the observation Y_{ij} only depends on the previous error $e_{i,j-1}$; dependence on the two previous errors would yield a *second-order* autoregressive process. The first-order autoregressive covariance is a process where the error term at the j^{th} occasion is a deterministic function of the error at the previous occasion (i.e. the recent past predicts the present), plus an additional (and independent) source of random error, w_{ij}: $e_{ij} = \rho e_{i,j-1} + w_{ij}$. For such a process, it can be shown that $Var(e_{ij}) = \rho^2$ and $Cov(e_{ij}, e_{ik}) = \sigma^2 \rho^{|j-k|}$.

The Markov model for the mean response of the filtration rate GFR over time, to be discussed in section 5, is an example of a first order autoregressive process.

Banded, exponential and hybrid models The *banded* covariance patterns make the assumption that the correlation is zero beyond some specified interval. It is possible to apply an banded pattern to any of the covariance pattern models considered so far. In longitudinal studies in the health sciences, it is rare for the correlation to decay to zero, even in studies where there is a lengthy period of follow up.

In the *exponential* covariance model, the correlation between any pair of repeated measures decreases exponentially with the time separation between them. This is a generalization of autoregression, suitable when the measurement times are not equally spaced over time. Let $t_{i1}, \ldots, t_{in}$ denote the observation times for the i^{th} subject and assume that the variance σ^2 is constant across measurements and subjects i. Then the correlation between observations Y_{ij} and Y_{ik} could be written as $Corr(Y_{ij}, Y_{ik}) = \rho^{|t_{ij}-t_{ik}|}$ and for the covariance we would have $Cov(Y_{ij}, Y_{ik}) = \sigma^2 \exp(-\theta|t_{ij} - t_{ik}|)$ for some $\theta \geq 0$.

A distinctive feature of the exponential model is that it assumes that the correlation is one if measurements for an individual are made repeatedly at the same occasion. This corresponds to the assumption that the responses are measured without error; an unrealistic assumption in most longitudinal studies in the health sciences.

In *hybrid* models, autoregressive and compound symmetry models are combined, thereby overcoming the less appealing aspects of these models for longitudinal data. In this model,

$$\begin{aligned} Var(Y_{ij}) &= \sigma_1^2 + \sigma_2^2, \\ Cov(Y_{ij}, Y_{ik}) &= \rho_1 \sigma_1^2 + \rho_2^{|t_{ij}-t_{ik}|} \sigma_2^2, \\ Corr(Y_{ij}, Y_{ik}) &= \frac{\rho_1 \sigma_1^2 + \rho_2^{|t_{ij}-t_{ik}|} \sigma_2^2}{\sigma_1^2 + \sigma_2^2}. \end{aligned}$$

This implies that the correlation between replicate measurements on an individual obtained at the same occasion is

$$\frac{\rho_1 \sigma_1^2 + \sigma_2^2}{\sigma_1^2 + \sigma_2^2},$$

which is less than one when $\rho_1 < 1$. The correlation no longer decays to zero but has a minimum of

$$\frac{\rho_1 \sigma_1^2}{\sigma_1^2 + \sigma_2^2},$$

which is greater than zero provided $\rho_1 > 0$.

4 Models for longitudinal data analysis applied to the GFR data set

In this chapter we apply the theory of longitudinal analysis, as discussed in chapter 3, to the GFR data set.

LUMC epidemiologists analyze the GFR data over time, modeling the non-anuric phase of the kidney (see section 2.4). Measurements *after* anury are discarded and the researchers assume that these measurements are missing

at random. This way, the extrapolated GFR measurements are allowed to become negative.

The assumption that the data in the anuric phase are MAR is not realistic; it can hardly be maintained that data are missing at *random* after such a deliberate act. This means that the MLE no longer has the nice properties listed in section 3.3.1.

LUMC epidemiologists start with performing an analysis of response profiles. When the curve of mean responses looks reasonably linear, they try to fit a linear model, but now with time as continuous covariate. If this is not the case, another approach to modeling has to be found.

We will discuss both techniques in this chapter.

4.1 Modeling GFR using analysis of response profiles

If Y_{ij} represents the filtration rate for patient i at time j, the mean GFR for patient i at visit j obeys the general model

$$\mu_{ij} = \mathbb{E}(Y_{ij}|X_{ij}) = \beta_1 X_{ij1} + \beta_2 X_{ij2} + \cdots + \beta_p X_{ijp} \quad i = 1, \ldots, N; j = 1, \ldots, n.$$

In our example, the regression coefficients β_k ($k = 1, \ldots, p$) refer to intercept, time, group or group×time interaction; $X_{ij1} \equiv 1$ and $X_{ij2} = 0$ if patient i was assigned to HD, 1, if the patient was assigned to PD. The other covariates are indicator variables referring to time or group×time interaction. This way, we can test for main effects of group and time or for a group×time interaction effect. E.g. when testing for group×time interaction effects the null hypothesis is of parallel mean response profiles; see section 3.4.1.

With $n = 6$ measurement occasions we would have $p = 2n + 2 = 14$ regression coefficients and the above expression for μ_{ij} would read as

$$\begin{aligned}(15)\quad \mu_{ij} = \beta_1 + \beta_2 \texttt{group}_i + \beta_3[\texttt{time} = 1] + \cdots + \beta_8[\texttt{time} = 6] \\ + \beta_9[\texttt{time} = 1] \times \texttt{group}_i + \cdots + \beta_{14}[\texttt{time} = 6] \times \texttt{group}_i,\end{aligned}$$

where $\texttt{group}_i = 1$ if the i-th patient was receiving HD therapy and $\texttt{group}_i = 0$ otherwise. This model is succinctly written as

$$\text{GFR} = \texttt{group + time + group*time.}$$

Table 3 lists the estimates for the regressions coefficients, calculated by SPSS. Let us try to reproduce the estimated means for GFR by using the estimates in table 3, as calculated by SPSS. We will do this manually for two time points. Add the estimated regression coefficients from the table, substituting in (15), for baseline and HD:

$$\begin{aligned}\beta_1 + \beta_2 \times 1 + \beta_3 \times 1 + \cdots + \beta_8 \times 0 + \beta_9 \times 1 \times 1 + \cdots + \beta_{14} \times 0 \times 1 = \\ = 2.475061 - 1.004013 + 3.428899 + 0.376508 \\ = 5.276455.\end{aligned}$$

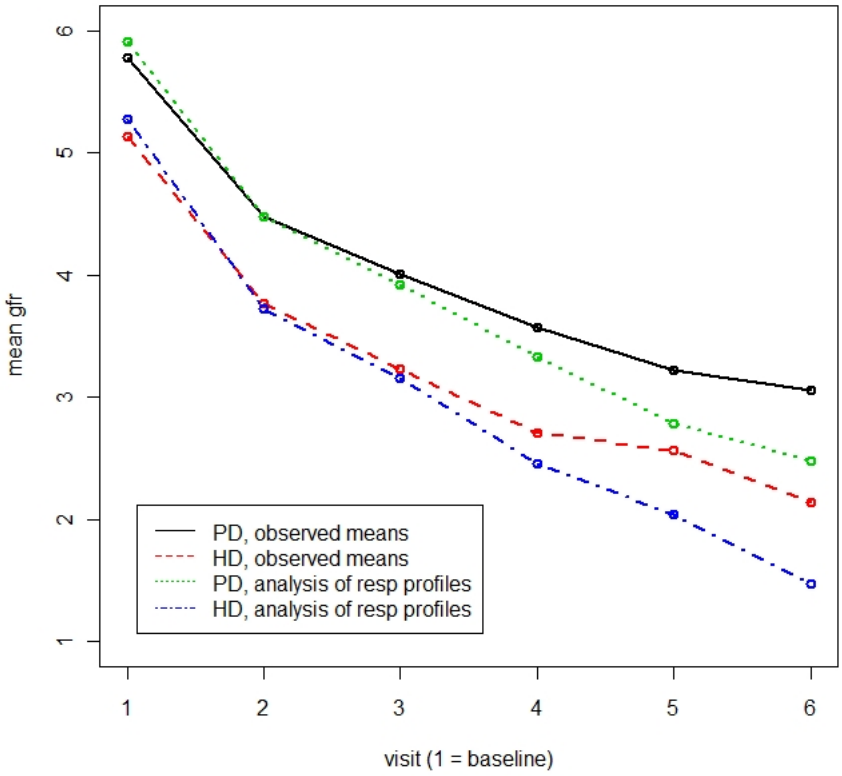

Figure 6: Observed means and means calculated by the analysis of response profiles.

Parameter	Therapy	Visit	Regr coeff	$\hat{\beta}$	Std. Error	p
Intercept			β_1	2.475061	0.154212	< 0.0001
Therapy	HD		β_2	-1.004013	0.212971	< 0.0001
Visit (1 = baseline)		1	β_3	3.428899	0.189173	< 0.0001
Visit		2	β_4	2.005964	0.160903	< 0.0001
Visit		3	β_5	1.447851	0.149859	< 0.0001
Visit		4	β_6	0.856431	0.124339	< 0.0001
Visit		5	β_7	0.305941	0.118182	0.01
Visit		6	β_8	0	0	.
Visit x therapy	HD	1	β_9	0.376508	0.260523	0.15
Visit x therapy	HD	2	β_{10}	0.248392	0.222019	0.26
Visit x therapy	HD	3	β_{11}	0.238306	0.209800	0.26
Visit x therapy	HD	4	β_{12}	0.132406	0.176525	0.45
Visit x therapy	HD	5	β_{13}	0.266403	0.168878	0.12
Visit x therapy	HD	6	β_{14}	0	0	.

Table 3: Estimated regression coefficients based on analysis of response profiles of GFR for baseline and first 5 visits

The same calculation for baseline and PD:

$$\begin{aligned}\beta_1 + \beta_2 \times 0 + \beta_3 \times 1 + \cdots + \beta_8 \times 0 + \beta_9 \times 1 \times 0 + \cdots + \beta_{14} \times 0 \times 0 \\ = 2.475061 + 3.428899 \\ = 5.90396.\end{aligned}$$

Visit 6 and HD:

$$\begin{aligned}\beta_1 + \beta_2 \times 1 + \beta_3 \times 0 + \cdots + \beta_8 \times 1 + \beta_9 \times 0 \times 1 + \cdots + \beta_{14} \times 1 \times 1 = \\ = 2.475061 - 1.004013 \\ = 1.471048.\end{aligned}$$

Visit 6 and PD:

$$\begin{aligned}\beta_1 + \beta_2 \times 0 + \beta_3 \times 0 + \cdots + \beta_8 \times 1 + \beta_9 \times 0 \times 0 + \cdots + \beta_{14} \times 0 \times 1 \\ = 2.475061\end{aligned}$$

It is possible to let SPSS do this job ("estimated marginal means"); the results are displayed in table 9. Figure 6 compares the estimates with the observed means. We notice that patients undergoing PD therapy (chronic) have a consistently higher value for the mean GFR than HD patients. This can be readily explained if we realize that PD therapy is usually given at an earlier stage, which by and large coincides with the patients being younger and healthier.

How do we explain the differences between the observed and the estimated means? Our model allows negative values for the filtration rate, whereas the observed values are non-negative by definition. This will result in a lower value for the estimated mean GFR across time. Next, we test the estimates for effects of group and time and for group×time interaction. It is obvious from figure 6 that the null-hypotheses of no group and/or no time effects have to be rejected; this is also born out by table 5. Figure 6 is inconclusive with respect to group×time interaction; indeed, according to table 4 the null hypothesis of no group×time interaction cannot be rejected. Table 6 lists the REML estimated (unstructured) covariance matrix for GFR,

Parameter	Numerator df	Denominator df	F	p
Visit	5	738.666	171.367	< 0.0001
Therapy	1	1128.244	30.914	< 0.0001
Therapy×visit	5	738.666	0.743	> 0.50

Table 4: Test of group×time interaction, based on the analysis of response profiles of GFR.

Parameter	Numerator df	Denominator df	F	p
Visit	5	744.719	173.241	< 0.0001
Therapy	1	1259.516	30.792	< 0.0001

Table 5: Test of main effects based on the analysis of response profiles for GFR.

which we assume to be equal for all patients (see section 3.2). The covariance between two measurements of the filtration rate decreases with time, as is to be expected with longitudinal data ([1], page 115).

The decreasing variance of the estimated mean is less easy to explain; one would expect greater variability in the data, due to missing data and drop out. The analysis of response profiles looks reasonably linear, suggesting the viability of a linear trend model.

4.2 Modeling GFR with linear trend over time

Linear trend over time is an example of parametric regression. This is not to be confused with *linear regression*, where we track a single patient and where no covariance is assumed between repeated measures. In SPSS we choose mixed models and select time as covariate.

We write the model as

$$\mu_{ij} = \mathbb{E}(Y_{ij}) = \alpha + \beta \texttt{time}_j + \gamma \texttt{group}_i + \delta \texttt{time}_j \times \texttt{group}_i \tag{16}$$

	Visit 1	Visit 2	Visit 3	Visit 4	Visit 5	Visit 6
Visit 1 (= baseline)	12.376362	4.894658	4.629420	3.984655	4.211130	4.020289
Visit 2	4.894658	9.356327	6.047510	5.227131	5.022917	4.243854
Visit 3	4.629420	6.047510	9.529696	6.120570	5.600444	5.336974
Visit 4	3.984655	5.227131	6.120570	8.342014	6.783093	6.385844
Visit 5	4.211130	5.022917	5.600444	6.783093	8.407177	6.714562
Visit 6	4.020289	4.243854	5.336974	6.385844	6.714562	7.985493

Table 6: Analysis of response profiles: REML estimate of the covariance matrix Σ of GFR, for baseline and first 5 visits.

where $\texttt{group}_i = 1$ if the i^{th} patient was assigned to HD and $\texttt{group}_i = 0$ if assigned to PD. $\texttt{time}_j$ denotes the j^{th} visit.

