W. Bandilla and F. Faulbaum (Eds.) (1997), SoftStat '97
Advances in Statistical Software 6, Stuttgart: Lucius & Lucius, 369-376

Description of methods of longitudinal data analysis with the help of an investigation of irregularities of the cardiac rhythm in animal experiments

E. Schuster, C. Schlesinger and H. Löster

Summary

Estimation methods such as maximum-likelihood (ML), restricted maximum-likelihood (REML) and generalized estimating equations (GEE) are introduced for longitudinal data analysis at the linear model with correlated errors. Their application will be demonstrated at an example of investigations of disturbances of cardiac rhythm in animal experiments.

1. Methods of Parameter Estimation for Longitudinal Data Models

The estimation methods that were used, maximum-likelihood (ML), restricted maximum likelihood (REML) and generalized estimating equations (GEE), are referred to Diggle et al. (1994) and Arminger (1995). They base on the general linear model with correlated errors for longitudinal data. Therefore, the following is introduced:

Y_{ij} = random variable of j-th response of i-th subject: j=1...n; i=1...m
y_{ij} = realized value of Y_{ij}
t_j = time at which Y_{ij} was measured (the same for all i)
$\mathbf{x}_{ij}$ = line vector of covariables (of length p)

$Y_{ij} = x_{ij1}\,\beta_1 + x_{ij2}\,\beta_2 + ... + x_{ijp}\,\beta_p + \epsilon_{ij}$

$Y_{ij} = \mathbf{x}_{ij}\,\beta + \epsilon_{ij}$

$\mathbf{Y}_i = \mathbf{X}_i\,\beta + \epsilon_i$ with $\mathbf{Y}_i = (Y_{i1},..., Y_{in})^T$ and n*p matrix $\mathbf{X}_i$

$\mathbf{Y} = \mathbf{X}\,\beta + \epsilon$ with $\mathbf{Y}=(\mathbf{Y}_1^T, ..., \mathbf{Y}_m^T)^T$ and nm*p matrix $\mathbf{X}$

assumptions:

- $E(\epsilon_{ij}) = 0$ for all i and all j
- $Var(\epsilon_i) = \mathbf{V}_0$ for all i
- ϵ_i and ϵ_j are independent for $i \neq j$

It follows:

$E(\mathbf{Y}) = \mathbf{X}\,\beta$

$Var(\mathbf{Y}) = \mathbf{V} = \mathbf{I} \otimes \mathbf{V}_0$ block diagonal.

1.1. Maximum Likelihood Estimation under Gaussian Assumptions

Assuming, the data has a multivariate normal distribution, the log-likelihood has to be maximized

$$L(\beta,\sigma^2,V_0) = -0.5[nm\ln\sigma^2 + m\ln|V_0| + \sigma^{-2}(y-X\beta)^T (I\otimes V_0)^{-1} (y-X\beta)] .$$

To given V_0 the maximization over β results in

$$\hat{\beta}(V_0) = (X^T (I\otimes V_0)^{-1} X)^{-1} X^T (I\otimes V_0)^{-1} y .$$

If this result is used the maximization of reduced log-likelihood over V_0 will remain

$L(V_0) = -0.5\ m\ \{n \ln(RSS(V_0)) + \ln|V_0|\}$ with $RSS(V_0) = (y - X\hat{\beta}(V_0))^T (I \otimes V_0)^{-1} (y - X\hat{\beta}(V_0))$

1.2. Restricted Maximum Likelihood Estimation

The basic idea of REML is to transform data, so that transformed data does not depend on β. REML is the maximization of likelihood for the transformed data.
To given V_0 the maximization over β has the results as above

$$\beta(V_0) = (X^T (I \otimes V_0)^{-1} X)^{-1} X^T (I \otimes V_0)^{-1} y .$$

If this result is used the maximization of restricted reduced log-likelihood over V_0 will remain

$$L^*(V_0) = -0.5\ m\ \{n \ln(RSS(V_0)) + \ln|V_0|\} - 0.5 \ln|X^T (I \otimes V_0)^{-1} X|$$

$$= L(V_0) - 0.5 \ln|X^T (I \otimes V_0)^{-1} X| \quad \text{with} \quad RSS(V_0) = (y - X\hat{\beta}(V_0))^T (I \otimes V_0)^{-1} (y - X\hat{\beta}(V_0)) .$$

Restricted log-likelihood $L^*(V_0)$ differs from log-likelihood $L(V_0)$ only by the additional term.

