Are Cox Regression Models a Valuable Tool for Social Stratification Research on Health?

Alessandro Procopio[1] Robin Samuel[2]

[1] University of Luxembourg
alessandro.procopio@uni.lu

[2] University of Luxembourg
robin.samuel@uni.lu

ESRA Conference
17/07/2019

UNIVERSITY OF LUXEMBOURG
Integrative Research Unit on Social and Individual Development (INSIDE)

Introduction

The rise of biological and social data

Recent Social Science studies include biomarkers measurements to understand social stratification processes on health outcome(Harris and Schorpp, 2018).

At the empirical level, social researchers can rely on an increasing number of *biosocial surveys* (National Research Council, 2008).

Research Question

How to analyze these different types of data?
How to exploit the information provided by these types of surveys?

Aim of the Study

Present a new specification of the Cox regression model when dealing with repeated measurements of the same individuals.

Research Strategy

1. Theory-based Monte Carlo Simulation on the Cox regression model with panel data.
2. Analyze how the model behaves in the context of unobserved heterogeneity, commmon issue in the Social Sciences.
3. Analyze the misspecification of the time modelling of the biomarker trajectory on the health outcome.

Time-Varying Cox Regression Approach

The classical approach

- The traditional approach to analyze a time-to-event response variable and a covariate measured over time is to include it as a time-dependent explanatory factor in the model (such as the biomarker trajectory).
- The Cox regression with panel data assumes, however, that the time-varying covariate (the biomarker) does not change until we get a new measurement. A strong assumption.
- Chen et al. (2004) demonstrated that the Cox regression with time-varying covariates returns biased estimates when the researcher is interested in causal effects of a determined treatment.

Proposed solutions

- In a first phase, the Two-Stage Model (Wulfsohn and Tsisatis, 1997) has been implemented. It consists of:
  - A running a mixed effect model
  - B predict the trajectory of the biomarker
  - C include the prediction to a survival model
- Currently, the model we want to propose to analyze social and biological data is the joint modelling approach.
- The main difference between them is that in the joint modelling the biomarker trajectory is not included as a prediction of the mixed effect model.
- But the longitudinal and the survival models are estimated simultaneously (Rizopoulos et al., 2008; Rizopoulos, 2014).

The model of interest

Joint modeling

Recently, the statistical literature improved the Two-Stage Model in a way that the mixed and the survival submodels are estimated *simultaneously*.

Let's take a look at the two submodels:

Random Intercept-Slope Submodel

$$m_i(t) = \mathbf{X_i}^T(t)\beta + \mathbf{Z_i}^T(t)\mathbf{b_i} + \epsilon_i(t)$$

Survival Submodel

$$h(t)^1 = h_{(0)}(t)exp[\beta\mathbf{X_i} + \alpha m_i(t)]$$

[1] $h(t) = \lim_{\delta \to \infty} \frac{P(t \leq T < t+\delta | T > t)}{\delta}$

Monte Carlo Simulation of the Joint Modelling

- Assume that a researcher conducts a study on a sample of 250 respondents over ten years. Let imagine that we have collected biological data through a biosocial survey for a defined m biomarker.
- Let imagine that the biomarker, let say the allostatic load, increases with age (young people manage stress levels better than the older) and this relationship is non-linear, it has a quadratic pattern.
- Assume that the socioeconomic position influences the level of allostatic load. For example, the rich have the resources to manage stress better than the poor.

Monte Carlo sets

The time scale

- In the statistical literature, it is known the Cox regression is sensible to the time scale specification (Thiébaut and Bénichou, 2004; empirical suggestion taken from Crowther et al., 2016).
- What kind of bias would we find in the estimates if we assume that the longitudinal trajectory of the biomarker is a linear function with the follow-up time, while it has a quadratic shape in reality?

Frailty/Heterogeneity

- In the epidemiological and social science literature, between-group frailties are increasingly taken into account in the data analysis process (for an empirical work: Zarulli et al., 2013).
- What kind of bias would we find in the estimates if we do not take into account the socioeconomic position?