Table 7 lists the regression parameters as calculated by SPSS. In table 8,

Parameter	Therapy	Estimate
Intercept (α)		5.038446
Time (β)		-0.115565
Group (γ)	HD	-0.692984
Time x group (δ)	HD	-0.011129

Table 7: Estimated regression coefficients for GFRwith time as a continuous parameter.

the REML estimate of the covariance matrix, based on linear regression, is diplayed. The values in the matrix differ from those in table 6, but overall the same pattern for the variance and covariance can be seen. The estimated

Time (month)	0	3	6	12	18	24
0	12.849629	4.402299	3.959676	3.238345	4.009179	4.234776
3	4.402299	9.398535	6.080113	5.231483	5.035871	4.345453
6	3.959676	6.080113	9.622230	6.155329	5.631603	5.451241
12	3.238345	5.231483	6.155329	8.330531	6.787102	6.499060
18	4.009179	5.035871	5.631603	6.787102	8.473863	6.997186
24	4.234776	4.345453	5.451241	6.499060	6.997186	8.487421

Table 8: Linear trend model: REML estimate of the covariance matrix Σ for GFR, for baseline and first 5 visits

means can now be calculated, in a manner similar to the calculations for the analysis of response profiles. The results are shown in table 9. Figure 7 compares the estimates with the observed values.

	Resp profiles		Linear trend	
Visit (Month)	HD	PD	HD	PD
1 (0)	5.276455	5.903960	4.345476	5.038460
2 (3)	3.725404	4.481025	3.965394	4.691765
3 (6)	3.157205	3.922911	3.585312	4.345070
4 (12)	2.459885	3.331491	2.825148	3.651680
5 (18)	2.043392	2.781002	2.064984	2.958290
6 (24)	1.471048	2.475061	1.304820	2.264900

Table 9: Estimated mean GFR.

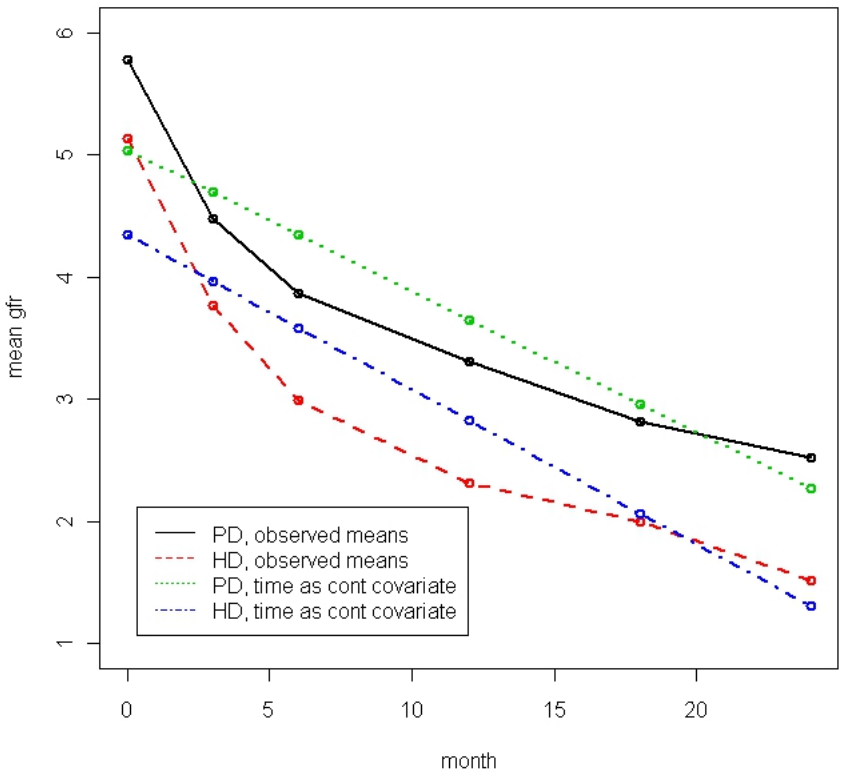

Figure 7: Observed means and the linear trend model.

Parameter	Numerator df	Denominator df	F	p
time	1	641.297	686.220	< 0.0001
Therapy	1	1293.032	20.370	< 0.001
Therapy×time	1	641.297	1.448	> 0.20

Table 10: Test of group×time interaction, based on the linear trend model for GFR.

As observed above (section 4.1), PD patients have a consistently higher value for the estimated mean GFR. And just as with the analysis of response profiles, figure 7 is inconclusive with respect to group×time interaction; indeed, according to table 10 the null hypothesis of no group×time interaction cannot be rejected. Figure 8 and table 9 compare the analysis of response profiles with linear trend.

4.3 Goodness of fit: comparing the models for response profiles and linear trend

Figures 6 and 7 give visual clues of how well both modeling approaches fit the data. But it is difficult to draw conclusions this way.

More is to be expected from a likelihood ratio test. If we denote the maximum likelihood for the analysis of response profiles and linear trend by respectively $\texttt{lik}_{resp}$ and $\texttt{lik}_{lin}$, then we know from theory ([1], page 97) that

$$-2\log\frac{\texttt{lik}_{resp}}{\texttt{lik}_{lin}} = -2\log\texttt{lik}_{resp} - (-2\log\texttt{lik}_{lin}) \sim \chi^2_{df} \tag{17}$$

with *df* := degrees of freedom, the difference in number of parameters between the models.

The maximum likelihood can be found in the output generated by SPSS, under the header "Information Criteria", but for a proper analysis we have to select Maximum Likelihood (ML) as Method in the Estimation panel for Linear Mixed Models, rather than REML which should be used if the two models have the same fixed parts and differ only in their random part(s); the random part in our model being the covariance matrix. The likelihood obtained using ML gives the *overall* likelihood, which is what we require here and which we use instead.

$$-2\log\texttt{lik}_{resp} - (-2\log\texttt{lik}_{lin}) = 23695.96 - 23852.857 = -156.897$$

The number of parameters for the profile model is 14 (table 3), but two regression parameters are zero and do not contribute to the dimensionality of the predicted hyperplane, so 12 degrees of freedom remain. From table 7 we have 4 parameters for the linear trend model. With quantile 156.897,

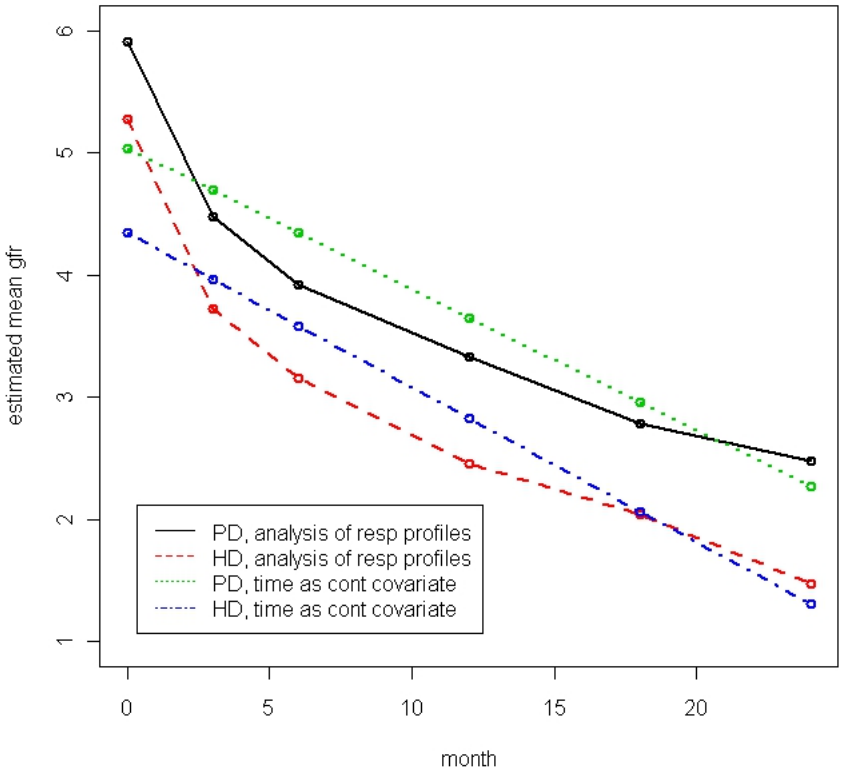

Figure 8: Estimated means: comparing the analysis of response profiles with the linear trend model.

the p-value for a χ^2 distribution with $12 - 4 = 8$ degrees of freedom is $p = 8 \times 10^{-13}$. This shows that the profile model fits the data significantly better than the linear trend model.

4.4 Limitations of response profiles and parametric curves

The two approaches to modeling the GFR have some obvious drawbacks:

- The epidemiologists suppress the zeros for the GFR in the anuric phase and then assume that the data are missing at random. Obviously they are not: when suppressing the data intentionally, they are certainly not missing at random! This has ramifications for the quality of the maximum likelihood estimates.
- Usually the observed and calculated mean response will not coincide; both models allow negative extrapolated values for the filtration rate, whereas the observed values are non-negative by definition. This will result in a lower value for the estimated mean GFR across time.

The (re)introduction of "absorbing zeros" for the GFR in section 6 offers a more realistic description of the data and contributes to more accurate calculations of the mean response.

This is done in combination with a *Markov model* which is comparatively simple to implement.

5 Markov models

5.1 Introduction. Why Markov models?

In section 3 our approach was to model the response Y_i for an individual i as

$$Y_i = (Y_{i1}, \ldots, Y_{in})' \sim N(\mu_i, \Sigma) \quad i = 1, \ldots, N,$$

with $\mu_i = (\mu_{i1}, \ldots, \mu_{in})'$ the vector of means for occasions $j = 1, \ldots, n$ and Σ the covariance matrix which is shared by all individuals i. The mean response μ_i is calculated by the analysis of response profiles or by fitting a parametric curve; see section 3.4.

The covariance is unstructered or a structure can be imposed as in section 3.5.

We now try to model an observation in terms of its values at preceding times. In the language of signal analysis, our model would be an *autoregression model of order 1*:

$$Y = AR(1) + terms$$

The advantage of this *Markov model* is the relative ease with which it can be executed.

The joint density $f(y_i)$ of Y_i can be written as

$$\begin{aligned} f(y_i) &= f(y_{i1}, \ldots, y_{in}) \\ &= f(y_{i1}, \ldots, y_{i,n-1}) f(y_{i,n}|y_{i1}, \ldots, y_{i,n-1}) \\ &= f(y_{i1}, \ldots, y_{i,n-2}) f(y_{i,n-1}|y_{i1}, \ldots, y_{i,n-2}) f(y_{i,n}|y_{i1}, \ldots, y_{i,n-1}). \end{aligned}$$

By induction

$$f(y_i) = f(y_{i1}) f(y_{i2}|y_{i1}) f(y_{i3}|y_{i2}, y_{i1}), \ldots, f(y_{i,n}|y_{i1}, \ldots, y_{i,n-1}).$$

We model

$$\begin{aligned} &f(y_{i1}) \\ &f(y_{i2}|y_{i1}) \\ &f(y_{i3}|y_{i1}, y_{i2}) \\ &\vdots \end{aligned}$$

If we assume that the Markov property holds,

$$f(y_{ij}|y_{i1}, \ldots, y_{i,j-1}) = f(y_{ij}|y_{i,j-1}),$$

then this yields for the density

$$f(y_i) = f(y_{i1}) \prod_{j=2}^{n} f(y_{ij}|y_{i,j-1}). \tag{18}$$

We try to find a simple model for $f(y_{ij}|y_{i,j-1})$, that gives a good description of the data, for example a linear regression model $\mathbb{E}(Y_{ij}|Y_{i,j-1}) = \alpha + \beta Y_{i,j-1}$.

We first show that the Markov model and the analysis of response profiles for modeling the mean are equivalent in case $n = 2$.

5.2 Comparing the analysis of response profiles and the Markov model for two time points

We consider repeated measurements on one subject, with response vector $Y = (Y_1, Y_2)'$, Y_1 and Y_2 being the responses at times $j = 1$ and $j = 2$. The response Y is bivariate normally distributed,

$$Y \sim N(\mu, \Sigma),$$

with $Y = (Y_1, Y_2)'$, $\mu = (\mu_1, \mu_2)'$, $\mu_1 = \mathbb{E}Y_1$, $\mu_2 = \mathbb{E}Y_2$ and $\Sigma = Cov((Y_1, Y_2)')$.

5.2.1 Mean and covariance of bivariate normal Y

The density is:

$$\begin{aligned} f(y) &= f(y_1, y_2) \\ &= \frac{1}{(2\pi)\sqrt{\det\Sigma}} e^{-\frac{1}{2}(y-\mu)'\Sigma^{-1}(y-\mu)}. \end{aligned}$$

The covariance is

$$\Sigma = \begin{pmatrix} \sigma_1^2 & Cov(Y_1, Y_2) \\ Cov(Y_1, Y_2) & \sigma_2^2 \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}. \tag{19}$$

5.2.2 The Markov model

The density can be written as

$$f(y) = f(y_1, y_2) = f(y_1) f(y_2|y_1). \tag{20}$$

We assume that

$$Y_1 \sim N(\mu_1, \sigma_1^2) \tag{21}$$

$$Y_2|Y_1 \sim N(\alpha + \beta Y_1, \sigma_\epsilon^2). \tag{22}$$

The expression for the distribution of $Y_2|Y_1$ is motivated by [4], Example B page 136 and Example A, page 140:

$$\mathbb{E}(Y_2|Y_1) = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(Y_1 - \mu_1), \tag{23}$$

with μ_1 and μ_2 as above. This suggests the following model for $Y_2|Y_1$:

$$Y_2|Y_1 = \alpha + \beta Y_1 + \epsilon, \quad \text{in which} \quad \epsilon \sim N(0, \sigma_\epsilon^2).$$

Note that the noise term ϵ is mandatory, otherwise Y_2 would be "hard wired" to Y_1. σ_ϵ^2 is unknown.