1.3. Pseudo-ML Estimation

It is assumed, that in the ML estimation the true conditional density of $\mathbf{Y}_i$ given $\mathbf{X}_i$ namely $f^*(\mathbf{Y}_i|\mathbf{X}_i)$ is known up to an unknown parameter vector β of dimension p.
Now partial misspecification of the density is assumed. The true but unkown density of $\mathbf{Y}_i$ given $\mathbf{X}_i$ namely $f^*(\mathbf{Y}_i|\mathbf{X}_i)$ has a conditional expected value $E_{f^*}(\mathbf{Y}_i|\mathbf{X})$. Furthermore, the researcher has specified - up to an unknown parameter vector β - a density $f(\mathbf{Y}_i|\mathbf{X}_i, \beta)$ where β is a parameterization of $E_f(\mathbf{Y}_i|\mathbf{X}_i, \beta) = \mathbf{X}_i \beta$. If a vector β_0 exists so that

$$E_{f^*}(\mathbf{Y}_i|\mathbf{X}_i) = E_f(\mathbf{Y}_i|\mathbf{X}_i, \beta_0),$$

then β_0 is also a parameterization of $E_{f^*}(\mathbf{Y}_i|\mathbf{X}_i)$. It is only the mean structur that is specified in this model in the parameter vector β_0. One can get pseudo-ML (PML) estimation of β_0 by maximization of the log-likelihood with the assumed density $f(\mathbf{Y}_i|\mathbf{X}_i, \beta)$. Since the assumed density may be misspecified-except of the mean structure-this estimation method is called pseudo-ML.
Gouriroux et al. (1984) show that PML estimation of β_0 based on the assumed density yields a consistent estimator $\hat{\beta}$ of β_0 if and only if the assumed density is a member of the univariate or multivariate linear exponential family. An important consequence of this result is that whenever the conditional mean structure is correctly specified one can use linear or nonlinear least squares to get consistent estimates of β_0 regardless of $Var(\epsilon_i)$ or any other property of the true distribution.
Assuming, V* is the estimation of the covariance matrix Var(Y) of the chosen model, whereas $\hat{V}$ is a consistent estimated covariance matrix for any covariance structure. As an example:

$$\hat{V} = I \otimes (y_i - \mathbf{X}_i \hat{\beta}) (y_i - \mathbf{X}_i \hat{\beta})^T.$$

Furthermore, it is provided that the mean value structure is correctly specified, that means the chosen model corresponds with the true model. The mean value parameter is then estimated consistently. Furthermore, the covariance matrix of the mean value parameter β can also be kept asymtotically consistent by the so-called robust or sandwich estimator even at error specific covariance structure

$$Var(\beta) = (\mathbf{X}^T V^{*-1} \mathbf{X})^{-1} (\mathbf{X}^T V^{*-1} \hat{V} V^{*-1} \mathbf{X}) (\mathbf{X}^T V^{*-1} \mathbf{X})^{-1}.$$

The derivated standard deviations should also be used at ML estimation, because they will also allow correct decisions if the model of covariance structure or distribution assumption is questionable. If, however, all preconditions for the chosen model can be fulfilled, then V* is a consistent estimator and can be used instead of $\hat{V}$ in the formula mentioned above.
Thus, the formula above simplifies to

$$\mathrm{Var}(\beta) = (X^T V^{*-1} X)^{-1}$$

of the so-called naive estimation.