Data Generation Mechanism

Longitudinal Model

$$m_i = .2 + .5(t) + .02(t)^2 + .085 * age + 0.1 * ses + e_{ij}$$

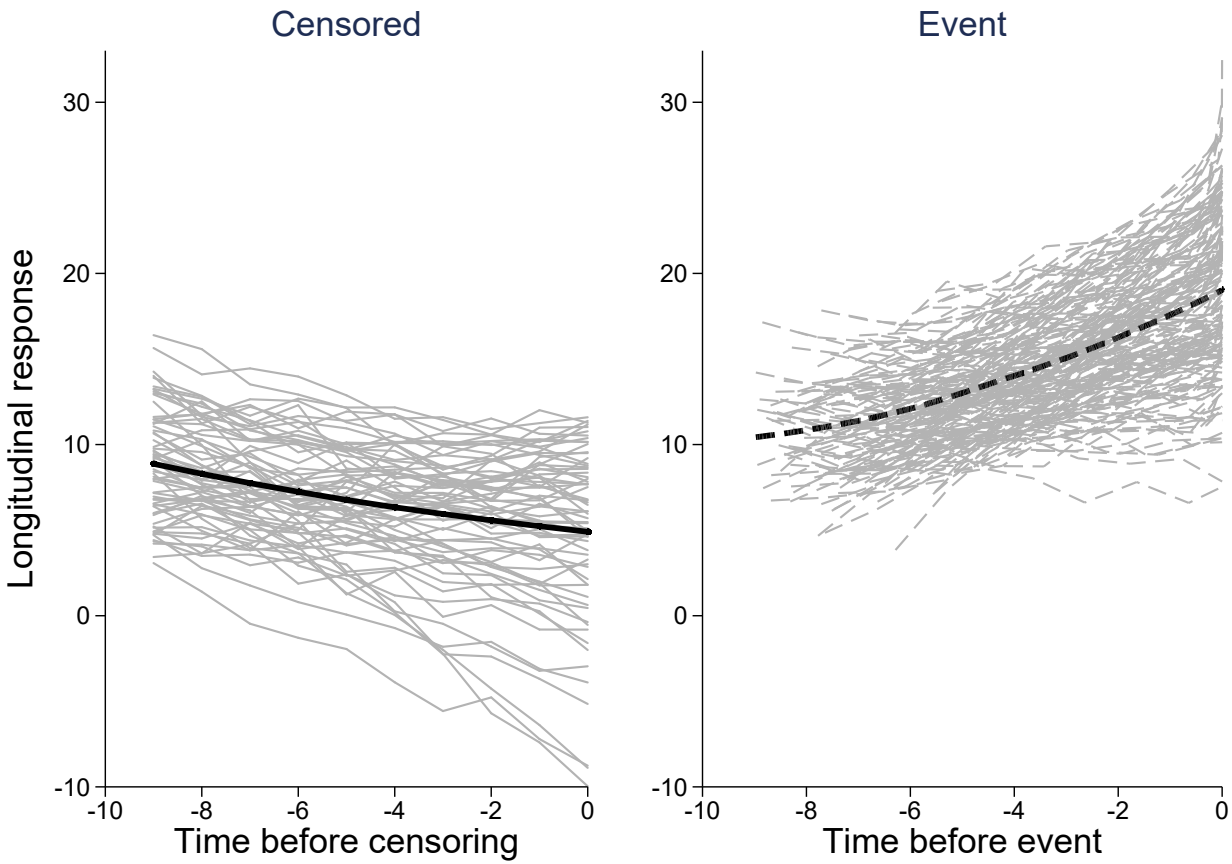
$$e_{ij} = \mathcal{N}(0, \Sigma) = \Sigma = \begin{bmatrix} \sigma^2_{00} & \\ \sigma^2_{01} & \sigma^2_{11} \end{bmatrix} \begin{cases} \sigma^2_{00} = 2.1 \\ \sigma^2_{11} = 1.07 \\ \sigma^2_{01} = .3 \end{cases}$$