The mean μ_2 is

$$\begin{aligned} \mu_2 = \mathbb{E}Y_2 &= \mathbb{E}(\mathbb{E}(Y_2|Y_1)) \\ &= \mathbb{E}(\alpha + \beta Y_1) \\ &= \alpha + \beta\mu_1. \end{aligned} \tag{24}$$

Let us calculate Σ in terms of σ_1, β and σ_ϵ.

$$Cov(Y_1, Y_2) = \mathbb{E}Y_1Y_2 - \mathbb{E}Y_1\mathbb{E}Y_2 = \mathbb{E}Y_1Y_2 - \mu_1\mu_2$$

$$Y_2Y_1 = (\alpha + \beta Y_1 + \epsilon)Y_1 = \alpha Y_1 + \beta Y_1^2 + \epsilon Y_1$$

$$\begin{aligned}\Rightarrow EY_2Y_1 &= \alpha\mu_1 + \beta\mathbb{E}Y_1^2 + \mathbb{E}\epsilon\mathbb{E}Y_1 \\ &= \alpha\mu_1 + \beta(VarY_1 + (\mathbb{E}Y_1)^2) \\ &= \alpha\mu_1 + \beta\sigma_1^2 + \beta\mu_1^2.\end{aligned}$$

Combining, we obtain

$$\begin{aligned}Cov(Y_1, Y_2) &= \mathbb{E}Y_1Y_2 - \mu_1\mu_2 \\ &= \alpha\mu_1 + \beta\sigma_1^2 + \beta\mu_1^2 - \mu_1(\alpha + \beta\mu_1) \\ &= \beta\sigma_1^2.\end{aligned}$$

We still have to calculate σ_2^2.

$$\sigma_2^2 = VarY_2 = Var(\alpha + \beta Y_1 + \epsilon) = \beta^2\sigma_1^2 + \sigma_\epsilon^2.$$

The covariance Σ is:

$$\Sigma = \begin{pmatrix} \sigma_1^2 & Cov(Y_1, Y_2) \\ Cov(Y_1, Y_2) & \sigma_2^2 \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \beta\sigma_1^2 \\ \beta\sigma_1^2 & \beta^2\sigma_1^2 + \sigma_\epsilon^2. \end{pmatrix} \tag{25}$$

In the following section we apply both methods to the GFR data.

5.2.3 Applying the models to the GFR data and comparing the results for two time points

We do not take the effect of therapy (HD of PD) into account.

The analysis of response profiles The model for the mean response is similar to (15), however, we are not interested in the (main) effect of group (therapy), nor in group×time interaction :

$$\mu_j = \tilde{\alpha} + \tilde{\beta}[j = 1], \quad j = 1, 2.$$

The SPSS estimates of the coefficients can be found in the first column of table 11. Note that before we convert the data set to long format, we first select only those patients which have non-missing data for both the first two visits (base line counts as the first visit, as usual). Only then we can expect the results for the estimated means to agree.

SPSS estimates (REML) the covariance matrix as

$$\widehat{\Sigma} = \begin{pmatrix} 12.604898 & 5.051108 \\ 5.051108 & 9.713908 \end{pmatrix}. \tag{26}$$

The Markov model The parameters estimated by SPSS can be found in table 11, second column; the results show that there is good (overall) agreement between the quantities.

Resp profiles		Markov approach	
$\hat{\mu}_1$ (*)	5.430272	$\hat{\mu}_1$	5.430272
$\hat{\mu}_2$ (*)	3.832424	$\hat{\mu}_2$	3.832425 (24)
		$\hat{\alpha}$	1.656374
		$\hat{\beta}$	0.400726
$\hat{\sigma}_1$	3.550338 (19)	$\hat{\sigma}_1$	3.550338
$\hat{\sigma}_2$	3.116714 (19)	$\hat{\sigma}_2$	3.118147 (25)
$\hat{\sigma}_\epsilon$	—	$\hat{\sigma}_\epsilon$	2.774658
$\widehat{Cov}(Y_1, Y_2)$	5.051108	$\widehat{Cov}(Y_1, Y_2)$	5.051111 (25)
$\hat{\rho}$	0.4564782 (19)	$\hat{\rho}$	0.4562686 (25)

(19) formula used. (*) $\hat{\mu}_2 = \hat{\hat{\alpha}}, \quad \hat{\mu}_1 - \hat{\mu}_2 = \hat{\hat{\beta}}$

Table 11: Estimates for the bivariate distribution $Y = (Y_1, Y_2)$: comparing response profiles and the Markov model.

6 Modeling GFR over time using Markov models in which GFR is censored

We found that for two time points, repeated measures (under the guise of response profiles) and Markov models are equivalent.

In the following we will extent Markov to more than two time points. An extra complication is, that the GFR can be not smaller than zero and remains zero when a patient becomes anuric.

6.1 Principles of latency and censoring

When modeling the filtration rate over time, we presume that the values are multivariate normally distributed. The GFR is by definition definite non-negative. When modeling the GFR in time with regression models which require negative values for GFR, we therefore presume a *latent variable*, normally distributed.

When renal function breaks down, GFR obtains the value 0 and enters an absorbing state, as discussed in the Introduction to this thesis. Over time, the number of patients whose GFR enter this absorbing state, increases, and so the frequency of the value 0 increases; see figure 9. The cdf of the observed values Y for the GFR is therefore continuous, but mass accumulates in the point $Y = 0$ and for $Y > 0$, Y is normally distributed. *Censoring* occurs when exact values for a random variable Y are known only outside certain intervals (see [5], chapter 3). The filtration rate GFR is the observed result of a left censored random variable. We observe $Y = \max(Z, 0)$ in which Z

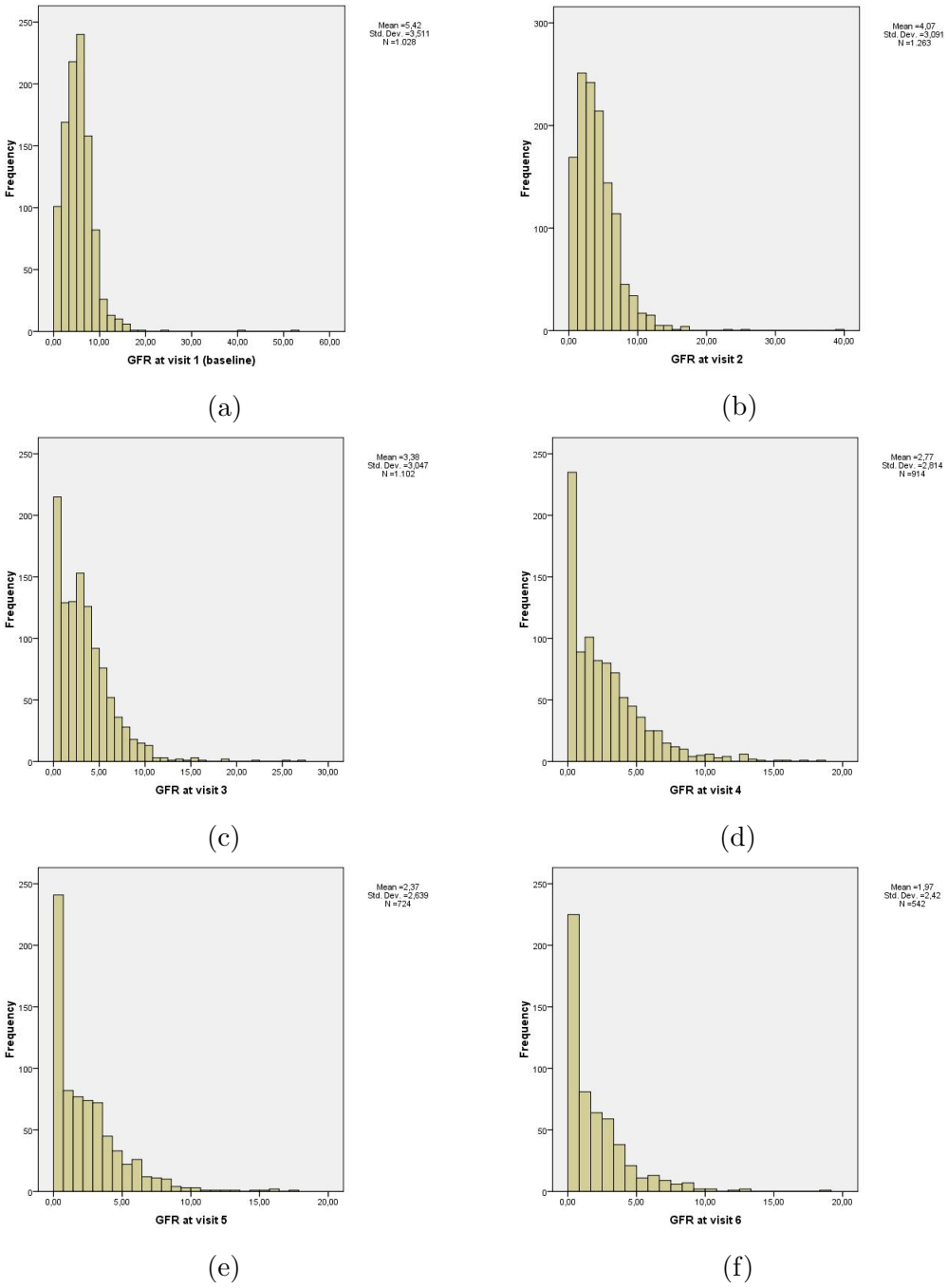

Figure 9: Frequency distributions of GFR at (a) visit 1 (= baseline), (b) visit 2, (c) visit 3, (d) visit 4, (e) visit 5, (f) visit 6. Data after cleaning up according to section 2.1.

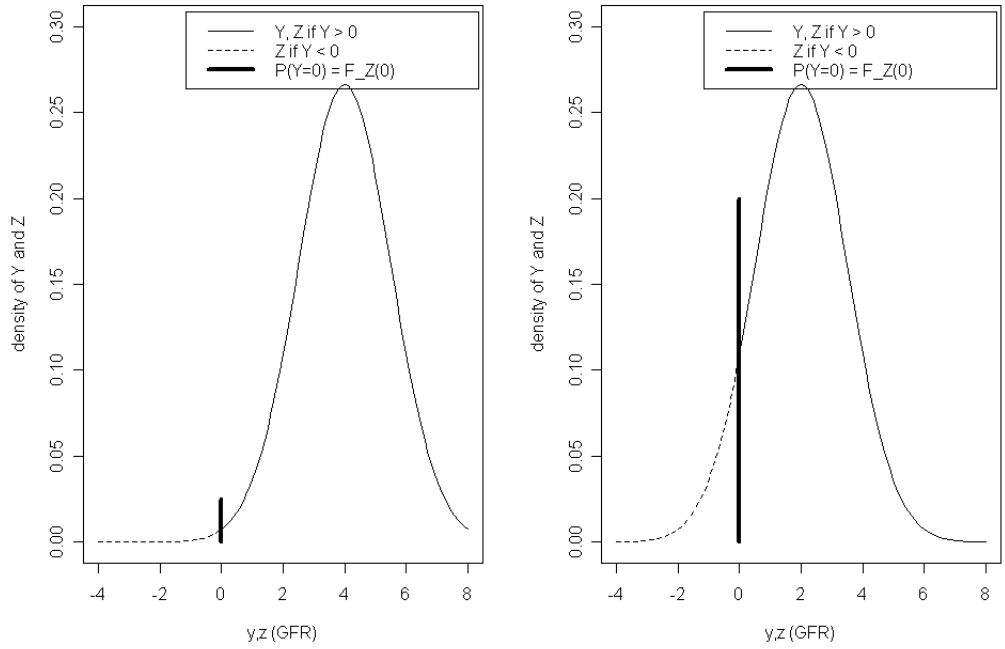

Figure 10: Illustrating latency and censoring of GFR (Y,Z). Left: densities of Y (observed GFR) and Z (latent to Y) at early visit of patients. Right: As left, but at later visit. The vertical bar measures the censored GFR.

is the latent variable:

$$Z > 0 \Rightarrow Y = Z \tag{27}$$

$$Z \leq 0 \Rightarrow Y = 0. \tag{28}$$

Figure 10 illustrates the concepts of censoring and latent variables: the vertical bar is a measure of the censored "mass" of the random variable Y.

6.1.1 The Method of Maximum Likelihood for the Markov model with censoring

We continue the discussion of section 5.1 and first consider the case without censoring.

There are N individuals labeled i, $i = 1, \ldots, N$ and n occasions j, $j = 1, \ldots, n$. Let Y_i be the response vector of subject i; $Y_i = (Y_{i1}, \ldots, Y_{in})$. If the measurements of different individuals are independent and $f(Y_i) =$

$f(Y_{i1}, \ldots, Y_{in})$ is the density, then the likelihood is

$$\texttt{lik}(\theta) = \prod_{i=1}^{N} f(Y_i; \theta) \tag{29}$$

$$= \prod_{i=1}^{N} \Big[f(Y_{i1}) \prod_{j=2}^{n} f(Y_{ij} | Y_{i,j-1}; \theta) \Big], \tag{30}$$

under the Markov assumption; see (18).

Now let Y be censored, with latent variable Z. Then the factors $f(Y_{ij}|Y_{i,j-1};\theta)$ in (29) assume the following values, depending on the "state" of $Y_{i,j-1}$ and Y_{ij}:

$$f(Y_{ij}|Y_{i,j-1};\theta) = \begin{cases} f_Z(Y_{ij}|Y_{i,j-1};\theta) & \text{if } Y_{ij} > 0 \text{ and } Y_{i,j-1} > 0, \\ F_Z(0|Y_{i,j-1};\theta) & \text{if } Y_{ij} = 0 \text{ and } Y_{i,j-1} > 0, \\ 1 & \text{if } Y_{ij} = 0 \text{ and } Y_{i,j-1} = 0, \\ 0 & \text{if } Y_{ij} > 0 \text{ and } Y_{i,j-1} = 0, \end{cases} \tag{31}$$

Diagram 11 explains the states in our Markov chain.