1.4. Parametric Models for Covariance Structure

All n (n+1) parameter of V_0 in ML or REML models mentioned above have to be estimated without a parametric model. If the number of the moments of time n is too high, it will be better for V_0 to use a parametric model of the covariance structure.
The following model assumptions can be made:

$$\epsilon_{ij} = U_i + W_i(t_j) + Z_{ij} \quad \text{with}$$

- $U_i \sim N(0, \nu^2)$ - random intercept
- $W_i(t)$ stationary process with $E(W(t))=0$ and $\mathrm{Cov}(W_i(t), W_i(s)) = \sigma^2 \rho(|t-s|) = \sigma^2 \exp(-\phi|t-s|^2)$ - serial correlation
- $Z_{ij} \sim N(0, \tau^2)$ - time independent measure error.

With the help of this, V_0 can be estimated with only a few parameters.

1.5. Generalized Estimating Equations for Mean Structures

If the assumed density of $\mathbf{Y}_i$ given $\mathbf{X}_i$ is the multivariate normal density with fixed covariance matrix V_0, the kernel of the pseudo-log-likelihood function may be written as

$$l(\beta) = (\mathbf{Y} - \mathbf{X}\beta)^T (I \otimes V_0)^{-1} (\mathbf{Y} - \mathbf{X}\beta) .$$

Now the pseudo-log-likelihood-function is differentiated in respect to parameters and the first derivative is set to zero

$$s(\beta) = \mathbf{X}^T (I \otimes V_0)^{-1} (\mathbf{Y} - \mathbf{X}\beta) = 0.$$

These score equations are called generalized estimating equations for the mean structure by Liang and Zeger (1986). Until now it is a special case of PML estimation for mean structures. But in the GEE approach by Liang and Zeger (1986) the working covariance matrix V_0 is parameterized as a function of β and a vector α of additional parameters with true value α_0 which are also estimated from the data. Therefore, the generalized estimating equations take the form

$$s(\beta) = \mathbf{X}^T (I \otimes V_0(\beta, \alpha))^{-1} (\mathbf{Y} - \mathbf{X}\beta) = 0.$$

It is to solve iteratively.
In the first step, V_0 is fixed to the unit matrix I. Then, β is estimated by solving $s(\beta) = 0$ yielding $\beta^{(1)}$ (the OLS approach). From $\beta^{(1)}$, residuals may be computed, which allow the estimation of α_0 by a consistent estimator $\alpha^{(1)}$. It follow $V_0^{(1)} = V_0(\beta^{(1)}, \alpha^{(1)})$ and the next iteration until convergence for variance stabilization. The parameters α are not of primary interest but are considered as nuisance parameters. The procedure is called GEE1 because only the mean value structure β is estimated which means the moments of first order. The GEE supply was illustrated for the linear model only. But it also holds true for other distributions from the linear exponential family. Chapter 2.2. deals with one of those examples. The information sandwich is used to estimate $\mathrm{Var}(\beta)$.

2. Illustration of Methods with the Help of an Investigation of Disturbances of Cardiac Rhythm in Animal Experiments

Damage to the heart caused by ischemia and reperfusion can be investigated using isolated hearts from test animals (models e.g. by NEELY and LANGENDORFF). Besides hemodynamic and metabolic changes, alterations of electrophysiological parameters (ECG) can be monitored.
In our investigations ischemic injury of the myocard was induced by a global ischaemia lasting 20 minutes without rest coronary flow ("no flow" ischemia), followed by a 60 minute reperfusion period. The extent of the ischemia injury does not only depend on the conditions of ischaemia, but also on the composition of the perfusion solution. Therefore, the hearts were firstly perfused with a standard perfusion solution for stabilization, followed by the solution that was to be tested, which contained either 0.4 or 1.2 mmol/l of sodium palmitate and with or without 5 mmol/l L-carnitine. L-carnitine is an essential component of the fatty acid transport system, which causes permeation of long-chain fatty acids through the inner membrane of mitochondria and causes a protective effect on the mitochondrial function and therefore on the whole myocard in the reperfusion phase. This has lead to controversies in literature.
Not only hemodynamic quantities such as left ventricular pressure, pressure rate product, contractility and relaxation velocities, and coronary flow were monitored and evaluated as parameters of the ischaemia injury in isolated hearts but also electrophysiological quantities (heart rate and ECG). The behaviour of the heart rate was tested before and after ischemia and in particular in the first 10 reperfusion minutes. The rhythm disturbances in the isolated hearts were monitored and evaluated in the early (1st - 10th minute) and late (10th - 60th minute) reperfusion phase with the help of a self- developed arrhythmia score.
In the following it will be investigated whether L-carnitine develops an influence on heart rate, incidence and degree of seriousness of rhythm disturbances in the reperfusion period of isolated hearts containing glucose and various fatty acid concentrations in the perfusion solution.
It turned out to be useful to evaluate the early and late reperfusion phase separately. Starting point was a model which contains all three factors sodium palmitate, L-carnitine, and time as well as corresponding interactions in each case.
The following calculations were done with "object-oriented software for the analysis of longitudinal data" (OSWALD) by David M. Smith and Peter J. Diggles in S-Plus which intergates the GEE-function by Liang and Zeger (1986).

