UNIVERSITY OF LUXEMBOURG
Department of Social Sciences

Social Mobility and Inequalities in Health.

Bayesian Perspectives on Mortality Risks and Methodological Implications for Statistical Inference.

Alessandro Procopio

2022

Abstract

Emerging and developed countries have experienced an unprecedented increase in life expectancy and a rapid shrink in mortality rates in the last thirty years. However, individuals at lower levels of socioeconomic status have higher chances of early disease onset or even death at a younger age. This association between individuals' social position and health outcomes persists in advanced welfare regimes and over time. Social science research has put considerable effort into understanding the underlying mechanisms of inequalities in health among individuals. In this Ph.D. thesis, I propose three different perspectives to understand better the dynamics that interconnect individuals' social conditions and health status, from static to a dynamic range of analysis. Chapter I analyzes the association between individuals' social position and mortality risks due to cardiovascular disease in the United Kingdom in 2012. I used the Bayesian framework to assess both inequalities between socioeconomic groups and differentials within those groups in levels of C-reactive protein (CRP) as a biological marker of cardiovascular disease and mortality risks. Chapter II analyzes the methodological issues encountered when studying social mobility, providing a deeper analysis of the Diagonal Reference Model, a statistical tool designed to overcome the identification problem. Lastly, Chapter III focuses on analyzing dropout processes in longitudinal design and proposes the Joint Modeling approach as a reliable tool for social scientists to gain additional insights on the issue of informative dropout. The main contributions of this Ph.D. thesis regard the innovative insight at the methodological level and an interdisciplinary view of inequalities in health. A more advanced methodological toolset prevent analytical pitfalls in social scientific research and statistical inference. The interconnections between social and medical sciences are essential to address policymakers better-refined policies to counteract the resurgence of health inequalities.

Keywords

Health Inequalities; Social Stratification and Mobility; Bayesian Regression Models; Diagonal Reference Model; Joint Modeling Approach.

Acknowledgement

I cannot begin to express my thanks to Professor Samuel, my supervisor. Throughout these intense years of research, he has always been supportive in the darkest days and enthusiastic about every minor achievement. I am extremely grateful to Robin as he believes in my skills much more than I do. If faith decides that I will be a Professor, he will be the perfect example. His motivation for letting me improve as a researcher and his kindness have been invaluable resources, without which most likely I could not complete this Ph.D. path. I would like to recognize also the invaluable assistance of my Ph.D. committee members. I am deeply grateful to Professor Conchita D'Ambrosio for her appreciation of my passion for statistical methods. Professor Kevin Ralston for his suggestions to revise this Ph.D. thesis (I am still embarrassed about how we met the first time). Professor Anja Leist, for her invaluable support and suggestions. I would like to express my sincere gratitude to the Youth Survey Team, particularly Caroline, Hamid, Lea, and Moritz, for the invaluable support they provided to me.

I am extremely grateful to the Center on Aging and the Life Course (CALC) at Purdue University, where I spent my last months as a Ph.D. student. I would like to thank Kenneth Ferraro, who allowed me to experience this extraordinary period. He has been of invaluable support throughout these months. I must also thank Olivia, who understood how difficult it could be for a legal alien to live in the US. I would also like to thank Madison for easing my life in analyzing the Health and Retirement Study data. I am deeply indebted to Mallory for revising the Conclusions section and deal with my terrible English. I would like to extend my sincere thanks to Shawn Bauldry and Trenton Mize for allowing me to present Chapter II at the Advanced Methods at Purdue.

This project would not have been possible without the support of my family: my mother Isabella, my father Giovanni, my sister Veronica and my aunt Grazia. I owe them a lot for their sacrifices for letting me reach this considerable achievement and for being supportive all the time. My deepest gratitude extends to my sister's family: Tom, Mary, and especially to Leo, my nephew. I promise I will be a reliable uncle to you.

I would like to recognize the invaluable assistance of my colleagues, particularly to Jason, Francisco, Tamara, Oksana and Oleksii.

I had the luck to meet many people who positively impacted my life and supported me throughout these years. I would like to give my special regards to Giulio, Amalia, Alicia, Francesca, Amin, Bianca, Matteo, Alessandro, Lorenzo, Elia, Giovanni, Oana, Maria and Claudia. You have all supported me throughout these years. I am profoundly indebted for this.