(31) yields as likelihood

(32)

$$\texttt{lik}(\theta) = \prod_{i=1}^{N} \Big[f_Z(Y_{i1})^{\delta_{i1}} (F_Z(0))^{1-\delta_{i1}} \prod_{j=2}^{n} f_Z(Y_{ij}|Y_{i,j-1};\theta)^{\delta_{ij}} F_Z(0|Y_{i,j-1};\theta)^{1-\delta_{ij}} \Big].$$

in which

$$\delta_{i1} = \begin{cases} 1 & \text{if } Y_{i1} > 0, \\ 0 & \text{if } Y_{i1} = 0. \end{cases} \tag{33}$$

and

$$\delta_{ij} = \begin{cases} 1 & \text{if } Y_{ij} > 0 \text{ and } Y_{i,j-1} > 0, \\ 0 & \text{if } Y_{ij} = 0 \text{ and } Y_{i,j-1} > 0. \end{cases} \tag{34}$$

The *log likelihood* L is

$$L(\theta) = \log \texttt{lik}(\theta), \tag{35}$$

We will use the likelihood to find estimates for the regression coefficients in the Markov model

$$Z_j | Z_{j-1} = \alpha + \beta Z_{j-1} + \epsilon, \quad j = 2, \ldots, 6; \quad \epsilon \sim N(0, \sigma_\epsilon^2). \tag{36}$$

Here $\theta = (\alpha, \beta, \sigma_\epsilon)$.

To obtain the maximum likelihoods estimates, this likelihood has to be optimized.

We start assuming that α, β and σ_ϵ are equal for all visits. Later we will examine if this is the case in our data set.

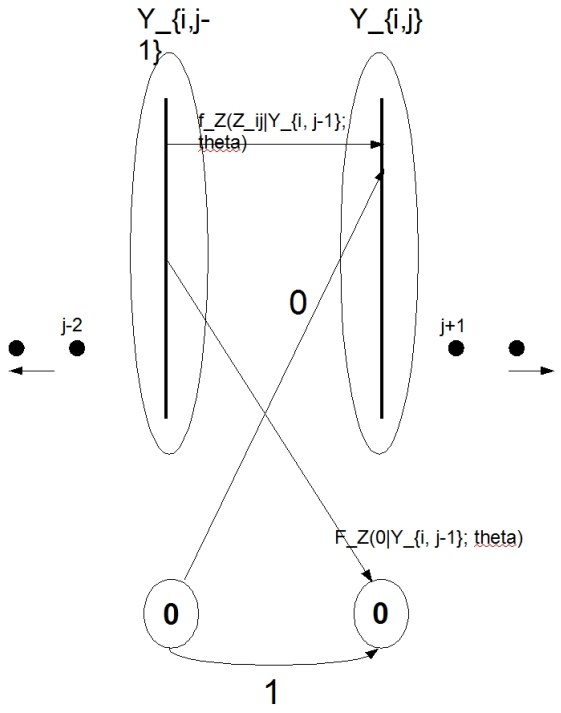

Figure 11: Diagram of transition densities

6.1.2 Estimating the regression coefficients in the Markov model

So far we used SPSS to perform the calculations, but this package is not suited for the kind of analysis we will now embark on. Therefore, we resort to the statistical package R. The script file can be found at the end of this document.

In order to find the regression parameters in our model, we can optimize the likelihood (32) using a numerical optimization procedure like for example the R function `optim`. However our problem with censored observations is similar to survival analysis with censored survival times. Therefore we can also use routines for parametric survival analysis like the R function `survreg`. Indeed, we performed several experiments with both routines and concluded that they give the same results. However, working with `survreg` has several advantages: we do not need carefully chosen initial values as with `optim`. Also, `survreg` estimates standard errors and returns a lot of diagnostic data, which are difficult to extract when optimizing with `optim`.

patient nr. (id_gw)	group (therap0)	time/visit (Index1)	Y_2 (gfr_cor)	Y_1 (lag_gfr)
1	2	1	0	.
1	2	2	0.943796	0
1	2	3	1.244697	0.943796
1	2	4	0.895639	1.244697
1	2	5	0	0.895639
2	2	1	1.888399	.
3	1	1	2.168828	.
3	1	2	2.171048	2.168828
3	1	3	1.562232	2.171048
3	1	4	0	1.562232

Table 12: Start of data set, after transposing to long format. (Variable names in parentheses.)

We will use the following basic format for `survreg`:

(37) `survreg(formula, dist="gaussian")`

On behalf of this analysis, the data set has to be transposed to long format, an operation we already performed in section 2.1. Thus, the data set has the format shown in table 12, in which entries for the first 3 patients and the first 6 visits are included: When performing linear regression *without* censoring to find the coefficients in (36), we use the R function `lm` and then `formula` would assume the basic form `Y_2` $\sim$ `Y_1`. *With* (right) censoring, `formula` assumes the more complicated command

(38) `Surv(Y_2, event)` $\sim$ `Y_1`

The response variable in `formula` is a *survival object* created by the R function `Surv`. In the language of survival analysis, the first argument to `Surv` is the *follow up time*, which in our study, with Y_2 and Y_1 being the actual measured value of GFR, is the positive value of Y_2. `event` is the *status indicator*, a binary vector: 0 if the subject is "alive" and 1 if "dead". In our study, 1 codes for $Y_2 > 0$ and 0 otherwise.

In survival analysis, variables can be right censored. We have left censored data. We accommodate left censored data by adding the switch `type="left"` as argument to the function `Surv`.

The second argument to `survreg` is the distribution of the survival time, in our study this is the distribution of the latent variable underlying GFR, which we assume to be normal. `survreg` then, returns information about the latent variable; with the coefficients returned, we can calculate $\mathbb{E}(Z_j)$ (given the value of Z_{j-1}).

6.1.3 Markov model with fixed regression coefficients: results over the first 6 visits.

Model without group variable We first consider the model without dividing the patients in therapy groups. This model has one intercept and one slope for all patients and is model 1 in table 15. If we execute command (37), the `summary` prints as:

```
> event<-as.numeric(Y_2 > 0)
> diagn <-survreg(Surv(Y_2, event, type="left")~Y_1, dist="gaussian")
> summary(diagn)

Call:
survreg(formula = Surv(Y_2, event, type = "left") ~ Y_1, dist = "gaussian")
             Value Std. Error     z        p
(Intercept) 0.505     0.0760  6.64 3.09e-11
Y_1         0.614     0.0143 42.96 0.00e+00
Log(scale)  0.962     0.0128 75.05 0.00e+00

Scale= 2.62

Gaussian distribution
Loglik(model)= -8131.4   Loglik(intercept only)= -8882.4
        Chisq= 1501.97 on 1 degrees of freedom, p= 0
Number of Newton-Raphson Iterations: 4
n= 3703
```

The output shows that Y_1 significantly contributes to Y_2.

Model for HD patients We repeat the analysis, but now select only patients with HD therapy:

```
> event<-as.numeric(Y_2 > 0)
> diagn <-survreg(Surv(Y_2, event, type="left")~Y_1, dist="gaussian")
> summary(diagn)

Call:
survreg(formula = Surv(Y_2, event, type = "left") ~ Y_1, dist = "gaussian")
             Value Std. Error     z         p
(Intercept) 0.406     0.1008  4.03  5.69e-05
Y_1         0.577     0.0198 29.16 6.12e-187
Log(scale)  0.954     0.0182 52.54  0.00e+00

Scale= 2.60

Gaussian distribution
Loglik(model)= -4155.9   Loglik(intercept only)= -4509.5
        Chisq= 707.27 on 1 degrees of freedom, p= 0
Number of Newton-Raphson Iterations: 4
n= 1945
```

As with the overall model, Y_1 significantly contributes to Y_2.

Model for PD patients

```
> event<-as.numeric(Y_2 > 0)
> diagn <-survreg(Surv(Y_2, event, type="left")~Y_1, dist="gaussian")
> summary(diagn)

Call:
survreg(formula = Surv(Y_2, event, type = "left") ~ Y_1, dist = "gaussian")
             Value Std. Error     z         p
(Intercept) 0.674     0.1144  5.89  3.88e-09
Y_1         0.637     0.0206 30.87 3.06e-209
Log(scale)  0.958     0.0181 53.00  0.00e+00

Scale= 2.61

Gaussian distribution
Loglik(model)= -3955.2   Loglik(intercept only)= -4336.8
        Chisq= 763.2 on 1 degrees of freedom, p= 0
Number of Newton-Raphson Iterations: 4
n= 1758

>
```

The output shows that Y_1 significantly contributes to Y_2.

Table 13 summarizes the output for the three models. Note the difference

Parameter	Overall estimate	Estimate for HD	Estimate for PD
α	0.505	0.406	0.674
β	0.614	0.577	0.637
σ_ϵ	2.62	2.60	2.61
log lik	-8131.4	-4155.9	-3955.2

Table 13: Markov model with fixed regression coefficients and censored GFR; first 6 visits

between the log likelihood for the overall estimate (-8131.4) and the sum for HD and PD $-4155.9 - 3955.2 = -8111.1$. The discrepancy can be explained if we realize that the overall model estimates a combined covariance matrix and combined effects for HD and PD and that the models for each group estimate a separate covariance matrix and separate effects.

Instead of separate analyses for the two groups, we can also differentiate between groups (therapies) in one overall analysis (model 2 in table 15). Therefore we extend model (36) with an interaction of (chronic) therapy

(HD, PD) with GFR, and modify the R command accordingly:

(39) `diagn <- survreg(Surv(Y_2, event, type="left")`
`~ Y_1*as.factor(therap0), dist="gaussian"),`

with `summary`:

```
> summary(diagn)

Call:
survreg(formula = Surv(Y_2, event, type = "left") ~ Y_1 * as.factor(therap0),
    dist = "gaussian")
                          Value Std. Error     z         p
(Intercept)              0.4047     0.1007  4.02  5.82e-05
Y_1                      0.5774     0.0198 29.14 1.06e-186
as.factor(therap0)2      0.2695     0.1519  1.77  7.60e-02
Y_1:as.factor(therap0)2 0.0592     0.0286  2.07  3.80e-02
Log(scale)               0.9561     0.0128 74.63  0.00e+00

Scale= 2.6

Gaussian distribution
Loglik(model)= -8111.1   Loglik(intercept only)= -8882.4
        Chisq= 1542.55 on 3 degrees of freedom, p= 0
Number of Newton-Raphson Iterations: 4
n= 3703

>
```

Note that the interaction between therapy and lag GFR is significant.

Let us verify if we can reconstruct the estimated regression coefficients for the aggregate model (36) separately, from the model with interaction.

First we verify the estimates in column HD (`therap0 = 1`):

$$\begin{aligned} Z_j &= 0.4047 + 0.5774 Z_{j-1} + 0 + 0 Z_{j-1} \\ &= 0.4047 + 0.5774 Z_{j-1}. \end{aligned}$$

For PD (`therap0 = 2`):

$$\begin{aligned} Z_j &= 0.4047 + 0.5774 Z_{j-1} + 0.2695 + 0.0592 Z_{j-1} \\ &= 0.6742 + 0.6366 Z_{j-1}. \end{aligned}$$

There is good agreement with the values in table 13. With respect to GFR PD patients start higher and also decrease slower with time.

6.1.4 The Markov model with censoring and separate intercept for each 6 time points

So far we have assumed one intercept and one slope for all visits. We will now check if this assumption is not too strong (model 3 in table 15).

The R command (39) is modified to accommodate a separate intercept for each visit:

(40)
```
diagn <- survreg(Surv(Y_2, event, type="left")
          ~ Y_1*as.factor(therap0)+ time, dist="gaussian"),
```

in which `time` represents the visit number and is treated as a continuous parameter (linear trend model).

```
> summary(diagn)

Call:
survreg(formula = Surv(Y_2, event, type = "left") ~ Y_1 * as.factor(therap0) +
    time, dist = "gaussian")
                           Value Std. Error     z         p
(Intercept)               0.5643     0.1701  3.32  9.08e-04
Y_1                       0.5736     0.0201 28.57 1.39e-179
as.factor(therap0)2       0.2751     0.1520  1.81  7.02e-02
time                     -0.0402     0.0345 -1.16  2.45e-01
Y_1:as.factor(therap0)2   0.0591     0.0286  2.07  3.85e-02
Log(scale)                0.9559     0.0128 74.62  0.00e+00

Scale= 2.6

Gaussian distribution
Loglik(model)= -8110.5   Loglik(intercept only)= -8882.4
        Chisq= 1543.9 on 4 degrees of freedom, p= 0
Number of Newton-Raphson Iterations: 4
n= 3703

>
```

From the output we may conclude that the main effect for time is not significant.

We combine the coefficients for effects and interaction for HD (`therap0 = 1`):

$$\begin{aligned} Z_j &= 0.5643 + 0.5736 Z_{j-1} + 0 + 0 Z_{j-1} \\ &= 0.5643 + 0.5736 Z_{j-1}. \end{aligned}$$

And for PD (`therap0 = 2`):

$$\begin{aligned} Z_j &= 0.5643 + 0.5736 Z_{j-1} + 0.2751 + 0.0591 Z_{j-1} \\ &= 0.8394 + 0.6327 Z_{j-1}. \end{aligned}$$

Table 14 summarizes the results for time dependent intercept. As stated above, the contributions of time to the intercept (γ) are not significant. This suggests that models without time varying intercept offer an adequate description.

