A Joint model of Longitudinal and Survival data with Covariates measured with Error

Azeez Adeboye*,
Prof. Y. Qin,
Dr. J. Ndege,
Dr. R. Mutambayi

Department of Statistics, University of Fort Hare, South Africa

61 Annual Conference of the South Africa Statistical Association, Nelson Mandela University, Port Elizabeth, South Africa,
27-29 November, 2019

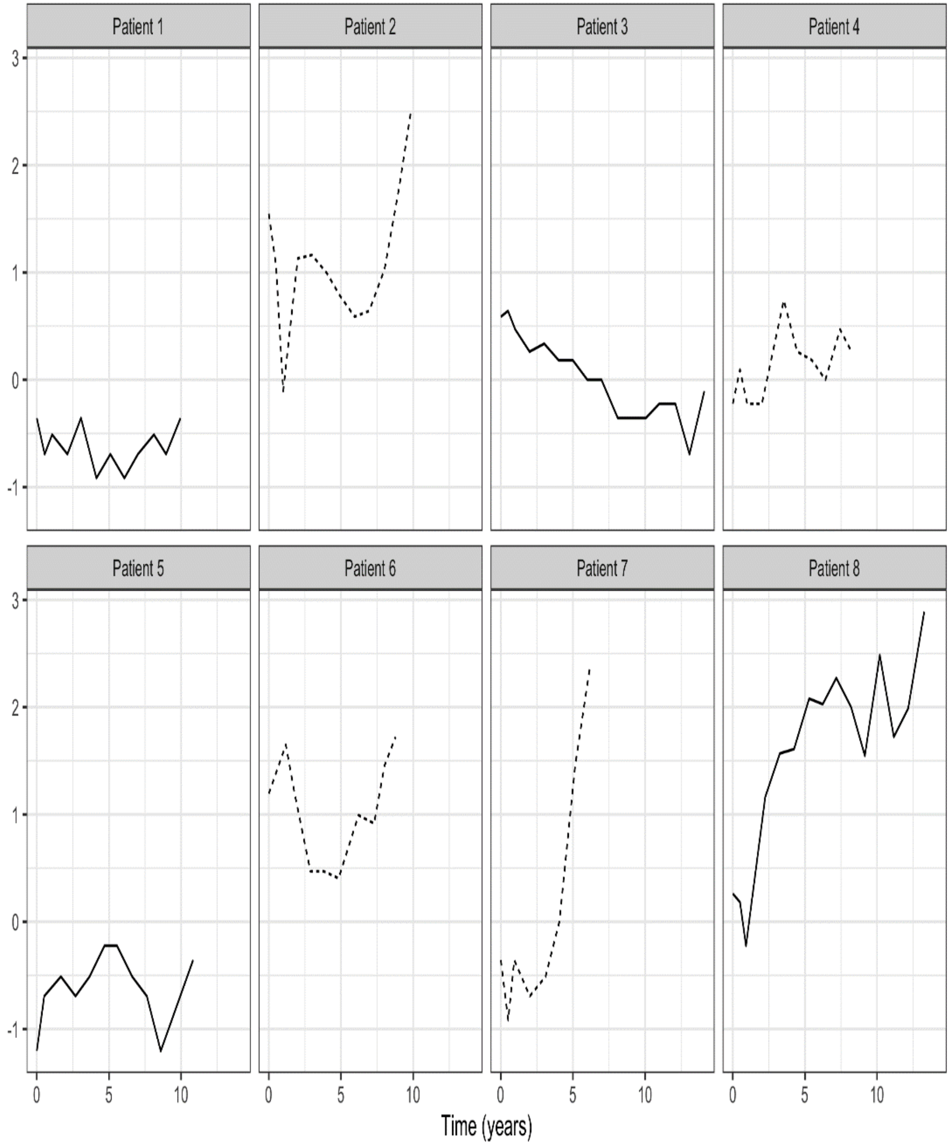
University of Fort Hare
Together in Excellence