Contents

Preamble

Chapters

Conclusions

Appendix

List of Tables

List of Figures

Preamble

1. Introduction

> *"Our knowledge about the world is never better than the data on which it is based."*
>
> *Breen & Jonsson (2005:235)*

Historical Evolution and Current Challenges in the Health Inequality Field

From the beginning of the XXth century in the developed and emerging countries, individuals' life expectancy considerably increased, and overall mortality rates rapidly shrunk (Elo, 2009; Ho & Hendi, 2018). While this remarkable improvement has benefited the entirety of the population, individuals with lower income, educational degree, or occupational status tend to experience health-related problems or even die at a younger age (Elo, 2009; Feinstein, 1993). This inequality in health between individuals in different social positions seems to persist in countries with advanced welfare regimes, and it appears to have widened (at least for some measures of health) over time (MacKenbach, 2020; Mackenbach, 2012). In the last three decades, social scientific literature on stratification and health inequalities strove to assess potential pathways that could explain the persistent association between socioeconomic status (SES) and health. As an example, Adler et al. (1994) wrote: "Despite recognition for decades of this fundamental association, the reasons for its existence remain largely obscure" (Adler et al., 1994, p. 613). In 2020, Yang et al. (2020) stated that: "Despite strong evidence for the socioeconomic status (SES) gradient in health, substantial gaps remain in understanding the social and biological mechanisms underlying these disparities"" (Yang et al., 2020, p. 15). Thus, the current challenge that social scientific research is coping with concerns the thorough comprehension of the underlying mechanisms and factors affecting this relationship (Elo, 2009; Wang & Hulme, 2021). To

achieve this objective, one of the most promising tracks of research nowadays regards the recent integration of biological processes into sociological inquiries of individuals' social conditions and health (Elo, 2009; Harris & Schorpp, 2018). Historically, social scientists tried to determine the association between SES and health, focusing on the relationship between poverty and health. In this research framework, the measurement of socioeconomic status consisted of classifying individuals either above or below the poverty line (Adler & Ostrove, 1999). This research framework assumed that the increasing wealth of individuals below the threshold would correspond to a cumulative, positive effect on health outcomes. On the contrary, increasing income for individuals falling above the poverty line would not result in significant improvements in health status.

From the seminal work of Marmot et al. (1991), social scientific research recognized that inequality in health is a more complex social phenomenon. It presents different complexities and non-linearities not captured by the threshold model governing this well-known association. Marmot et al. (1991) assessed, with the Whitehall Study[1], that health-related outcomes improved and mortality rate decreased at every higher occupational grade considered, challenging the threshold model (Adler & Ostrove, 1999). Social science shifted its focus, moving from a theoretical model that considered whether the individuals were above or below the poverty line to elaborating on potential pathways underlying the social gradient of health. From this strand of research, two main theories arose as conceptual guidance for the empirical research on stratification and health. The social selection and the social causation theories provided two competing explanations to assess the causal direction of the link between socioeconomic status and health. Social causation theory assumes that SES influences health status (Link & Phelan, 1995; Phelan et al., 2010).

In social selection theory, health status favors upward social mobility and increases the chance of being in higher social classes (Blane et al., 1993). Through the years, social causation theory seemed to be the most prominent (Adler & Ostrove, 1999). However, from Lundberg (1991), social selection regained interest among social scientists, specifically in its indirect form (Haas, 2006). The indirect form of social selection states that both health outcomes and social mobility are affected by an antecedent, common factor (Lundberg, 1991). In the current research, indirect selection theory is carried forward to the life-course perspective, which focuses on the detrimental effects of deprived social conditions during early childhood of individuals on the observed health outcomes at later life stages (Corna 2013; Ferraro et al. 2016; Ferraro and

Shippee 2009; Haas 2006; Yang et al. 2020). The remarkable innovation that both social selection and social causation theories have provided to the social scientific community regards the paradigmatic shift that these theories have imprinted. Social selection and social causation theories shifted the focus to the potential mechanisms that social scientists could use to interpret and explain the social gradient of health. A low but steadily increasing number of investigations in the social sciences started to focus on the mechanisms underlying the association between social stratification and health outcomes.

The Rise of the Biomarker Use in Social Science Research

This section elaborates on the increasing use of biological markers in social research as objective health measures because of their crucial role in detecting, even at the pre-symptom stage, diseases that are associated with individuals' social position (Davillas et al., 2019; Harris & Schorpp, 2018).
In the survey research field, biomarkers are collected and derived from the gathering and analyzing different biospecimens (such as blood, saliva, or urine samples) following standard procedures performed by trained personnel. The biological markers signal the abnormal pathophysiological processes and indicate potential health risks and the future insurgence of specific diseases. For example, high C-reactive Protein levels indicate a state of chronic inflammation and insurgence of cardiovascular disease. CD4 cells counts inform the progression of AIDS disease. Low serum albumin levels, a nutritional marker of malnutrition, indicate high chances of chronic kidney disease.

In recent years social science research progressively interested in the interconnections between the social environment and the physiological answers to deprived social conditions. The intuition of exploring how social conditions get under the skin has been established earlier in social sciences. Indeed, social scientists started to claim the potential benefits during the same period in which social selection and social causation theories arose in the sociological literature. The potential advantages that social scientists may gain from the integration of biological explanations into sociological research started to be claimed in sociology (Udry 1995; Levine 1995; Freese, Li, and Wade 2003) and in demography (National Research Council 2001; Crimmins, Kim, and Vasunilashorn 2010). Specifically, these contributions claimed that the use of objective measures of health (such as the collection and analysis of biospecimens) would enable social scientists to thoroughly model the individual effects

related to experienced life events and social environment throughout stages of the life course (Harris and Schorpp 2018). For instance, with the inclusion of biomarkers in sociological inquiry, researchers can deepen the scientific understanding of how "timing, duration, and magnitude of particular social exposures such as poverty or social isolation uniquely shape physiological states and health trajectories" (Harris and Schorpp 2018, p. 364). With the expansion of biosocial surveys, social scientists nowadays can empirically test hypotheses by analyzing socio-biological data. This expansion has been made possible by technological advances in survey field collection of physiological data, which has become more feasible due to noninvasive, low-cost procedures for gathering specimens (such as saliva, blood, or urine samples) from the respondents (National Research Council, 2008; Hobcraft, 2009; Harris and Schorpp, 2018). Contemporaneously, social sciences progressively pushed for a shift from uni- to an interdisciplinary perspective of understanding social phenomena (Jacobs & Frickel, 2009).