Gompertz-Cox parametric model

$$\begin{aligned} h(t \mid \beta_i) = & \, exp(-16) + exp(1.5)t \\ & + exp[.40(\beta_{0i} + \beta_{1i}t) \\ & + .02(t)^2 + .085 * age + 0.1 * ses] \end{aligned}$$

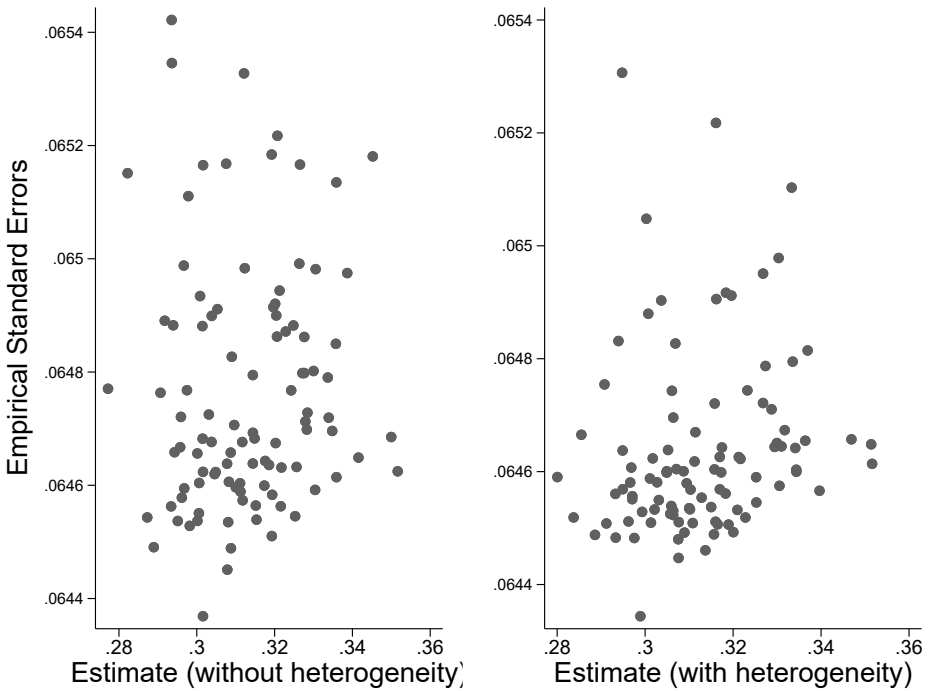
Baseline hazard function taken from Bender et al. (2005). Baseline mortality rate λ reparametrized as: $\lambda = exp(\gamma^*)$, see Van den Hout and Muniz-Terrera (2016)