Content

1. Introduction
2. Motivating data application
3. Joint Model specification
4. Real data application
5. Conclusion

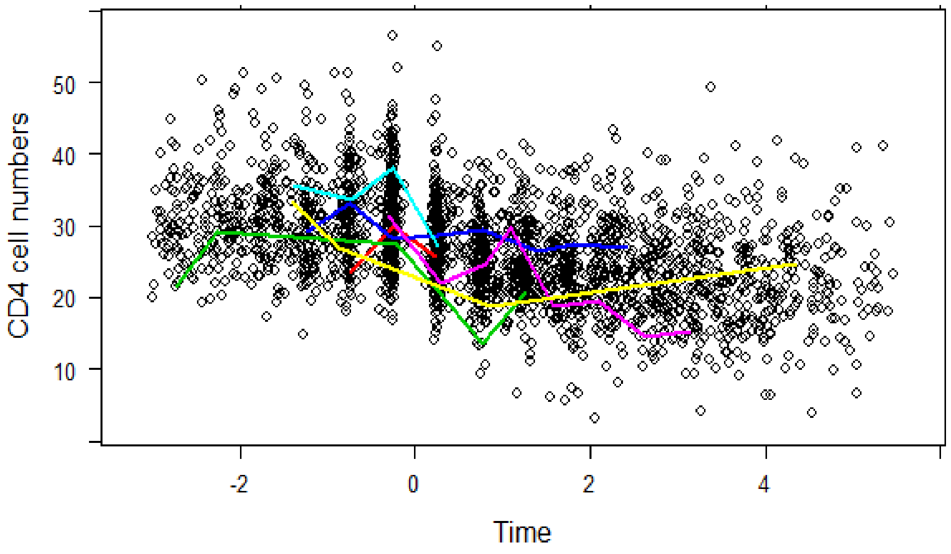
1. Introduction

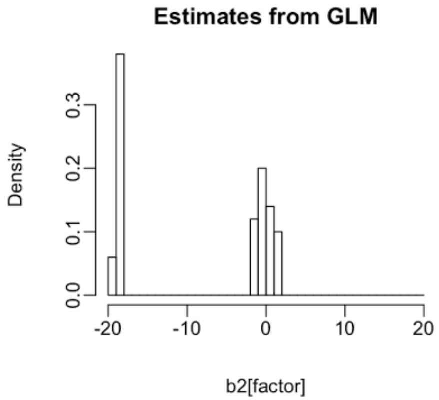
- It commonly in clinical research to collect information both on longitudinal measurements and a time-to-event for each participant.
- Longitudinal process that characterize the relationship between the survival and the longitudinal process and other covariates.
- Longitudinal process follows a linear mixed effects model and that the survival process through a proportional hazards model.
- Time-dependent covariates are measured intermittently and often with error
- Some research methods have been advocated to reduce the bias such as Regression calibration (RC) and corrected score (CS) approach
- Likelihood-based approaches are consistently used and it is efficient with normality assumption on both the underlying random effects and the measurement error.

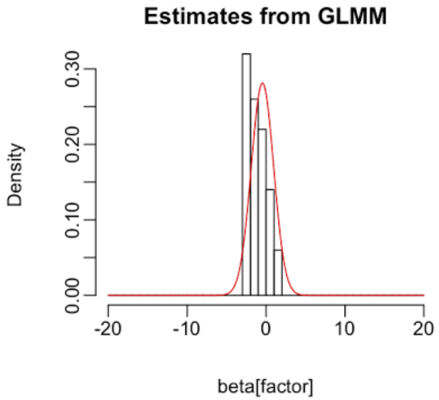


University of Fort Hare
Together in Excellence

- Normality assumption may be violated due to measurement error (within-subject random error) when jointly model the longitudinal and survival data.
- Also, misspecification of the covariance structure could lead to erroneous statistical inference.
- Violation of assumption of mutual independence of within-subject errors, could lead to bias estimation
- Our primary interest is to characterize the relationship between survival and the covariates, which includes the covariate consisting of current value of the relative time-dependent longitudinal process and other covariates which may or may not be time-dependent.
- The current value of longitudinal process is treated as one of important covariates of the survival regression model.
- In this study, sufficient statistics is used for the conditional score estimation and conditional estimating equations for the parameter inference