2.1. Mean Response of Heart Rate

In each case, values at the beginning of global ischemia (-20 min) were defined as 100% to a uniform standard and the reperfusion results were related to that proportionally to compare reperfusion values of the heart rate at individually different values at the beginning.
At the heart rate the mechanic heart rate was valid. This lead to the following: a ventricle tachycardia lasting more than one minute, which got along without or with minimal pressure development was calculated with a heart rate of 0/min in this minute because an estimation of the heart rate of e.g. 1000/min would not be very useful.
In the early phase, the model with all three factors showed that both the main effect and all interaction effects with L-carnitine have no significant impact on the heart rate. Therefore, the next used model was a REML-model which was calculated with the factors sodium palmitate and time only. The most important results can be found in table 1. Next to the parameters the naive z-value and the robust z-value are given in brackets (see chapter 1.3). The z-values have an asymptotic standard normal distribution. Thus, they have to be absolutely higher than 1.96 to get a significant difference from zero for the parameters at 5% error probability.

Table 1. Models of heart rate in the early reperfusion phase.

parameter (naive z robust z)	intercept	Na-palmitate	time	Na-palmitate time
REML	35.59 (7.37)	-11.07 (-1.62)	3.46 (5.72)	2.21 (2.59)
ML	35.59 (7.43)	-10.93 (-1.61)	3.46 (5.67)	2.19 (2.52)
GEE (M=9)	35.86 (6.88 6.03)	-12.35 (-1.67 1.82)	3.27 (4.48 4.21)	2.60 (2.52 2.76)
GEE (exchangeable)	35.31 (8.54 5.79)	-14.49 (-2.48 -2.11)	3.53 (8.95 4.34)	2.73 (4.88 2.77)

The ML approach results in similar parameters (s. table 1). Parameters were also estimated with GEE under the assumption of Gaussian distribution and stationary M-dependent of the correlation structure. Stationary means, for time points with the same distance an equal correlation is assumed. M=1 means, that only adjacent time points are correlated. Thus, in the estimated working correlation only the diagonals that are directly next to the main diagonal are filled with the same values, whereas all the other fields are zero. At M=2 both side diagonals (to each side) are filled with one value in each case. A similar model as REML shows a possibly high M (here 9 for 10 moments of time). As the parameters drop according to a regulation to side diagonals of working correlation matrix at REML, REML needs one parameter only. The last model that is illustrated in table 1 shows that all correlations are assumed to be equal, which means they are called exchangeable or compound symmetry. Similarity of estimations in table 1 is related to their