Graphical visualization of the simulated data

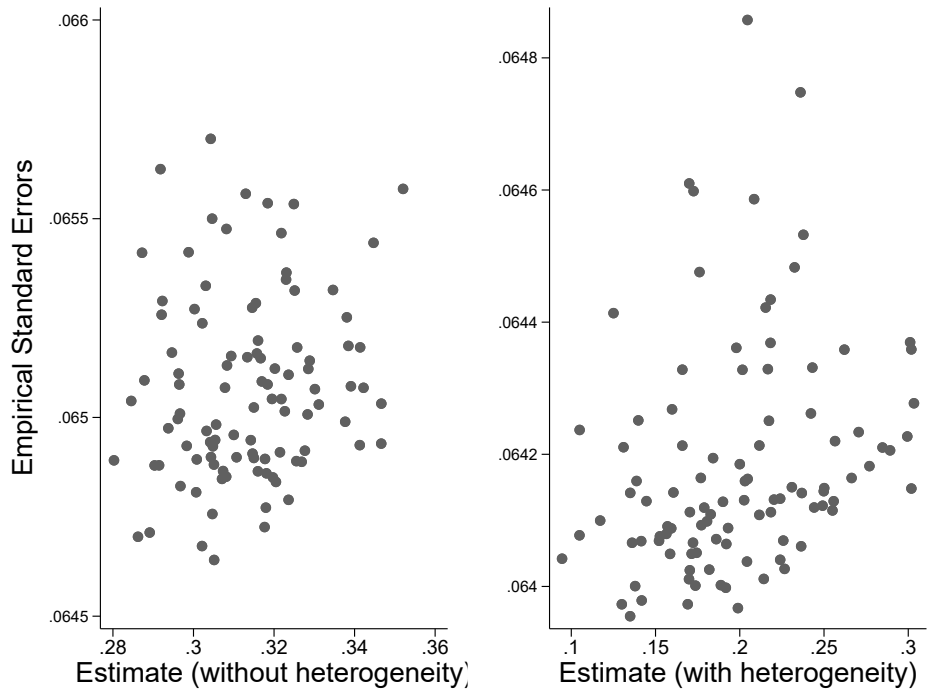


Polynomial Trajectory: Correlation coefficient

ρ when U.H.=0.1

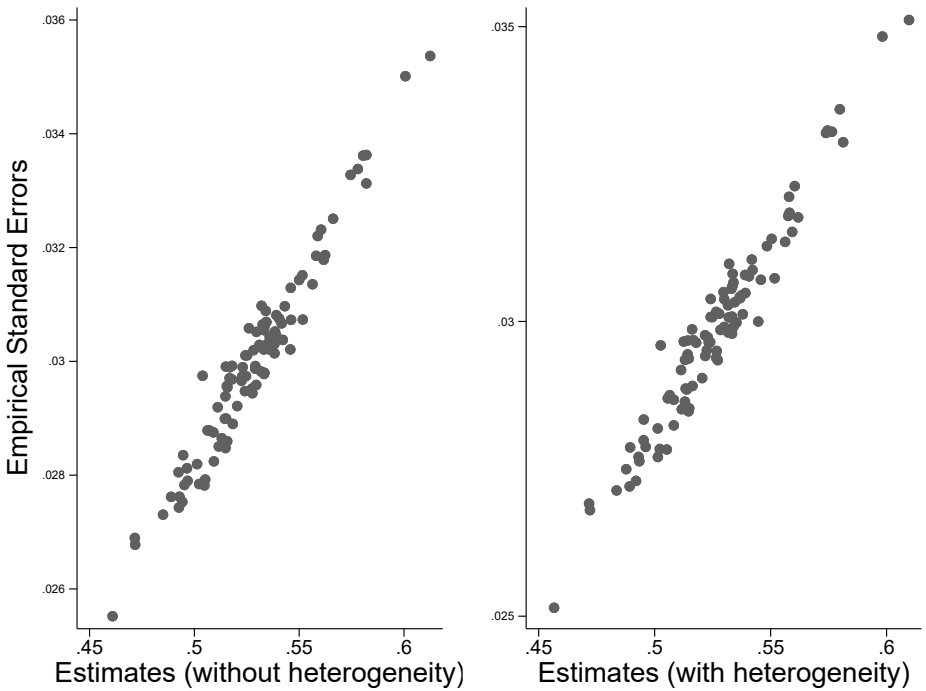


ρ when U.H.=3

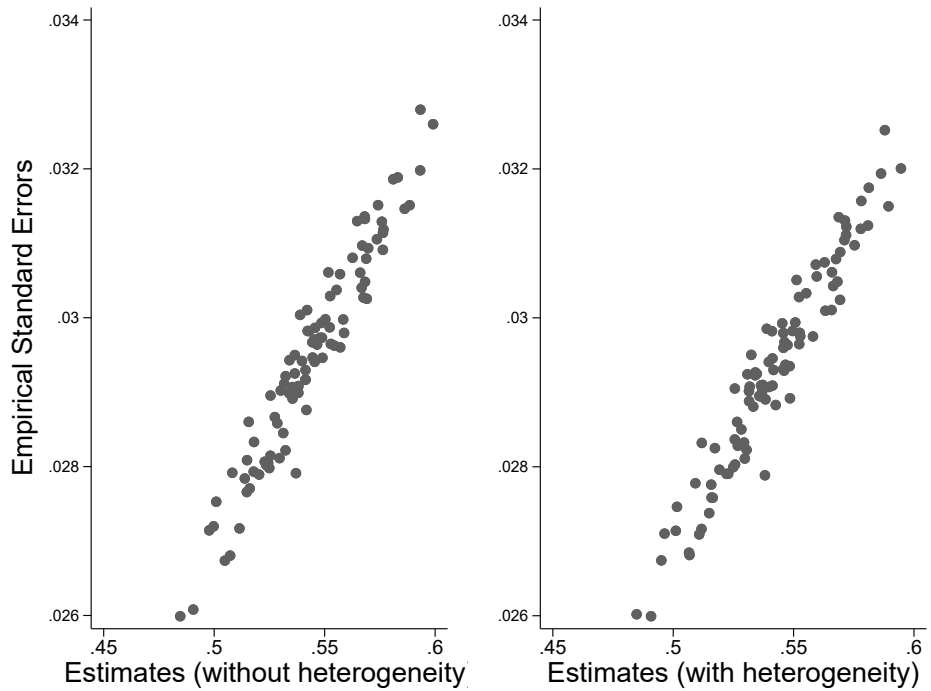


And the association parameter

α when U.H.=0.1

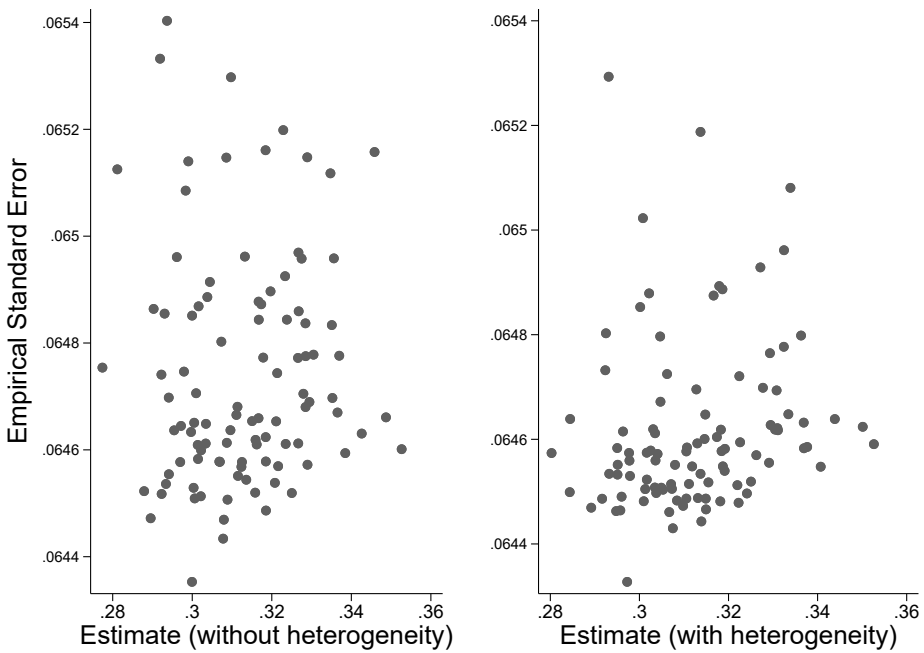


α when U.H.=3

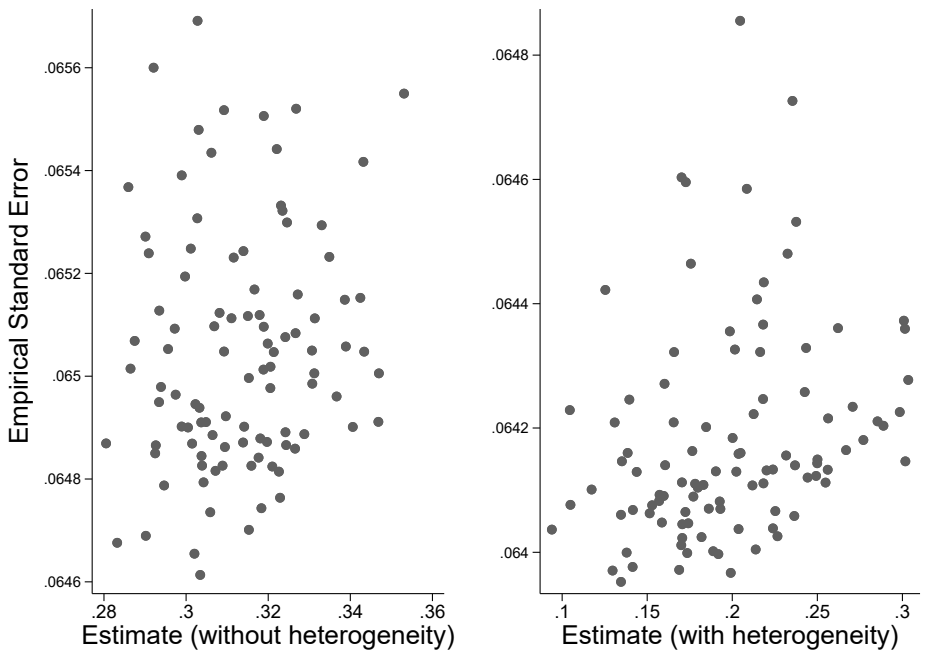


Linear trajectory: correlation coefficient

ρ when U.H.=0.1

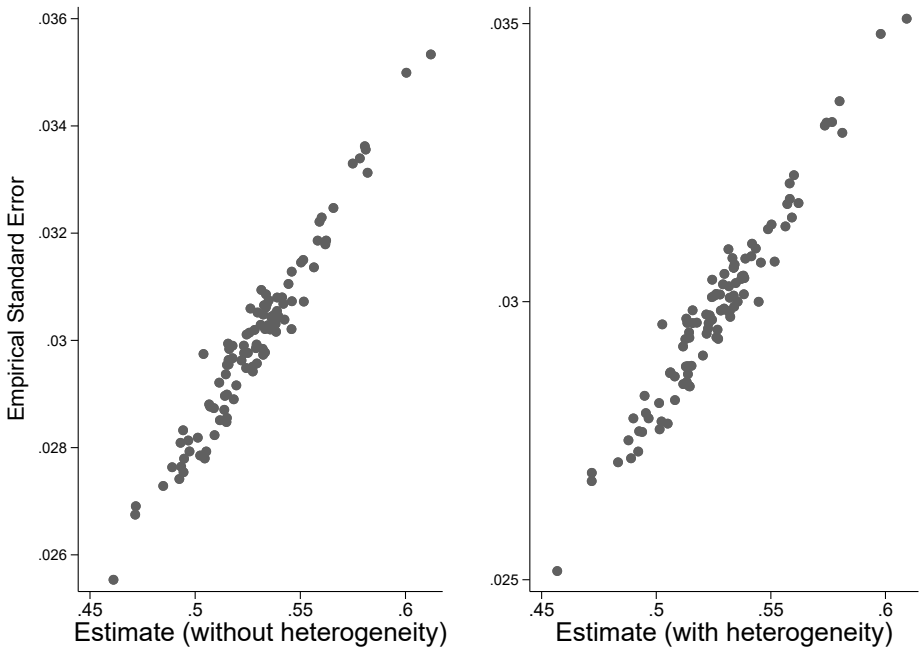


ρ when U.H.=3

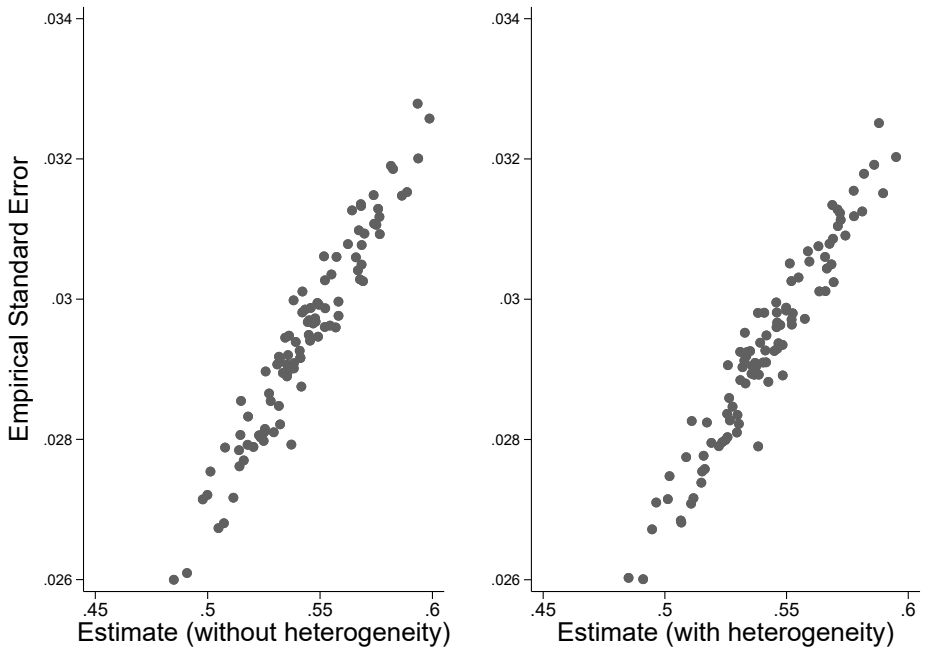


And the association parameter

α when U.H.=0.1



α when U.H.=3



Conclusions

- The association parameter ρ that captures the correlation between the fixed and random effects is on average around the true model.
- However, stability toward the true parameter over the replications present higher variance and bigger empirical standard errors.