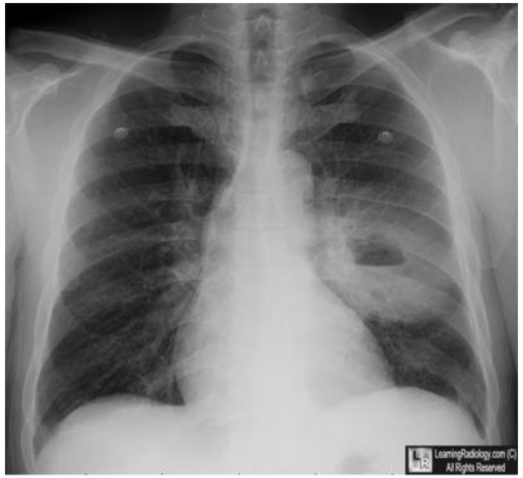

Estimates from GLM

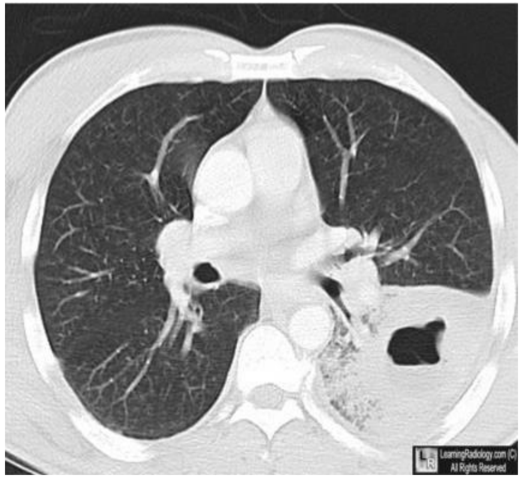

Estimates from GLMM

University of Fort Hare
Together in Excellence

2. Motivating data

- The link between tuberculosis (TB) and Renal function has been known for more than 40 years, but the interaction between these 2 diseases is still poorly understood.
- Dialysis and renal transplant patients appear to be at a higher risk of TB, in part related to immunosuppression along with socioeconomic, demographic, and comorbid factors.
- Meanwhile, TB screening and diagnostic test performance is suboptimal and limited evidence to guide protocols.
- Given the increasing prevalence of renal function in TB endemic areas have significant public health implications, especially in low- to middle-income countries
- To address this situation, a clear understanding of the relationship between these 2 conditions needs to be established, and consistent, evidence-based screening and treatment guidelines need to be developed.