Parameter	Estimate for HD	Estimate for PD
α	0.5643	0.8394
β	0.5736	0.6327
γ	-0.0402 ($p > 0.20$)	-0.0402 ($p > 0.20$)

Table 14: As table 13 but with time dependent intercept.

Model	-2 log likelihood	# regression parameters
1. One intercept, one slope (36), (38)	16262.8	2
2. Interaction therapy*Y_1 (39)	16222.2	4
3. Idem, with time as linear covariate (40)	16221	5

(36) Equation

Table 15: Likelihoods for Markov models with censoring

6.1.5 Which model fits the data best?

Table 15 lists the likelihoods of the models we encountered so far. We conduct an analysis of the models along the line of section 4.3, using equation (17). Let us compare models 3 and 2 in table. With quantile $16222.2 - 16221 = 1.2$, the p-value for a χ^2 distribution with $5 - 4 = 1$ degrees of freedom is $p = 0.27$, far from significant.

Comparing models 3 and 1, we find as quantile: $16262.8 - 16221 = 41.8$. For a χ^2_3-distribution, this is highly significant, so we reject model 1.

Therefore *we adopt model 2*, equation (39). It shows that with respect to GFR PD patients start higher and also decrease slower with time. This is also reflected by figure 3.

7 Evaluating the performance of the selected model

In section 6.1.5 we arrived at a model which we think best describes the mean GFR. How does it compare with the *observed* means shown in figure 3?

Because model (36) assumes that the GFR for each visit has a normal distribution — which it has not, see section 2.1—, the values for the expectations thus calculated for visits $j = 2, \ldots, 6$, come out to low. This is not the case though for the expectation at baseline ($j = 1$): if $\mathbb{E}Y_1$ is the expectation of the observed GFR at baseline and if $\mathbb{E}Z_1$ is the expectation of the (normal) variable Z_1 latent to Y_1 (see figure 10), then we may equate $\mathbb{E}Z_1 = \mathbb{E}Y_1$, because patients who are already anuric at baseline were removed from the data set.

For visit $j = 2$ the $\mathbb{E}Y_2$ is calculated as follows (the data base at this point is in long format as described in section 2.1). Equation (44) (appendix B)

calculates the expectation for the truncation Y of the latent normal distribution $Z \sim N(\mu, \sigma)$. To obtain $\mathbb{E}Y_2$ we substitute in (44):

$$\begin{aligned}\mu &= \mathbb{E}(Z_2|Z_1) = \alpha + \beta Z_1\\ \sigma &= \text{the standard deviation returned by the output of model (39)}\\ \mathbb{E}Y &= \mathbb{E}(Y_2|Y_1),\end{aligned}$$

in which α, β and σ depend on the therapy (HD or PD) to which the patient is subjected. We then have

$$(41)\quad \mathbb{E}(Y_2|Y_1) = \mathbb{E}(Y_2|Z_1) = \frac{\sigma}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\alpha+\beta Z_1}{\sigma}\right)^2} + (\alpha + \beta Z_1)(1 - F_{Z_2|Z_1}(0)),$$

in which $F_{Z_2|Z_1}$ is the cumulative distribution function of $Z_2|Z_1 \sim N(\alpha + \beta Z_1, \sigma^2)$. From this

$$(42)\quad \mathbb{E}Y_2 = \mathbb{E}(\mathbb{E}(Y_2|Y_1)) = \mathbb{E}\left(\frac{\sigma}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\alpha+\beta Z_1}{\sigma}\right)^2}\right) + (\alpha + \beta \mathbb{E}Z_1)(1 - F_{Z_2}(0)).$$

Another complication arises if we want to compute $\mathbb{E}(Y_{j+1}|Y_j)$ for $j \geq 2$. Then we have to take into account that in case of kidney failure at visit j, that is, GFR $= 0$ at visit j, it remains in that state from then on. This can be expressed as:

$$(43)\qquad \begin{aligned}\mathbb{E}(Y_{j+1}|Y_j) &= \mathbb{E}(Y_{j+1}|Z_j) && \text{if} \quad Y_j > 0\\ &= 0 && \text{if} \quad Y_j = 0.\end{aligned}$$

We can however find $\mathbb{E}Y_2$ in (42), by using numerical techniques. Either by calculating the integrals with classical numerical methods or by simulation. Solving the integrals by numerical methods requires quite some effort for which we do not have time.

We can also employ simulation techniques, as follows: Draw $Z_1 = Y_1$ from a normal distribution. Then draw Z_2 from the conditional distribution $Z_2|Z_1$, determine Y_2 from $Y_2|Y_1 = Y_2|Z_1$ (see (42)). If $Y_2 = 0$, then $Y_3 = 0$. Otherwise calculate Y_3 by drawing Z_3 from $Z_3|Z_2$, etc.

If this is repeated a relevant number of times, it is possible to calculate the averages $Y_1, Y_2, Y_3, \ldots$ from the simulation. We performed this simulation by simulating a data set with 10000 patients whose GFR at baseline are normally distributed, with as parameters the observed mean GFR and observed standard deviation at baseline. This was done separately for each therapy. E.g., for HD the R code for the first 3 visits reads as follows:

```
N <- 10000
Y_1<- rnorm(N, mean=EY1_HD, sd=sdY_HD ) #Z_1=Y_1

Z_2 <- aHD + bHD*Y_1 + rnorm(N, 0, sdInter) # Z_2|Z_1 ~ a + bZ_1 + eps
```

```
Y_2 <- pmax(Z_2,0)
EY_2 <-mean(Y_2)) #E(Y_2|Y_1

Z_3 <- aHD + bHD*Z_2 + rnorm(N, 0, sdInter)
Y_3 <- ifelse(Y_2 > 0, pmax(Z_3,0)  ,0)
EY_3 <-mean(Y_3)
```

Note how conditional expectation (43) is implemented by the `ifelse` construction. `sdInter` is the standard deviation found in the diagnostics of command (39) and the regression parameters are from section 6.1.3.

The results are plotted in figure 12. One sees that the simulated GFR for PD starts higher and decreases slower with time, consistent with earlier observations.

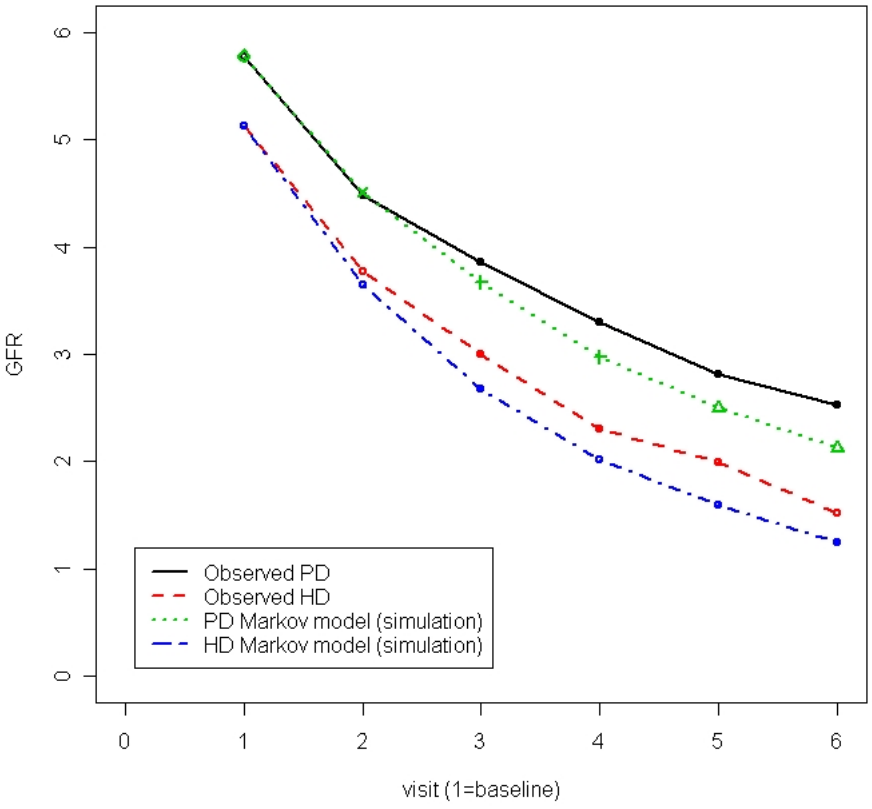

Figure 12: Observed means compared with simulated Markov model.

Can we improve on this result? A possible explanation for the difference between observed and predicted filtration rate, is drop out due to death of patients. Probably these are persons in a bad condition with GFR$= 0$. This

amounts to selective drop out which is why the observed GFR probably is somewhat to high.

This means that by replacing the missing values of the observed GFR of anuric patients with the value zero, our model corrects for missing data, as shown by figure 13.

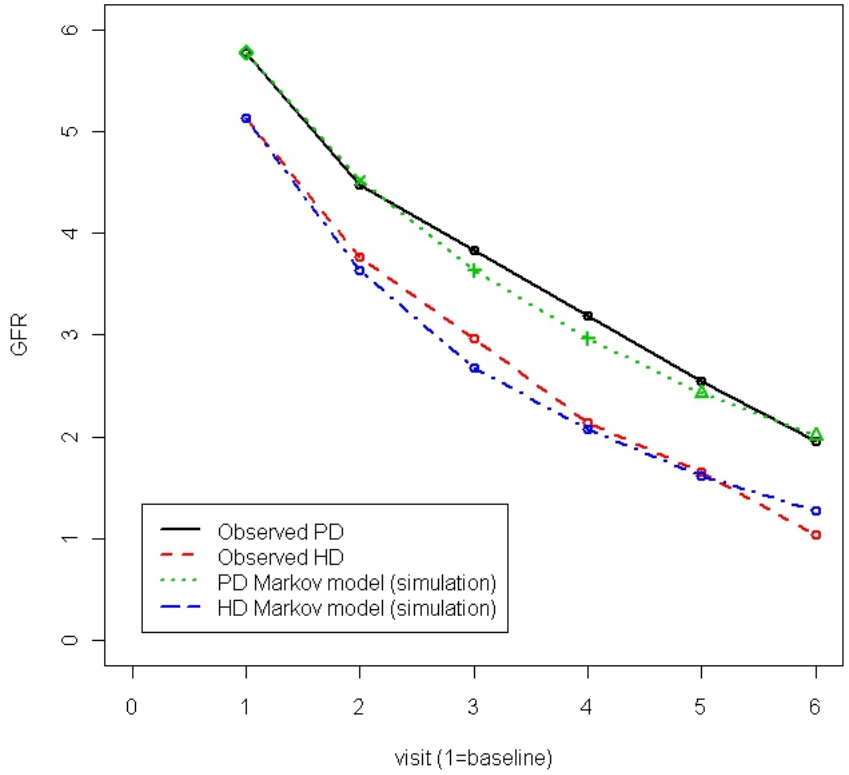

Figure 13: Observed means compared with simulated Markov model.

8 Summary and Conclusions

Patients undergoing kidney dialysis experience a deteriorating performance of their kidneys. The best index of kidney function is the Residual Glomerular Filtration Rate, GFR for short, a continuous response variable.

In this thesis we discuss several statistical techniques to model the decrease of the mean GFR over time, in particular the effect of (chronic) therapy (hemodialysis HD or peritoneal dialysis PD) on the baseline value of the GFR and the speed with which the filtration rate decreases. The models discussed

are examples of *Linear Mixed Models*, which are likelihood based and which are robust against data missing at random. This requires that we tranpose the data set (of patients) to long format.

In behalf of the statistical analysis, a patient is considered *anuric*, when GFR $= 0$ at two consecutive visits. From then on, the GFR at each visit is considered missing (at random), rather than set to zero; this allows the GFR to be extrapolated and assume negative values.

From visit 6 on, the effect of therapy on GFR becomes less unambiguous due to increased patient drop out. Therefore, we restricted our analysis to the first 6 observations (including the baseline values) to make computations easier.

We started with an *Analysis of Response Profiles*, a saturated model, in which all main effects (time and group, that is: visit and therapy) and all interactions (time*group) are considered. This is implemented by the SPSS command MIXED, with time as category and with unstructured covariance matrix. Plotting the GFR for both groups shows a reasonably linear decrease with time without a noticeable group*time interaction (figure 6). This is born out by a statistical analysis (table 4).

This approach results in many regression parameters which have to be communicated to researchers (table 3).

We considered the same saturated model, but now with time as continuous covariate, which we called *linear trend over time*. Just as with the analysis of response profiles, there is no clear group*time interaction (figure 7) and this is confirmed by table 10.

Though both models are not equivalent statistically, the linear trend model requires considerable fewer parameters to communicate to others, see table 7.

The statistical methods above have some drawbacks. The data in the anuric phase are considered missing at random. Of course they are not, considering that suppressing data can hardly be considered a random act! This has ramifications for the quality of the maximum likelihood estimates.

These drawbacks are addressed by a comparatively simple *Markov model*, in which we try to model an observation Y_{ij} for the GFR of patient i at time j linearly in terms of its value $Y_{i,j-1}$ at the preceding visit ("lagged GFR"). In behalf of this analysis, the data in the anuric phase are no longer considered missing, but are set to zero ("absorbing zeros", "censored GFR" of "censored observations"). Now, the GFR can no longer be considered normally distributed, and we have to work with a normally distributed latent variable, latent to the GFR.

The state space in this Markov "chain" consists of the values for Y_{ij} and $Y_{i,j-1}$. The two possible "states" are 0 and "positive" and are represented in equation (31). Note that we are talking not so much about transition probabilities, but rather transition *conditional expectations*.