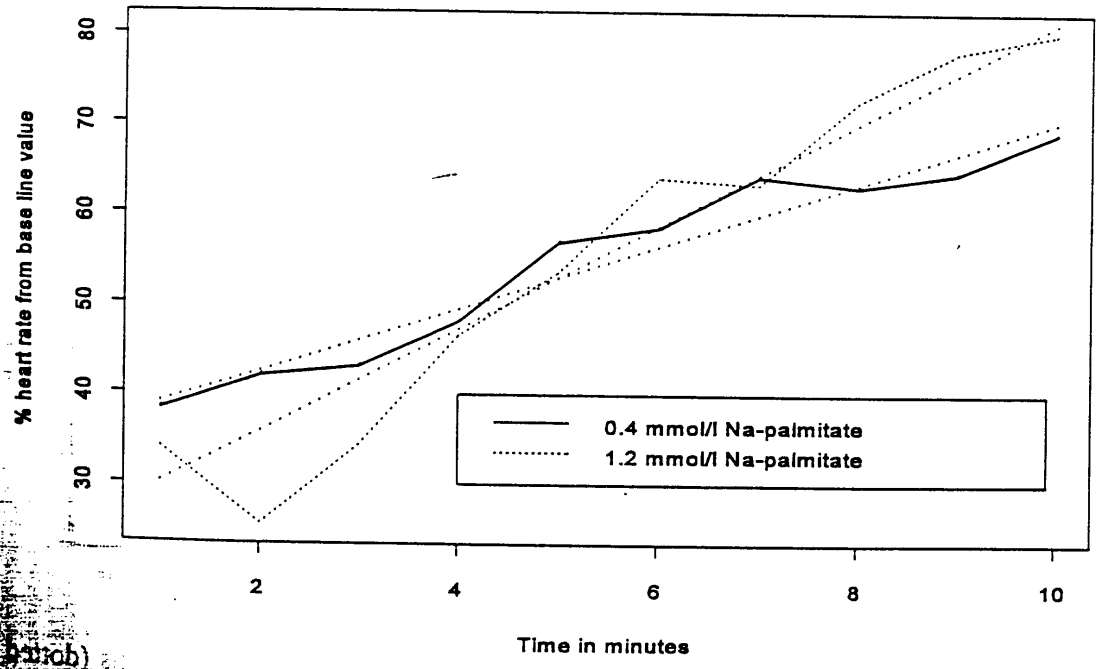

Figure 1. Mean values of heart rate grouped by Na-palmitate and REML estimations (dotted lines) in the early reperfusion phase.

asymptotic consistency, but there was no convergence for stationarity with M=1. An advantage of the integrated GEE-function in OSWALD is to give not only naive but also robust z-values.
In the early phase of reperfusion the heart rate recovers significantly faster at higher concentrations of sodium palmitate than at lower concentrations. Figure 1 shows the mean value courses in the sodium palmitate groups and their corresponding REML-estimators

y_l = 35.59 + 3.46 * time
y_h = 35.59 - 11.07 + (3.46+2.21) * time.

A higher heart rate which is kept during the late phase is achieved at the end of the early phase by the faster increase at higher sodium palmitate. It is only Na-palmitate that remains an influencing factor because neither time nor L-carnitine have significant influence on the heart rate in the late phase. Analogous variants as above are summarized in table 2. Even M=1 has useful results, although they are between 0.677 and 0.921 with M=8 estimated correlations.
This is the reason why even the GEE-model with compound symmetry gives good estimators with a correlation of 0.813. Figure 2 illustrated mean value courses in Na-palmitate groups in the late phase. The corresponding REML-estimators are shown by dotted horizontal straight lines.

Table 2. Models of heart rate in the late reperfusion phase.

parameter (naive z, robust z)	intercept	Na-palmitate
REML	72.03 (19.11)	16.50 (3.10)
ML	72.02 (19.68)	16.52 (3.19)
GEE (M=8)	72.08 (19.98 17.53)	16.51 (3.24 3.27)
GEE (M=1)	71.09 (29.88 16.12)	18.18 (5.40 3.40)
GEE exchangeable	72.41 (19.82 17.29)	16.06 (3.11 3.11)