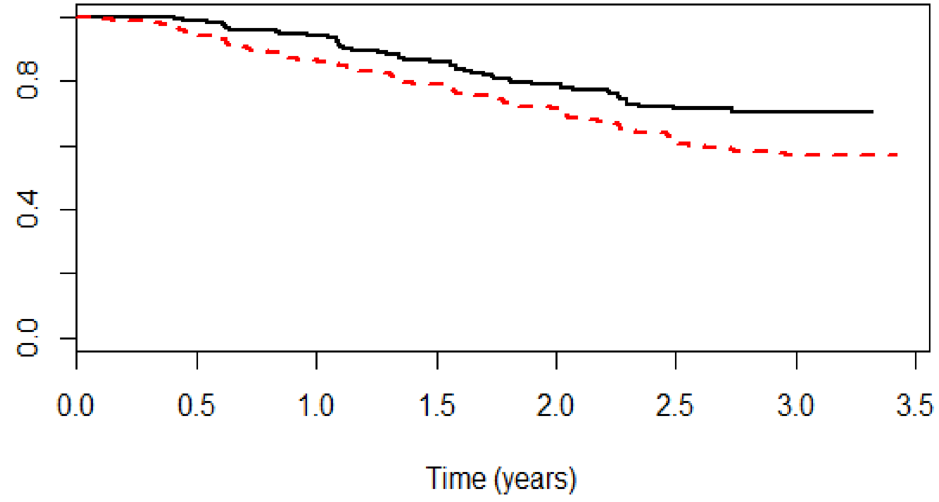


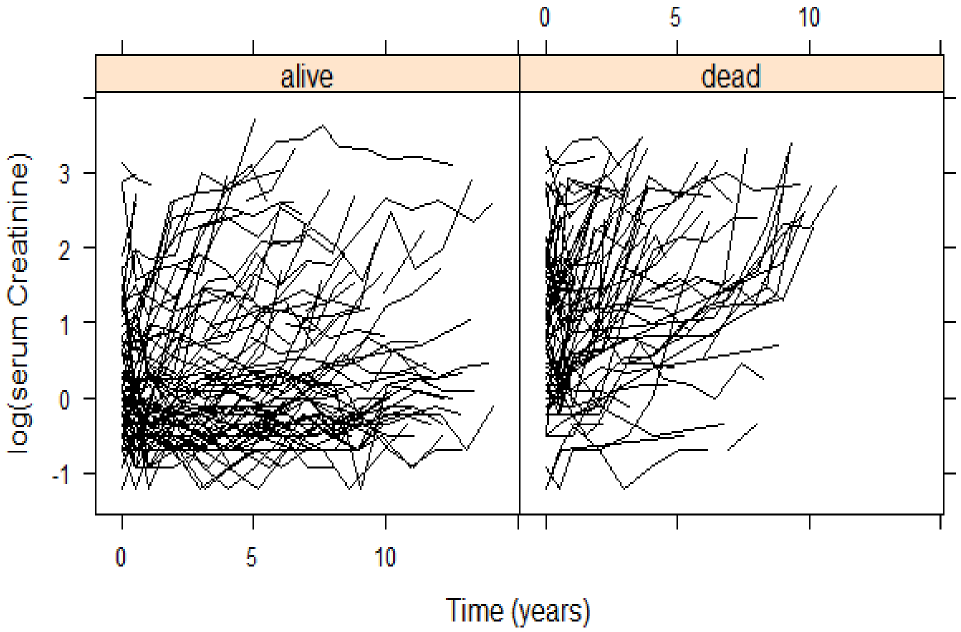


Motivating data cont.

- Dataset from Longitudinal Cohort TB-ESRD Clinical Study (Grey Hospital)

- 2987 patients were observed in the study
- N= 612 TB patients with impaired renal function was sampled from 2008 - 2018.
- Each patient assigned to a treatment regime to compare the treatment effects and the survival end-point with a CD4 count ≥50% decrease
- Time-progression to survival end-point and covariates-age, weight, CD4 count, drug on each patient about every 3 months.
- Epidemiology: Increased incidence and prevalence of TB in ESRD and dialysis patients, rate of OIs and mortality
- CD4 biomarker is a good surrogate for treatment effect but may be subjected to a considerable measurement error.
- Covariance within-subject error in joint model of longitudinal-survival data of $log_{10}CD4$ count of TB patients with impaired renal





University of Fort Hare
Together in Excellence

3. Joint Model specification

For simplicity, we assume a single longitudinal process, but generalization to multiple cases should be feasible and without much difficulties. The longitudinal data are assumed to follow a linear mixed effects model with time polynomial function, but the generalization to that with additional baseline covariates is straightforward:

$$W_i(t_{ij}) = X_i(t_{ij}) + \varepsilon_i(t_{ij})$$

$$W_i(t_{ij}) = b_{io} + b_{i1}t_{ij} + b_{i2}t_{ij}^2 + .. + b_{iq}t_{ij}^q + \varepsilon_{ij}, \ i = 1, 2, .., n; \quad j = 1, 2, .., m_i,$$

Given $t_i = (t_{i1}, .., t_{ini})'$, the within-subject error $\varepsilon_i = (\varepsilon_{i1}, .., \varepsilon_{ini})'$ are assumed to be normally distributed with mean zero and covariance Σ_i, and they are independent of random effects $\alpha_i = (\alpha_{i1}, .., \alpha_{ini})'$.