- The α parameter, that captures the association between the biomarker trajectory and survival chances, presents a smoother linear pattern than the longitudinal ρ.
- That means that the empirical standard errors are much narrower to the estimate.
- Moreover, it is rather "robust" to unobserved heterogeneity.
- The only problematic set, coherently with previous studies arises when we misspecify the time of measurements **and** unobserved heterogeneity is present. Specifically, the correlation coefficients between the random and the fixed effects are downwardly biased in the longitudinal submodel.

Thank you for your attention

References I

Bender, R., Augustin, T., and Blettner, M. (2005). Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine*, 24(11):1713–1723.

Chen, F., Paxton, P., Bollen, K. A., Curran, P. J., and Kirby, J. (2004). Monte Carlo Experiments: Design and Implementation. *Structural Equation Modeling: A Multidisciplinary Journal*, 8(2):287–312.

Crowther, M. J., Andersson, T. M., Lambert, P. C., Abrams, K. R., and Humphreys, K. (2016). Joint modelling of longitudinal and survival data: Incorporating delayed entry and an assessment of model misspecification. *Statistics in Medicine*, 35(7):1193–1209.

Harris, K. M. and Schorpp, K. M. (2018). Integrating Biomarkers in Social Stratification and Health Research. *Annual Review of Sociology*, 44(1):361–386.

National Research Council (2008). *Biosocial Surveys*. The National Academies Press, Washington, DC.

Rizopoulos, D. (2014). Joint modeling of longitudinal and time-to-event data. *Handbook of Missing Data Methodology*, pages 117–136.

References II

Rizopoulos, D., Verbeke, G., and Molenberghs, G. (2008). Shared parameter models under random effects misspecification. *Biometrika*, 95(1):63–74.

Thiébaut, A. C. and Bénichou, J. (2004). Choice of time-scale in Cox's model analysis of epidemiologic cohort data: A simulation study. *Statistics in Medicine*, 23(24):3803–3820.

Van den Hout, A. and Muniz-Terrera, G. (2016). Joint models for discrete longitudinal outcomes in aging research. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 65(1):167–186.

Wulfsohn, M. S. and Tsisatis, A. A. (1997). A Joint Model for Survival and Longitudinal Data Measured with Error. *Biometrics*, 53(1):330–339.

Zarulli, V., Marinacci, C., Costa, G., and Caselli, G. (2013). Mortality by education level at late-adult ages in turin: A survival analysis using frailty models with period and cohort approaches. *BMJ Open*, 3(7):e002841.

Appendix: K-M Survivor Functions