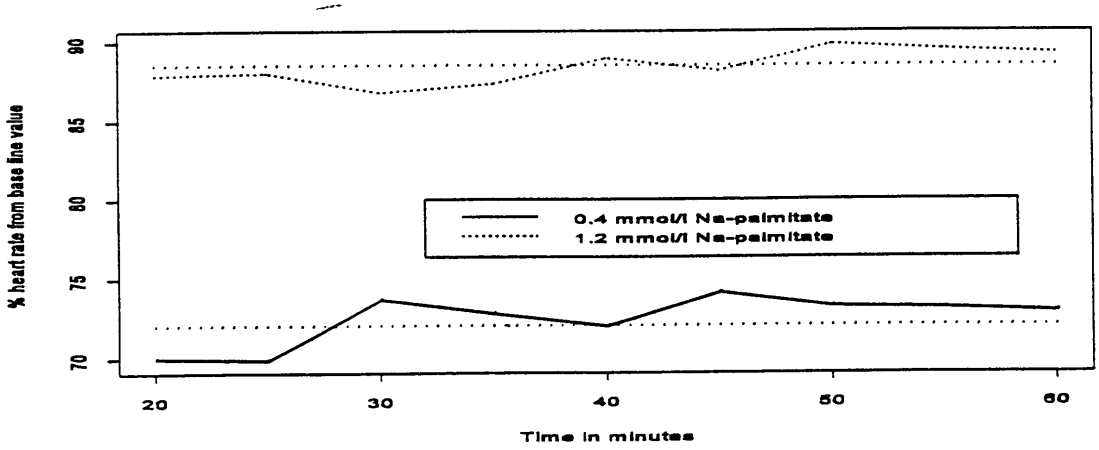

Figure 2. Mean values of heart rate grouped by Na-palmitate and REML estimations (dotted lines) in the late reperfusion phase.

2.2. GEE Models for Degree of Tachycardia

For registration and quantification of disturbances of the cardiac rhythm in the reperfusion phase an arrythmia score was developed which subdivided tachycardiac and bradycardiac irregularities of the rhythm into 4 different degrees in each case. Here, only the change of tachycardia-scores of the early phase of reperfusion can be illustrated. It is only possible to monitor the score in four degrees. Thus, a GEE-model with Poisson distribution and logarithmic link function was useful, i.e.
$\ln(E(y_i)) = \mathbf{X}_i\beta$ or $E(y_i) = \exp(\mathbf{X}_i\beta)$.
In the early phase, the model with all three factors showed that both the main effect and all interaction effects with Na-palmitate have no significant impact on the tachycardia. Another term exp(-time) was taken because models with L-carnitine and time did not show sufficient adjustment.

Table 3. Models for degree of tachycardia in the early reperfusion phase.

parameter (naive z-robust z)	intercept	L-carnitine	time	exp(-time)	L-carnitine*time	L-carnitine*exp(-time)
GEE (M=10)	0.923 (8.58 10.05)	0.177 (1.20 1.14)	-0.049 (-4.57 -6.71)	0.749 (2.56 2.40)	-0.003 (-0.18 -0.26)	-0.101 (-0.25 -0.22)
GEE (M=11)	0.944 (10.64 12.87)	0.149 (1.62 2.03)	-0.050 (-6.86 -10.23)	0.694 (3.48 2.94)		

The results are given in the first row of table 3. This table was done analogously to table 1 (compare chapter 2.1.) Interaction terms can be left out because the last two columns of table 3 which contain interaction terms have only very low z-values. The solution of the developed model