To specify the interrelationship, we assumed the hazard of failure is related to covariates Xi(t) and Zi through a Cox proportional hazards model, that is,

$$\lambda_i(t) = \lim_{dt \to 0} \frac{P\{t \le T_i < t + dt \mid T_i \ge t, \bar{Y}_i(t), Z_i(t)\}}{dt} = \lambda_0(t) \exp\{\gamma Y_i(t) + \eta^T Z_i(t)\}$$

In this study, we would like to investigate the influence of violating the assumption of **mutual independence of error** on the inference for survival parameters and no distribution assumption is made on random effects.

Conditional Score Estimator

Tsiatis and Davidian (2001) proposed a conditional score (CS) estimator in respect of unknown regression parameters in the proportional hazards model when the underlying time-dependent covariate follows a linear mixed model.

$\hat{Z}_i(t)_{ols}$ represents the OLS estimate of $Z_i(t)$

The longitudinal for all subject *i* at time *t* i.e.

$$Z_i(t)_{gls} = \bar{t}^T (D_{i,t}^T D_{i,t})^{-1} D_{i,t}^T W_{i,t}$$

Where vector $\bar{t} = (1, t, .., t^p)'$, $D_{i,t}$ is the design matrix and $W_{i,t}$ is the longitudinal measurement.

A sufficient statistic of Xi(t) is proposed, provided the parameters are known:

$$S_i(t,\gamma,\Sigma_{i,t}) = \gamma\sigma^2_{\hat{X}_i(t)_{ols}} dN_i(t) + \hat{Z}_i(t)_{ols},$$

Given the sufficient statistic, the conditional hazard function turns out to be

$$\lambda_1(t \mid S_i(t,\gamma) = \lambda_0(t)\exp\left(\gamma S_i(t,\gamma) - \frac{1}{2}\gamma^2\sigma^2_{\hat{X}_i(t)_{ols}} + \eta^T Z_i\right)$$

Unbiased estimating equations for (γ, η)′ are proposed based on the conditional hazard

$$\sum_{i=1}^{n} \int \begin{pmatrix} S_i(t,\gamma) \\ Z_i \end{pmatrix} \left(dN_i(t) - \lambda_1(t \mid S_i(t,\gamma))dt\right) = 0$$

Generalized Conditional Score Estimator

We propose generalized least squares estimator for the unknown underlying covariate Xi(t), which has smaller variation than that of OLS estimator.

$\hat{Z}_i(t)_{gls}$ represents the GLS estimate of $Z_i(t)$

Measurements taken by using all the longitudinal data up to and including time t, that is, we have

$$\hat{Z}_i(t)_{gls} = \bar{t}^T (D_{i,t}^T \Sigma_{i,t}^{-1} D_{i,t})^{-1} D_{i,t}^T \Sigma_{i,t}^{-1} W_{i,t}$$

A sufficient statistic proposed for GSC parameters as:

$$S_i(t,\gamma,\Sigma_{i,t}) = \gamma\sigma^2_{\hat{X}_i(t)_{gls}} dN_i(t) + \hat{Z}_i(t)_{gls},$$

conditional intensity process is defined as

$$\lim_{dt\to 0} dt^{-1} P(dN_i(t) = 1 \mid S_i(t,\gamma,\Sigma_{i,t}), Y_i(t)$$

$$= \lambda_0(t)\exp\left(\gamma S_i(t,\gamma,\Sigma_{i,t}) - \frac{1}{2}\gamma^2\sigma^2_{\hat{X}_i(t)_{gls}} + \eta^T Z_i\right) Y_i(t),$$

the estimating equations can be expressed as

$$V_1(\psi_k) = \sum_{i=1}^{n} \int \left\{ \Upsilon_i(t,\psi_k) - \frac{E_{0i}(t,\gamma,\eta,\Sigma_{i,t})}{E_0(t,\gamma,\eta,\Sigma_t)} \right\} dN_i(t) = 0.$$

4. Real Data Application

- We excluded the treatment and BMI variable from the final proportional hazard model (not significant)
- $log_{10}CD4$, age and gender were included in the final PH model
- GCS approaches, especially the one with Cholesky decomposition, provide stronger coefficients of $log_{10}CD4$ than that of CS approach
- The estimators of GCS with AR(1) covariance are almost the same as CS approach, but the coefficient of $log_{10}CD4$ of GCS approach with modified Cholesky decomposition of covariance are much stronger.