In order to find the regression parameters in the simple linear recursive

Markov model (36), we note that the problem with censored observations is similar to survival analysis with censored survival times. We used the R package (function `survreg`) to find the regression coefficients. We investigated 3 models: with one intercept and one slope, with main effects and interaction of therapy and lag GFR and models with interaction and time as linear covariate, that is, a separate intercept for each visit. Analysis of the diagnostic data (table 15) shows that the second model, with interaction but without separate intercept, suits are our purpose well. *It shows that with respect to the* GFR PD *patients start higher and decrease slower with time.* From a clinical perspective, this can be explained if we realize that PD therapy is usually given at an earlier stage, which by and large coincides with the patients being younger and healthier.

We then compared this model for the mean GFR with the observed means. The naive approach is to plug in the observed GFR at baseline in the linear recursive model and then recursively calculate the values for the other visits. That way, the calculated filtration rate would come out to low. The correct procedure is to calculate the mean value for the truncated latent variable (see figure 10 and appendix B). We can numerically calculate the integrals involved or we can resort to simulation. We chose the last option, and simulated a data set with 10000 patients whose GFR at baseline are normally distributed.

This thesis is the fruit of an internship at the LUMC which gave me the opportunity to get acquainted with several facets of modern statistical practice: Linear Mixed Models, SPSS, programming in R etc. However, after being for years immersed of in a sea of abstract mathematical thought, the transition to a more data oriented environment took some adapting.

9 Suggestions for further research

- Study the effect of other covariates on the GFR, e.g. age or age brackets, Primary Renal Disease (see table 1), etc,
- Consider covariance pattern models other than "unstructured" (see section 3.5.2),
- Transform the data: maximum likelihood is reasonably robust, but, with time, patients leave the study and data become more and more skewed (figure 9). To increase the accuracy of the estimates, transforming the data should be considered, e.g. take the square root of the GFR,
- Consider higher order regression. In this thesis we considered only first order regression (36) when we delt with the Markov model. It might be interesting to involve earlier visits: $Y_j = \alpha + \beta_{j-1}Y_{j-1} + \beta_{j-2}Y_{j-2} + \cdots + \beta_{j-k}Y_{j-k} + \epsilon$,

- In the simulation, a fixed standard deviation is assumed. This may not be the case,
- In the state diagram, transition probabilities (see (31)) are calculated in a simple consistent way; however, this may not be the correct model.

A Some statistical background

The expression for a first order linear regression model with fixed covariate x is:

$$Y = \alpha + \beta x + \epsilon,$$

with noise term $\epsilon \sim N(0, \sigma_\epsilon^2)$. The estimated respons is

$$\hat{Y} = \hat{\alpha} + \hat{\beta} x.$$

If the covariate is a random variable $X \sim N(\mu, \sigma_X^2)$, we write

$$Y = \alpha + \beta X + \epsilon.$$

The (conditional) expectations and variances are:

$$\begin{aligned} \mathbb{E}(Y|X) &= \mathbb{E}(\alpha + \beta X + \epsilon|X) \\ &= \mathbb{E}(\alpha|X) + \mathbb{E}(\beta X|X) + \mathbb{E}(\epsilon|X) \\ &= \alpha + \beta X + 0 \\ \mathbb{E}Y &= \mathbb{E}(\mathbb{E}(Y|X)) = \alpha + \beta\mu \end{aligned}$$

$$\begin{aligned} Var(Y|X) &= 0 + \beta Var(X|X) + Var(\epsilon|X)) = \sigma_\epsilon^2 \\ VarY &= \beta^2\sigma_X^2 + \sigma_\epsilon^2 \end{aligned}$$

So, conditionally,

$$Y|X \sim N(\alpha + \beta X, \sigma_\epsilon^2),$$

and marginally:

$$Y \sim N(\alpha + \beta\mu, \beta^2\sigma_X^2 + \sigma_\epsilon^2).$$

B Calculation of the mean of a left censored normal variable

Let Y be the left censored normal variable and Z its latent variable, with density $f_Z(z)$ and distribution function $F_Z(z)$

$$\begin{aligned} Z &\sim N(\mu, \sigma^2) \\ Y &\sim max(Z, 0). \end{aligned}$$

We calculate the mean $\mathbb{E}Y$:

$$\begin{aligned}\mathbb{E}Y &= \int_0^\infty z f_Z(z) dz \\ &= \frac{1}{\sigma\sqrt{2\pi}} \int_0^\infty (z-\mu) e^{-\frac{1}{2}(\frac{z-\mu}{\sigma})^2} dz + \frac{\mu}{\sigma\sqrt{2\pi}} \int_0^\infty e^{-\frac{1}{2}(\frac{z-\mu}{\sigma})^2} dz.\end{aligned}$$

and we obtain

$$\mathbb{E}Y = \frac{\sigma}{\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{\mu}{\sigma})^2} + \mu(1 - F_Z(0)). \tag{44}$$

C SPSS syntax and R scripts

C.1 SPSS syntax

```
*================================================================.
*Mathematics thesis G.G.A. Westhoff, student nr 0140481, Leiden University.
*Modelling GFR.
* Supervisor Dr. S. LeCessie, Med Stat, LUMC.
*================================================================.

* ********** Load data set (raw data).
GET  FILE='H:\MasterProject\werkbestand.sav'.
*GET  FILE='D:\UniLeiden\MasterProject\Database\werkbestand.sav'.

DATASET NAME DataSet2 WINDOW=FRONT.

/keep= id_gw bmi0 diures0 gesl0 gewich0 gfr_cor0 kahnkl0 lengte0 lf_str0
rook0 str_th0   therap0 u24_vo0 u_100ml u_misl mtpnt1 dagmtp1 gfr_cor1 mtpnt2 dagmtp2 gfr_cor2
 mtpnt3 dagmtp3 gfr_cor3
 mtpnt4 dagmtp4 gfr_cor4
 mtpnt5 dagmtp5 gfr_cor5
 mtpnt6 dagmtp6 gfr_cor6
 mtpnt7 dagmtp7 gfr_cor7
 mtpnt8 dagmtp8 gfr_cor8
 mtpnt9 dagmtp9 gfr_cor9
 mtpnt10 dagmtp10 gfr_co10
 mtpnt11 dagmtp11 gfr_co11
 mtpnt12 dagmtp12 gfr_co12
 mtpnt13 dagmtp13 gfr_co13
 mtpnt14 dagmtp14 gfr_co14
 mtpnt15 dagmtp15 gfr_co15
 uitred dagen dood dag3m.
DATASET NAME DataSet2 WINDOW=FRONT.

*Section {S:exclusion}.
*checkin for typos and other errors in the data, .
IF  (lengte0<100) lengte0=lengte0+100.
EXECUTE.
IF  (bmi0>60) bmi0=10000*gewich0/(lengte0*lengte0).
execute.

* Force gfr_cor0=0 if diures < 200 ml/24h:.
```

```
if (diures0 < 200)  gfr_cor0 = 0.
if (diures1 < 200)  gfr_cor1 = 0.
if (diures2 < 200)  gfr_cor2 = 0.
if (diures3 < 200)  gfr_cor3 = 0.
if (diures4 < 200)  gfr_cor4 = 0.
if (diures5 < 200)  gfr_cor5 = 0.
if (diures6 < 200)  gfr_cor6 = 0.
if (diures7 < 200)  gfr_cor7 = 0.
if (diures8 < 200)  gfr_cor8 = 0.
if (diures9 < 200)  gfr_cor9 = 0.
if (diures10 < 200)  gfr_co10 = 0.
if (diures11< 200)  gfr_co11 = 0.
if (diures12 < 200)  gfr_co12 = 0.
if (diures13 < 200)  gfr_co13 = 0.
if (diures14 < 200)  gfr_co14  = 0.
if (diures15 < 200)  gfr_co15 = 0.


********** Exclude patients (from data set).
* (1) for which urine never was collected (variable diures missing on all visits, even though the patient partici
* or (2) Patient is already anuric at the start of the therapy, that is, has gfr missing or equal to 0 at the fir
FILTER OFF.
USE ALL.
SELECT IF (~((gfr_cor0 < 0 & gfr_cor1< 0 & gfr_cor2< 0 & gfr_cor3 < 0 &  gfr_cor4 < 0 &
    gfr_cor5< 0 & gfr_cor6< 0 & gfr_cor7< 0 & gfr_cor8< 0 & gfr_cor9< 0 & gfr_co10<0 & gfr_co11<0 &
        gfr_co12<0 & gfr_co13<0 & gfr_co14<0 & gfr_co15<0) | (gfr_cor0=0 & gfr_cor1=0)|(gfr_cor0<0 &
     gfr_cor1<0))).
EXECUTE.

********** Switch on the two following lines, when comparing two approaches for two time points (see below).
*select if (not(missing(gfr_cor0)) & not(missing(gfr_cor1))).
*execute.

********** Table {T:baseline_data2} General patient characteristics at base line after exclusion
* Basic Tables.

CROSSTABS
  /TABLES= therap0 BY kahnkl0 gesl0 pnkl40
  /FORMAT=AVALUE TABLES
  /CELLS=COUNT ROW
  /COUNT ROUND CELL.

TABLES
  /FORMAT BLANK MISSING('.')
  /OBSERVATION lf_str0 gfr_cor0 bmi0 lengte0 gewich0 diures0
  /TABLES (lf_str0 + gfr_cor0 + bmi0 + lengte0 + gewich0 + diures0)
  BY therap0 > (STATISTICS)
  /STATISTICS
  mean( )
  stddev( )
 min( )
 max( ).

*Average age,  gfr0, bmi0,  length0, weight0, Residual renal function0,  etc. *all* patients.
DESCRIPTIVES VARIABLES=lf_str0 gfr_cor0 bmi0 lengte0 gewich0 diures0
  /STATISTICS=MEAN STDDEV min max.

*Testing.
T-TEST GROUPS=therap0(1 2)
  /MISSING=ANALYSIS
  /VARIABLES=gfr_cor0 lf_str0 bmi0 lengte0 gewich0
  /CRITERIA=CI(.95).
```

```
* Chi square for categorical variables at baseline.
* Prim Renal Disease.
CROSSTABS
  /TABLES=therap0 BY pnkl40
  /FORMAT=AVALUE TABLES
  /STATISTICS=CHISQ
  /CELLS=COUNT
  /COUNT ROUND CELL.

*Comorbidity.
CROSSTABS
  /TABLES=therap0 BY kahnkl0
  /FORMAT=AVALUE TABLES
  /STATISTICS=CHISQ
  /CELLS=COUNT
  /COUNT ROUND CELL.
********** END table {T:baseline_data2} .

********* Figure {F:histo_gfr} (in section {S:lat_cens}) ****************.
GRAPH
  /HISTOGRAM=gfr_cor0.

GRAPH
  /HISTOGRAM=gfr_cor1.

GRAPH
  /HISTOGRAM=gfr_cor2.

GRAPH
  /HISTOGRAM=gfr_cor3.

GRAPH
  /HISTOGRAM=gfr_cor4.

GRAPH
  /HISTOGRAM=gfr_cor5.

*** End fig.


********************** Transform to long format: variables become cases  ************************.
**  Note: Index1 is variable containing the vsist number (1 =baseline).
compute mtpnt0=0.
compute dagmtp0=0.

VARSTOCASES
  /MAKE gfr_cor FROM gfr_cor0 gfr_cor1 gfr_cor2 gfr_cor3 gfr_cor4 gfr_cor5 gfr_cor6 gfr_cor7
    gfr_cor8 gfr_cor9 gfr_co10 gfr_co11 gfr_co12 gfr_co13 gfr_co14 gfr_co15
    /Make dagmtpnt FROM dagmtp0 dagmtp1 dagmtp2 dagmtp3 dagmtp4 dagmtp5 dagmtp6 dagmtp7 dagmtp8 dagmtp9 dagmtp10
                       dagmtp11 dagmtp12 dagmtp13 dagmtp14 dagmtp15
 /Make diures FROM diures0 diures1 diures2 diures3 diures4  diures5 diures6 diures7 diures8 diures9 diures10 diur
diures14  diures15
  /INDEX=Index1(16)
  /KEEP=id_gw bmi0 gesl0 gewich0 kahnkl0 lengte0 lf_str0 rook0 str_th0  therap0 u24_vo0 u_100ml
    u_misl mtpnt1  uitred dagen dood dag3m
  /NULL=KEEP.

********** Figure (F:mean_1428)  Observed mean gfr, excluding patients who are anuric at baseline.
* Includes anurische patients; see section {S:anuric}.
* Visits with missing data not yet removed.
GRAPH
```

```
  /LINE(MULTIPLE)=MEAN(gfr_cor) BY Index1 BY therap0.

********** Figure (F:participants):  Number of patients with a gfr >= 0 measurement, excluding patients who are
temporary.
select if (not(missing( gfr_cor))).
FREQUENCIES VARIABLES=Index1
  /HISTOGRAM
  /ORDER=ANALYSIS.

* ********** Look for patients which are anuric from certain visit:.
* get rid of missing data.
select if (not(missing(gfr_cor))).
execute.

******** We also have to calculate the observed mean (displayed in figure {F:mean_1428}) in R;.......
*.....to that end we present the data to R by way of the following file:.
save outfile = 'H:\MasterProject\werkbest_long_with_zeroes.sav'.
*save outfile = 'D:\UniLeiden\MasterProject\Database\werkbest_long_with_zeroes.sav'.

*sort cases by id_gw  Index1. LAG_GFR is set to missing on encountering new patient.
if (lag(id_gw) = id_gw)  lag_gfr = lag(gfr_cor).
execute

* Set new variable BOOL to 1, if two consecutive visits are zero.
compute bool=((gfr_cor=0) & (lag_gfr=0)).
execute.

* Increment CUMBOOL as long as same patient  and visits have zero GFR.
* Careful! This is not matter of pure parallelism; SPSS scans the lines top to bottom and....
*........ compares LAG with CURRENt in a looping fashion! I always get mixed up!!!.
compute cumbool=0.
if (lag( id_gw)=id_gw) cumbool = lag(cumbool)+bool.
execute.

********* Select first 6 visits.
select if (Index1 < 7).
EXECUTE.


**********Figure {F:first_time_anuric0}. Number of patients who are for the first time anuric.
compute cumbool_1=0.
if (cumbool=1) cumbool_1=1.
execute.

* We still have to label X-  and Y-axis:.
* X: Visit (1 = baseline); Y: # of patients for the first time anuric.
GRAPH
  /BAR(SIMPLE)=SUM(cumbool_1) BY Index1.
****** END Figure {F:first_time_anuric0}.

********** Table {T:prob_not_anuric} : Probability for a patient to become anuric T:.
* Count patients becoming anuric at any point in time:.
CROSSTABS
  /TABLES=Index1 BY cumbool
  /FORMAT=AVALUE TABLES
  /CELLS=COUNT
  /COUNT ROUND CELL.
**********.

*******!!!!! Only execute in behalf fig {F:rho_j}.
*select if ((cumbool<=1) ).
*execute.
```

```
*save outfile = 'D:\UniLeiden\MasterProject\patients_becoming_anuric'.
*save outfile = 'H:\MasterProject\patients_becoming_anuric'.
*save outfile ='D:\onderwijs\afstudeerproject gerard\patients_becoming_anuric'.
******** END "Only.....".

******* Select only patients in non-anuric stage.
select if (cumbool=0).
execute.

save outfile = 'D:\UniLeiden\MasterProject\Database\werkbest na pat selectie_data_to_cases.sav'.
*save outfile = 'H:\MasterProject\werkbest na pat selectie_data_to_cases.sav'.


***************************** Chapter {S:lin_models} *****************************.
*Table {T:est_coeff}: Estimated regression coefficients based on analysis of response:.
* First check for significant group*time interaction.
* Procedure MIXED of SPSS fits the Linear Mixed model (LMM) which is direct generalization.....
* ......of the LM to repeated measurements. The method is likelihood based....
*.....(Course Rep Measures, page 50).

recode Index1(1=0)(2=3)(3=6)(4=12)(5=18)(6=24) into time.

MIXED gfr_cor BY Index1 therap0
  /CRITERIA=CIN(95) MXITER(100) MXSTEP(5) SCORING(1) SINGULAR(0.000000000001) HCONVERGE(0,
    ABSOLUTE) LCONVERGE(0, ABSOLUTE) PCONVERGE(0.000001, ABSOLUTE)
  /FIXED=Index1 therap0 Index1*therap0 | SSTYPE(3)
  /METHOD=ML
/PRINT=R SOLUTION
  /REPEATED=Index1 | SUBJECT(id_gw) COVTYPE(UN)
  /EMMEANS=TABLES(Index1*therap0) .

* Table {T:fixed_main}.
* Skip the interaction and then check for main effects in group and time.
MIXED gfr_cor BY Index1 therap0
  /CRITERIA=CIN(95) MXITER(100) MXSTEP(5) SCORING(1) SINGULAR(0.000000000001)
HCONVERGE(0, ABSOLUTE) LCONVERGE(0, ABSOLUTE) PCONVERGE(0.000001, ABSOLUTE)
  /FIXED=Index1 therap0 | SSTYPE(3)
/METHOD=ML
  /REPEATED=Index1 | SUBJECT(id_gw) COVTYPE(UN).

************* Table{T:est_coeff_cont}: Estimated regression coefficients for gfrwith time as a continuous.
MIXED gfr_cor BY therap0 WITH Index1
  /CRITERIA=CIN(95) MXITER(100) MXSTEP(5) SCORING(1) SINGULAR(0.000000000001) HCONVERGE(0,
    ABSOLUTE) LCONVERGE(0, ABSOLUTE) PCONVERGE(0.000001, ABSOLUTE)
  /FIXED=therap0 Index1 therap0*Index1 | SSTYPE(3)
  /METHOD=ML
  /PRINT=R SOLUTION
  /REPEATED=Index1 | SUBJECT(id_gw) COVTYPE(UN).
* /EMMEANS=TABLES(Index1*therap0) .

*********** End chapter {S:lin_models} ***********.


************************** Chapter { S:markov} M A R K O V  M O D E L S ********************.
********  Section {S:est_Y_1Y_2_gfr} : .
*Comparing the analysis of response profiles and the Markov model for FIRST TWO VISITS. Do NOT distinguish betwee
*Make sure that database contains only two time points.
*Results in table {T:gfr_est_2}.

*GET file = 'H:\MasterProject\werkbest na pat selectie_data_to_cases.sav'.
GET file='D:\UniLeiden\MasterProject\Database\werkbest na pat selectie_data_to_cases_2.sav'.
```

```
********* Y_2 = \alpha + \beta*tijd.
*time as factor rather than linear covariate!! (niet: MIXED lead WITH time!!).

MIXED gfr_cor BY Index1
  /CRITERIA=CIN(95) MXITER(100) MXSTEP(5) SCORING(1) SINGULAR(0.000000000001) HCONVERGE(0,
    ABSOLUTE) LCONVERGE(0, ABSOLUTE) PCONVERGE(0.000001, ABSOLUTE)
  /FIXED=Index1 | SSTYPE(3)
  /METHOD=REML
  /PRINT=R SOLUTION
  /REPEATED=Index1 | SUBJECT(id_gw) COVTYPE(UN)
 /EMMEANS=TABLES(Index1) .

*******  Markov: Y_2 recursief in Y_1.
*f(Y_1, Y_2)   = f(Y_1). F(Y_2|Y_1).

REGRESSION
/DESCRIPTIVES MEAN STDDEV CORR SIG N
 /MISSING LISTWISE
 /STATISTICS COEFF OUTS R ANOVA
/CRITERIA=PIN(.05) POUT(.10)
/NOORIGIN
/DEPENDENT gfr_cor
/METHOD=ENTER lag_gfr.

*Bepaal nu gemiddelde van Y_1=gfr_cor0:.
DESCRIPTIVES VARIABLES=gfr_cor
  /STATISTICS=MEAN STDDEV MIN MAX.
```