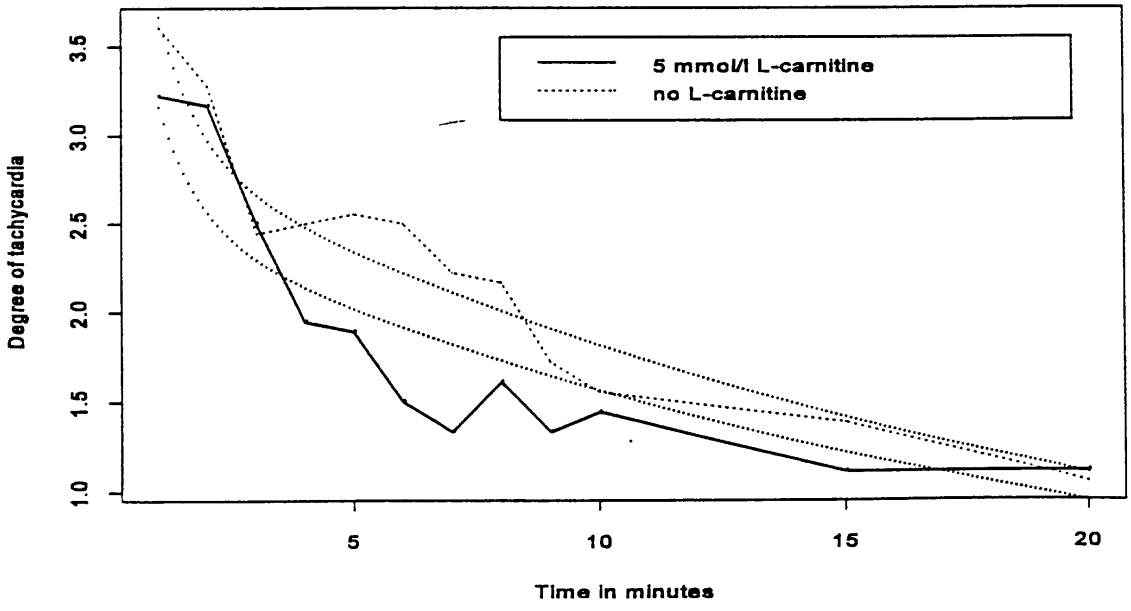

Figure 3. Mean degree of tachycardia grouped by L-carnitine and GEE estimations (dotted lines) in the early reperfusion phase.

is given in the second row of table 3. In this case, at three parameters robust z-vaules are higher than naive z-values in respect to the amount.
Besides a statistically determined influence of both time parameters robust z-values also show an impact on L-carnitine at a niveau of 5%. That means, the tachycardia normalized faster in the group with L-carnitine. The corresponding graph of the estimated GEE-model are shown in Figure 3.

y_c = exp(0.944 - 0.050*time + 0.694*exp(-time)) with L-carnitine and
y_n = exp(0.944 + 0.149 - 0.050*time + 0.694*exp(-time)) without L-carnitine.

In the late phase of reperfusion non of the investigated factors had a significant influence on the development of the tachycardia.

References

Arminger, G. (1995). Specification and Estimation of Mean Structures: Regression Models. In G. Arminger, C.C. Clogg & M.E. Sobel (Eds.), *Handbook of Statistical Modeling for the Social and Behavioral Sciences* (pp. 77-183), New York and London: Plenum Press.

Diggle, P.J., Liang, K.Y.,& Zeger, S.L. (1994). *Analysis of Longitudinal Data.* Oxford: Oxford University Press.

Gourieroux, C.& Monfort, A. (1993). Pseudo-Likelihood Methods. In G. Maddala, C. Rao & H. Vinod (Eds.), *Handbook of Statistics, volume 11* (pp. 335-362). Amsterdam: Elvesier.

Gourioux, C., Monfort, A.,& Trognon, A. (1984). Pseudo maximum likelihood methods: theory, *Economica,* 52, pp. 682-700.

Liang, K.Y.,& Zeger, S. (1986). Longitudinal data analysis using generalized linear models. *Biometrica*, 73, pp.13-22.

Ludwig, F.,& Tutz, G. (1994). *Multivariate statistical modelling based on generalized linear models.* New York, USA: Springer-Verlag.

Schuster, E., Kaltenhäuser, S.,& Häntzschel, H. Longitudinale Datenanalyse in der Auswertung einer Studie zur "Pathogenese der Rheumatoiden Arthritis" (in press).

Ziegler, A.,& Arminger, G. (1996). Parameter estimation and regression diagnostics using generalized estimating equations. In F. Faulbaum & W. Bandilla (Eds), *SoftStat '95 Advances in Statistical Software 5* (pp. 229-237). Stuttgart: Lucius & Lucius.

Ziegler, A., Kastner, Ch., Grömping, U.,& Blettner, M. (1996). Die Generalized Estimating Equations: Herleitung und Anwendung, *Informatik, Biometrie und Epidemiologie in Medizin und Biologie* **27** (2), 69-91.