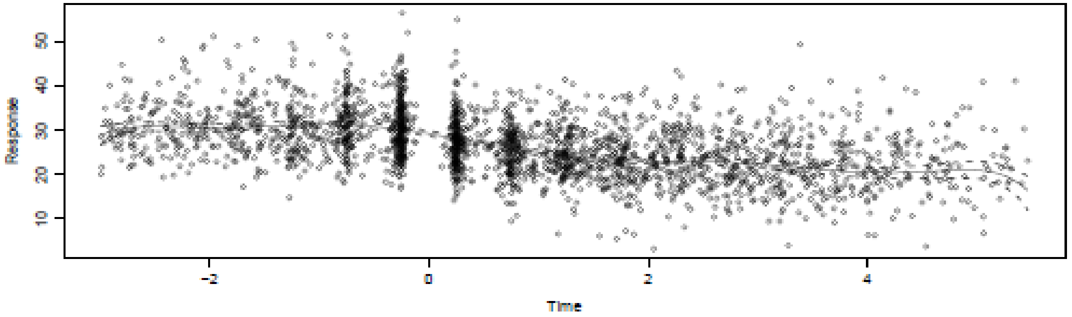
Table 6: Statistical Inference for survival analysis of real data results for the CS and GCS approaches

	CS		**GCS-AR(1)**		**GCS-CD**
Covariates	**MEst**	**MLE**	**MEst**	**MLE**	**MLE**
$log_{10}CD4$	-3.124 (0.231)	-2.052 (0.228)	-2.282 (0.221)	-2.269 (0.211)	-2.442 (0.255)*
Age	0.036 (0.009)	0.036 (0.009)	0.036 (0.009)	0.036 (0.009)	0.036 (0.009)*
Gender	0.418 (0.213)	0.419 (0.212)	0.426 (0.206)	0.426 (0.205)	0.314 (0.292)

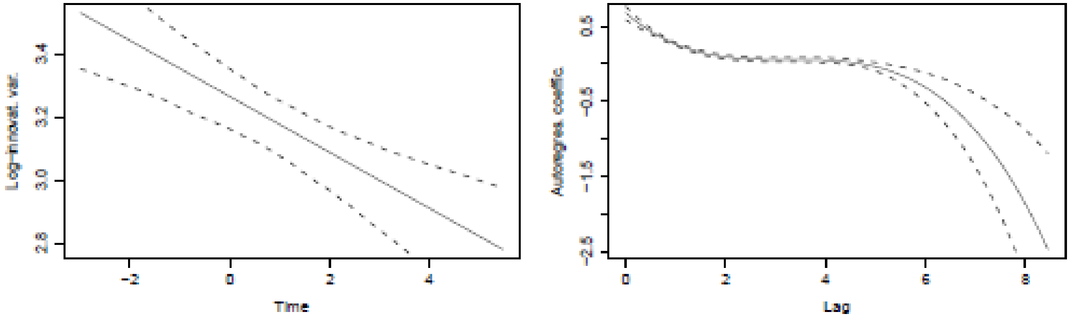
Values inside the parenthesis are the standard deviation, * indicates the significance of covariate of 95% Wald confidence interval.

- Age has significant positive effect on the hazard rate, that is, the older the TB patients with impaired renal function, the higher possibility of experiencing death.
- GCS approach of MEst tends to slightly overestimate the survival parameters, but GCS approach of MLE has a very good statistical performance and provides virtually unbiased estimates, and the coverage probabilities also show the validity of approach.