C.2 R script

```
#=============================================================
########## Masterproject G.G.A. Westhoff 2008/2009
#Modelling Repeated Measurements of Renal
#Function during Haemodialysis
#with cut off due to complete kidney failure
#
#University Leiden
# Supervisors Dr. S. LeCessie (LUMC), Prof. DR. R. Gill (Snellius Instute)
#=============================================================
#!!!!This script  is divided in 3 broad sections: Calculations, Figures, Functions.

#===================== Loading packages and reading data set===================.
require("foreign")
require("survival")

#SPSS-data set "werkbest na pat selectie_data_to_cases.sav" is the....
#.... is the work horse of this project and is the data set "werkbestand.sav" ...
#... after cleaning up and transposing to long format.
filepath="H://MasterProject//werkbest na pat selectie_data_to_cases.sav"
#filepath="D://UniLeiden//MasterProject//Database//werkbest na pat selectie_data_to_cases.sav"
#filepath="H://MasterProject//werkbest_long_with_zeroes.sav"

gfr<-read.spss(file=filepath,use.value.labels=FALSE,to.data.frame=TRUE,
max.value.labels=Inf,
 trim.factor.names=FALSE)

#Then execute functions at end of script file!!

#======================== CALCULATIONS ========================================.
#==========Section {S:anuric} Defining patients as anuric
filepath="H://MasterProject//patients_becoming_anuric"
```

```
#filepath="D://UniLeiden//MasterProject//Database//patients_becoming_anuric"

result<-read.spss(file=filepath,use.value.labels=FALSE,to.data.frame=TRUE,
max.value.labels=Inf,
 trim.factor.names=FALSE)

(visit<-result$INDEX1)
(max_visit<-max(visit))
#N=size original data set (short format) excl patients which are anuric at baseline; ....
#...see SECTION {S:exclusion}

freq<-table(visit,result$CUMBOOL)
p_j <- freq[,2]/(freq[,1]+freq[,2])
(rho_j<-cumprod(1 - p_j))
barplot(rho_j, xlab= "Visit (1 = baseline)",
ylab="rho_j Estimated probability to be not yet anuric",
names.arg=c("1", "2", "3", "4", "5" ,"6"))



#Ch {S:gfr_zeroes} "Modeling GFR over time using Markov models in which GFR is censored"


#===========Section: Markov models with censored GFR
gfr2 <- gfr[!is.na(gfr$LAG_GFR),] # Throw out rows with missing values.

#Y_2<-gfr2$GFR_COR #
#gfr_cor <- Y_2
#Y_1 <-gfr2$LAG_GFR #
#lag_gfr <- Y_1
#Index1  <- gfr2$INDEX1
#therap0 <- gfr2$THERAP0

# ======== ============.
#Aggregate model (without group var):
diagn <-gfr_survreg(gfr2$LAG_GFR, gfr2$GFR_COR )#Idem
diagn
summary(diagn)

#Same analysis, but select patients HD=1
gfrHD<- gfr[(gfr$THERAP0==1)&!is.na(gfr$LAG_GFR),] # Throw out rows with missing values.
(diagnHD <-gfr_survreg(gfrHD$LAG_GFR, gfrHD$GFR_COR ))#Idem
summary(diagnHD)
#diagnHD[[1]]: regression coeffs
aHD <- diagnHD[[1]][1]# surv regr coeffs: intercept
bHD <- diagnHD[[1]][2] # Idem slope
sdHD <- diagnHD[[8]] # stand dev of latent var Z1, Z2,...,Z6.

#Same analysis, but select patients PD=2
gfrPD <- gfr[(gfr$THERAP0==2)&!is.na(gfr$LAG_GFR),] # Throw out rows with missing values.
(diagnPD <-gfr_survreg(gfrPD$LAG_GFR, gfrPD$GFR_COR ))#Idem
#diagnPD[[1]]
aPD <- diagnPD[[1]][1]# surv regr coeffs: intercept
bPD <- diagnPD[[1]][2] # Idem slope
sdPD <- diagnPD[[8]] # stand dev of latent var Z1, Z2,...,Z6.


#Interaction between GFR and therapy:
(diagnInter <-gfr_survreg2(gfr2$LAG_GFR, gfr2$GFR_COR , gfr2$THERAP0))
summary(diagnInter)
sdInter<-diagnInter[[8]]
```

```
#As 'gfr_survreg2' with TIME as continuous param: =====
(diagnTime <-gfr_survreg3(gfr2$LAG_GFR, gfr2$GFR_COR , gfr2$THERAP0, gfr2$INDEX1))
summary(diagnTime)

#??? fooPred<-data.frame(THERAP0, INDEX1,PredPD, PredHD)

#========= Section: Performance of the selected model=============.
filepath3="H://MasterProject//werkbest_long_with_zeroes.sav"
#filepath3="D://UniLeiden//MasterProject//Database//werkbest_long_with_zeroes.sav"

gfr3<-read.spss(file=filepath3,use.value.labels=FALSE,to.data.frame=TRUE,
max.value.labels=Inf,
 trim.factor.names=FALSE)
gfr3<-gfr3[gfr3$INDEX1 < 7,]
#attach(gfr3)
summary(gfr3)

#Observed as in fig {F:mean_1428}:
(aggr_mean<-tapply(gfr3$GFR_COR, list(gfr3$INDEX1, gfr3$THERAP0), mean)) #
(aggr_sd <- tapply(gfr3$GFR_COR, list(gfr3$INDEX1, gfr3$THERAP0), sd)) #

(EY1_HD <- aggr_mean[1,1])# EZ1=EY1. Z1 is latent of Y1, the truncation of Z1.
(sdY_HD <- aggr_sd[1,1])

(EY1_PD <- aggr_mean[1,2])#
(sdY_PD <- aggr_sd[1,2])

#Simulation for ******HD =1*******:
#Simulate N patients.
N <- 10000

Y_1<- rnorm(N, mean=EY1_HD, sd=sdY_HD ) #Z_1=Y_1

Z_2 <- aHD + bHD*Y_1 + rnorm(N, 0, sdInter) # Z_2|Z_1 ~ a + bZ_1 + eps
Y_2 <- pmax(Z_2,0)
EY_2 <-mean(Y_2))#E(Y_2|Y_1

Z_3 <- aHD + bHD*Z_2 + rnorm(N, 0, sdInter)
Y_3 <- ifelse(Y_2 > 0, pmax(Z_3,0)  ,0)
EY_3 <-mean(Y_3)

Z_4 <- aHD + bHD*Z_3 + rnorm(N, 0, sdInter)
Y_4 <- ifelse(Y_3>0, pmax(Z_4,0)  ,0)
(EY_4 <-mean(Y_4))

Z_5 <- aHD + bHD*Z_4 + rnorm(N, 0, sdInter)
Y_5 <- ifelse(Y_4 > 0, pmax(Z_5,0)  ,0)
(EY_5 <-mean(Y_5))

Z_6 <- aHD + bHD*Z_5 + rnorm(N, 0, sdInter)
Y_6 <- ifelse(Y_5>0, pmax(Z_6,0)  ,0)
(EY_6 <-mean(Y_6))

# Transport to FIGURE section below:
(PredHD.trunc <- c(EY1_HD, EY_2, EY_3, EY_4, EY_5, EY_6))

#****** PD=2 *********
Y_1<- rnorm(N, mean=EY1_PD, sd=sdY_PD ) #Z_1=Y_1
mean(Y_1)
EY1_PD

Z_2 <- aPD + bPD*Y_1 +rnorm(N, 0, sdInter) # Z_2|Z_1 ~ a + bZ_1 + eps
```

```
Y_2 <- pmax(Z_2,0)
(EY_2 <-mean(Y_2))#E(Y_2|Y_1)

Z_3 <- aPD + bPD*Z_2 + rnorm(N, 0, sdInter)
Y_3 <- ifelse(Y_2 > 0, pmax(Z_3,0)  ,0)
(EY_3 <-mean(Y_3))

Z_4 <- aPD + bPD*Z_3 + rnorm(N, 0, sdInter)
Y_4 <- ifelse(Y_3>0, pmax(Z_4,0)  ,0)
(EY_4 <-mean(Y_4))

Z_5 <- aPD + bPD*Z_4 + rnorm(N, 0, sdInter)
Y_5 <- ifelse(Y_4 > 0, pmax(Z_5,0)  ,0)
(EY_5 <-mean(Y_5))

Z_6 <- aPD + bPD*Z_5 + rnorm(N, 0, sdInter)
Y_6 <- ifelse(Y_5>0, pmax(Z_6,0)  ,0)
(EY_6 <-mean(Y_6))

# Transport to FIGURE section below:
(PredPD.trunc <- c(EY1_PD, EY_2, EY_3, EY_4, EY_5, EY_6))

#==================================FIGURES=========================================.
#======== Figure{F:latent_vars} "Illustrating latency and censoring"
#
old_par <- par(mfrow = c(1,2))
par(mfrow =c(1,2))

theta<-c(2, 1.5)#mean and sigma of Y
y<-seq(0, 8, .1)
f_Y<-dnorm(y, theta[1], theta[2])
z <- seq(-4,0,.1)
f_Z <- dnorm(z, theta[1], theta[2])
plot(seq(-4, 8,.1), seq(0, .3, .0025), type="n", xlab="y,z (GFR)", ylab="density of Y and Z")
points( c(0,0), c(0,.2), lty=1, type="l", lwd=4) #vertical bar measuring censored zeroes
#points( c(0,0), c(0,.4), lty=1, type="l", lwd=1) #vertical axis (
points(y, f_Y, type="l") #Observed GFR
points( z, f_Z, lty=2, type="l") # Latent GFR
legend( x="topright", c("Y, Z if Y > 0", "Z if Y < 0",
  "P(Y=0) = F_Z(0)"), lty=c(1,2,1), lwd=c(1,1,4), inset=.01)

# As above, but give GFR higher mean; this will decrease the value P(Y=0)
theta<-c(4, 1.5) #mean and sigma of Y
y<-seq(0, 8, .1)
f_Y<-dnorm(y, theta[1], theta[2])
z <- seq(-4,0,.1)
f_Z <- dnorm(z, theta[1], theta[2])
plot(seq(-4, 8,.1), seq(0, .3, .0025), type="n", xlab="y,z (GFR)", ylab="density of Y and Z")
points( c(0,0), c(0,.025), lty=1, type="l", lwd=4) #vertical bar measuring censored zeroes
#points( c(0,0), c(0,.4), lty=1, type="l", lwd=1) #vertical axis (
points(y, f_Y, type="l") #Observed GFR
points( z, f_Z, lty=2, type="l") # Latent GFR
legend( x="topright", c("Y, Z if Y > 0", "Z if Y < 0",
  "P(Y=0) = F_Z(0)"), lty=c(1,2,1), lwd=c(1,1,4), inset=.01)
# ============ END demonstrating concepts of latency and ================

#======== Table {T:est_coeff_cont}, cols time as cont covariate =========
a<- 5.980107
b<- -0.624122
c<- -0.623639
d<- -0.055291
```

```
tijd <- 1:6

grp <- 1 # HD
est_hd_cont<- a + b*tijd + c*grp + d*tijd*grp

grp<-0
est_pd_cont<- a + b*tijd + c*grp + d*tijd*grp