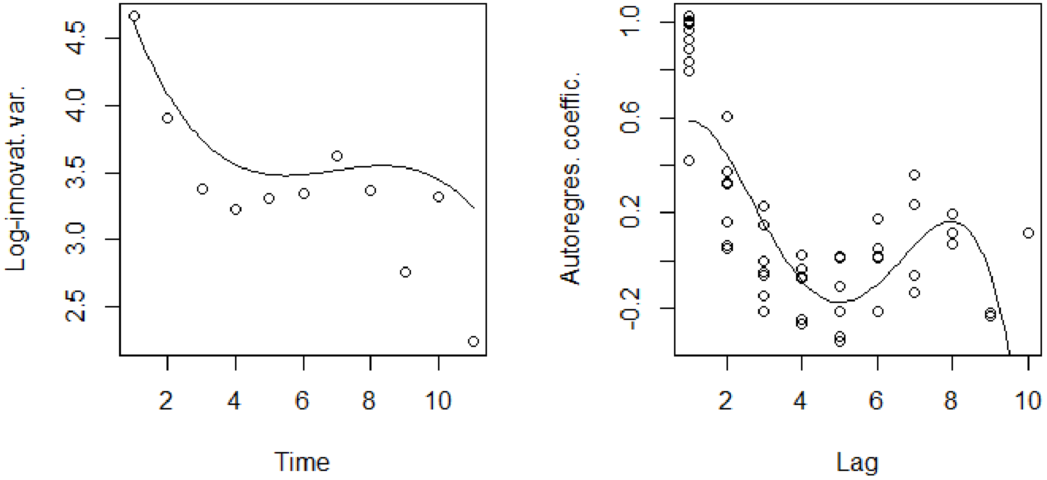
The MLE approach for modified Cholesky decomposition based on the covariance within-subject was fitted in a similar way to compare the fitted models using the log-likelihood of the estimates. The covariance matrix of the fitted model produced the fitted curves with 95% confidence interval using sandwich bootstrap method.



The mean fitted curve with log-variance, AR(1) and the 95% confidence interval in Figure 2 showed that there is decreasing association of log-variance fitted with respect to time and curve shape of AR(1) fitted coefficient with time lag.



The fitted curve for the real data shows that the fitted polynomial function curvature shape is captured well and indicate a good fit for autoregressive coefficients (AR(1)) in examining the AR(1) coefficient versus time lag between the measurements and the fitted curve



5. Conclusion

- GCS-CD (modified Cholesky decomposition) suggests that the capturing of the covariance within-subject are more accurate using Cholesky decomposition approach than simple AR(1).
- It should be worth utilising classical likelihood-based approach to further investigate the impact on survival analysis if the assumption of independence of random errors is violated.

Thanks for listening

References

Aalen, Odd. 1978. “Nonparametric Inference for a Family of Counting Processes.” *The Annals of Statistics* 6(4):701–26.

Anastasios A . Tsiatis. 1990. “Estimating Regression Parameters Using Linear Rank Tests for Censored Data.” *The Annals of Statistics* 18(1):354–72.

Crowther, Michael J., Keith R. Abrams, and Paul C. Lambert. 2012. “Flexible Parametric Joint Modelling of Longitudinal and Survival Data.” *Statistics in Medicine* 31(30):4456–71.

Rizopoulos, Dimitris, Laura A. Hatfield, Bradley P. Carlin, and Johanna J. M. Takkenberg. 2014. “Combining Dynamic Predictions From Joint Models for Longitudinal and Time-to-Event Data Using Bayesian Model Averaging.” *Journal of the American Statistical Association* 109(508):1385–97.

Tsiatis, Anastasios A. and Marie Davidian. 2001. “A Semiparametric Estimator for the Proportional Hazards Model with Longitudinal Covariates Measured with Error.” *Biometrika* 88(2):447–58.

Tsiatis, Anastasios A. and Marie Davidian. 2004a. “Joint Modelling of Longitudinal and Time-to-Event Data: An Overview.” *Statistica Sinica* 14(3):809–34.

Tsiatis, Anastasios A. and Marie Davidian. 2004b. “Joint Modelling of Longitudinal and Time to Event Data: An Overview.” *Statistica Sinica* 14(3):809–34.