#Ditto, time as category (response profiles) table {T: est_mean_gfr}
est_hd_catg <- c(5.276455, 3.725404, 3.157205, 2.459885, 2.043392, 1.471048)
est_pd_catg <- c(5.903960, 4.481025, 3.922911, 3.331491, 2.781002, 2.475061)

# ==== Figure {F:obs_est_resp_profiles} compare observed means with response profiles
plot(1:6, 1:6, type="n", xlab="visit (1 = baseline)", ylab="mean gfr")
points(1:6,  aggr_mean[,1], lty=2, lwd=2, col=2, type="o") #
points( 1:6,  aggr_mean[,2], lty=1, lwd=2, col=1, type="o")#
points( 1:6, est_hd_catg,lty=4, lwd=2, col=4, type="o")#
points( 1:6, est_pd_catg,lty=3, lwd=2, col=3, type="o")#
legend(x="bottomleft", c("PD, observed means", "HD, observed means", "PD, analysis of resp profiles",
"HD, analysis of resp profiles"), col=1:4, lty=1:4, inset=.05)

#===== Figure {F:obs_est_cont} compare observed means with time as cont covariate
plot(1:6, 1:6, type="n", xlab="visit (1 = baseline)", ylab="mean gfr")
points(1:6,  aggr_mean[,1], lty=2, lwd=2, col=2, type="o") #
points( 1:6,  aggr_mean[,2], lty=1, lwd=2, col=1, type="o")#
points( 1:6, est_hd_cont,lty=4, lwd=2, col=4, type="o")#
points( 1:6, est_pd_cont,lty=3, lwd=2, col=3, type="o")#
legend(x="bottomleft", c("PD, observed means", "HD, observed means", "PD, time as cont covariate",
"HD, time as cont covariate"), col=1:4, lty=1:4, inset=.05)

#===== Figure {F:est_mean_gfrplot} estimates response profiles and lin trend model.
plot(1:6, 1:6, type="n", xlab="visit (1 = baseline)", ylab="estimated mean gfr")
points(1:6, est_hd_catg, lty=2, lwd=2, col=2, type="o") #
points( 1:6, est_pd_catg, lty=1, lwd=2, col=1, type="o")#
points( 1:6, est_hd_cont,lty=4, lwd=2, col=4, type="o")#
points( 1:6, est_pd_cont,lty=3, lwd=2, col=3, type="o")#
legend(x="bottomleft", c("PD, analysis of resp profiles", "HD, analysis of resp profiles",
"PD, continuous time", "HD, continuous time"), col=1:4, lty=1:4, inset=.05)


#!Coefficients from table {T:lik_regression}
(coeffHD<-c(diagnHD[[1]],aggr_mean[1,1]))#HD=1
(PredHD.cens<-PredMarkov(coeffHD)) #Predicted Means
(coeffPD<-c(diagnPD[[1]],aggr_mean[1,2])) #Idem
(PredPD.cens<-PredMarkov(coeffPD))

oldpar <- par(mfrow=c(2,2))

#=== Figures describing Markov models with censoring ======
#=== Figure {F:Observed_Modeled} Obs and modeled means
plot(c(0,rownames(aggr_mean)), 0:6, type="n", xlab="visit (1=baseline)", ylab="GFR")
points( row.names(aggr_mean), aggr_mean[,1], lty=2, lwd=2, col=2, type="o") # Observed 1=HD
points( row.names(aggr_mean), aggr_mean[,2], lty=1, lwd=2, col=1, type="o")#Observed PD
points( row.names(aggr_mean), PredHD.cens,lty=4, lwd=2, col=4, type="o")# Calaculated HD
points( row.names(aggr_mean), PredPD.cens, aggr_mean[,2],lty=3, lwd=2, col=3, type="o")#Calculated PD
legend(x="bottomleft", c("Observed PD", "Observed HD", "PD Markov model with censored GFR",
"HD Markov model with censored GFR"), col=1:4, lwd=2, lty=c(1,2,3,5),inset=.05)

# ======== Censored zeroes, with truncated distrib
#(PredHD.trunc <- EY_trunc (PredHD.cens, aggr_sdHD))# Erik's gemiddelden.
#(PredPD.trunc <- EY_trunc (PredPD.cens, aggr_sdPD))
```

```
#PredHD.trunc
#PredHD.cens

#==== Figure {F:Observed_Modeled_Max}. EY_1 is observed mean.
plot(c(0,rownames(aggr_mean)), 0:6, type="n", xlab="visit (1=baseline)", ylab="GFR")
points( row.names(aggr_mean), aggr_mean[,1], lty=2, lwd=2, col=2, type="o") # Observed 1=HD
points( row.names(aggr_mean), aggr_mean[,2], lty=1, lwd=2, col=1, type="o")#Observed PD
points( row.names(aggr_mean), PredHD.trunc,lty=4, lwd=2, col=4, type="o")# Calaculated HD
points( row.names(aggr_mean), PredPD.trunc, aggr_mean[,2],lty=3, lwd=2, col=3, type="o")#Calculated PD
legend(x="bottomleft", c("Observed PD", "Observed HD", "PD Markov model (simulation)",
"HD Markov model (simulation)"), col=1:4, lwd=2, lty=c(1,2,3,5), inset=.05)


#==== Figure {F:max_cens}
plot(c(0,rownames(aggr_mean)), 0:6, type="n", xlab="visit (1=baseline)", ylab="GFR")
points( row.names(aggr_mean), PredHD.cens, lty=2, lwd=2, col=2, type="o") # Observed 1=HD
points( row.names(aggr_mean), PredPD.cens, lty=1, lwd=2, col=1, type="o")#Observed PD
points( row.names(aggr_mean), PredHD.trunc,lty=4, lwd=2, col=4, type="o")# Calaculated HD
points( row.names(aggr_mean), PredPD.trunc, aggr_mean[,2],lty=3, lwd=2, col=3, type="o")#Calculated PD
legend(x="bottomleft", c("PD censored GFR", "HD censored GFR", "PD truncated Gaussian",
"HD truncated Gaussian"), col=1:4, lwd=2, lty=c(1,2,3,5), inset=.05)

par(oldpar)

#lty (0=blank, 1=solid, 2=dashed, 3=dotted, 4=dotdash, 5=longdash, 6=twodash)
#or as one of the character strings "blank", "solid", "dashed", "dotted", "dotdash",
#"longdash", or "twodash", where "blank" uses invisible lines (i.e., does not draw them

#========================== *NO* censored zeroes ==========================
#coeffHD<-c( 1.271, 0.519, - 0.086,  5.201058)#a, b, c, Y_1
#PredHD<-PredMarkov(coeffHD)
#coeffPD<-c( 1.058,0.598 , - 0.014, 5.800045 ) #Idem
#PredPD<-PredMarkov(coeffPD)

#plot(c(0,rownames(aggr_mean)), 0:6, type="n", xlab="visit (1=baseline)", ylab="filtration rate gfr")
#points( row.names(aggr_mean), aggr_mean[,1], lty=2, type="o") # 1=HD
#points( row.names(aggr_mean), aggr_mean[,2],type="o")#PD
#points( row.names(aggr_mean), PredHD,lty=4, type="o")#HD
#points( row.names(aggr_mean), PredPD, aggr_mean[,2],lty=3, type="o")#PD
#legend(x="bottomleft", c("PD", "HD", "PD Markov model(*), NO censored zeroes",
#  "HD Markov model(*), NO censored zeroes",
#  "*: Y_i = alpha + beta Y_{i-1} + ganma*i + epsilon"), lty=1:4, inset=.05)

#================
#==================== END FIGURES ==============================

#============================== FUNCTION DEFINITIONS ==================================
gfr_survreg <- function(Y_1, Y_2){
# Survival regression  Y_2 ~Y_1
#Formula's {E:survreg} and {E:formula}
event<-as.numeric(Y_2 > 0)
survreg(Surv(Y_2, event, type="left")~Y_1, dist="gaussian")
}

gfr_survreg2 <- function(Y_1, Y_2, therap0){
#As 'gfr_survreg', but with interactions.
#
# Survival regression  Y_2 ~Y_1*group
#Formula {E:survreg2}
event<-as.numeric(Y_2 > 0)
survreg(Surv(Y_2, event, type="left")~Y_1*as.factor(therap0), dist="gaussian")
}
```

```
gfr_survreg3 <- function(Y_1, Y_2, therap0, time){
#As 'gfr_survreg2', but with time as linear covariate.
#Formula {E:survreg2_time}
event<-as.numeric(Y_2 > 0)
survreg(Surv(Y_2, event, type="left")~Y_1*as.factor(therap0) + time, dist="gaussian")
}

gfr_survreg_time <- function(Y_1, Y_2, time){
#foo2
#As gfr_survreg, but include time as covariate
event<-as.numeric(Y_2 > 0)
#return(event)
#Z_2 <- -Y_2; Z_1 <- -Y_1 #Negate, so data become right censored
#return(Surv(Z_2, event))
survreg(Surv(Y_2, event, type="left")~Y_1 + time, dist="gaussian")
#survreg(Surv(Z_2, event)~Z_1 + time, dist="gaussian")
}

PredMarkov<-function(coeff){
#Recursive definition.
a<-coeff[1]; b<-coeff[2];Y_i_1<-coeff[3]
visits<-6
Y_i<- rep(0,visits)
Y_i[1]<-Y_i_1
for (i in 2:6){
  Y_i[i]<-a +b*Y_i_1
  Y_i_1<-Y_i[i]
  }
  return(Y_i)
}

EY_trunc <-function(mu, sigma)
{
(sigma/(sqrt(2*pi)))*exp(-.5*(mu/sigma)^2) + mu*(1-pnorm(0,mean=mu,sd=sigma))
}
#======================== END FUNCTIONS ==========================
```

D Glossary of terms

References

[1] Fitzmaurice, G.M., Laird, N.M and Ware, J.H. (2004). *Applied Longitudinal Analysis*, Wiley Interscience.

[2] Diggle, P.J., Liang, K.-Y. and Zeger, S.L. (1994). *Analysis of Longitudinal Data*. New York: Oxford University Press.

[3] Boeschoten, Dr. E.W. (2003). *Necosad Eindverslag*, http://necosad.nl/files/reports/final_report_necosad.pdf.

[4] Rice, J.A. (1995). *Mathematical Statistics and Data Analysis*, 2-nd ed. Duxbury Press

[5] Klein, J.P. and Moeschberger, M.L. (1997). *Survival Analysis*. New York: Springer