Integrating physiological, objective health measures on social surveys has a mutual benefit to both biomedical - such as social epidemiology and public health research - and sociology - particularly medical sociology. The mutual benefit mainly concerns data collection and sample representativeness on which the study's external validity relies. Biomedical studies are interested in collecting thoroughly detailed biological measures; however, the study design usually targets small or non-representative groups of participants in the clinical trial. In this perspective, inference to the entire population and external validity is limited to the case of the study. Additionally, in biomedical studies, individuals' SES has been commonly treated as a mere characteristic of the participants in the trial, thus not considering the substantial importance of social conditions as a driving force that shapes inequalities in health (Harris & Schorpp, 2018; McEwen, 2015). Integrating biomarkers in social surveys prioritizes the random sampling and representativeness of the drawn sample from the population of interest. In this framework, the acquisition of relevant physiological measures represents a secondary step through which it is possible to expand the survey range. Hence, medical researchers might benefit from a representative study design with the inclusion of a wide range of physiological measures. From a sociological point of view, the advantage is the possibility of mapping biological reactions to the social environment, social events, and exposure to social conditions.

Even more, sociologists can signal to biomedical research that SES plays a determinant role in shaping health inequalities (Harris & Schorpp, 2018). The disadvantages related to the inclusion of biological measures in social

science surveys can hit three main aspects of the survey design. Firstly, the field collection of biospecimens and personnel training increases the costs of surveys. Secondly, collecting biospecimens is an additional source of burden for the respondent. Lastly, the advantage in prediction accuracy of individuals' specific mortality can be achieved by considering one single biomarker per biological system (such as C-reactive protein and fibrinogen for the cardiovascular system or HDL levels for the metabolic functions) rather than composite measures for mortality prediction (such as Allostatic Load; Glei et al. 2014). Despite these costs, surveys that integrate biological measures of individuals' physiological functioning are increasing in number over time (National Research Council 2008; Harris 2010; Harris and Schorpp 2018), as the gains outpace the drawbacks.

Including biological markers in social science research can have theoretical and methodological advantages. The theoretical advantage relies on the possibility to analyze how the social environment can affect the physiological functioning of individuals. The methodological advantage relies on avoiding the inherent subjectivity of self-reported health measures and predicting socially-graded disease insurgence before their symptomatic manifestation.

Literature Gaps

Even if the use of biomarkers as objective measures of health risks is growing in the field of public health (Karimi et al., 2019; R. S. Liu et al., 2017) and aging research (Ferraro et al., 2016; Ferraro & Shippee, 2009; Piazza et al., 2010; Yang et al., 2020), social scientific literature lacks adequate analytical strategies that involve biological pathways to explain the association between social stratification and health outcomes. In particular, previous research literature focuses on the differentials between groups categorized in the SES scale, for instance, comparing individuals with different educational degrees, occupational statuses, or income levels. That means social research focused on the inter-group differences (or the contextual effects), while it lacked a thorough insight into the intra-group differences or compositional effects, i.e., the differences among individuals within the same social group.

A second gap in the sociological literature outlined in this manuscript regards how social stratification and, in a dynamic perspective, social mobility are critical factors in an empirical model. It is practical, at this stage, to distinguish between the concepts of social stratification and social mobility. In examining the relationship between social stratification and health, researchers

are interested in the static association between the differentials in health outcomes according to individuals' class position (Marmot et al., 1991, Adler and Ostrove, 1999). Social researchers measure the differentials in health outcomes among individuals according to their social position observed at a specific point in time and context. Thus, it returns a static depiction of the association as the researcher cannot ascertain the evolutionary trend of this association. Conversely, investigating the relationship between social mobility and health assumes a dynamic change perspective on the individuals' class position and the relative health outcome pattern (Power et al., 1996; Manor et al., 2003). Even if the dynamic perspective of the association between social position (and its evolution over time) and health might be appealing, this approach introduces additional complexities at the empirical stage of the research. Indeed, empirical social scientists have put a remarkable effort to apply and test the theoretical frameworks using statistical models to analyze the consequences of social mobility on health.

A particularly cumbersome problem relates to the identification problem. The identification problem arises when the researcher includes indicators of origin, destination, and mobility in the same regression model. Technically, this analytical strategy introduces linear dependency (i.e., perfect collinearity) between the covariates. Indeed, individuals' social mobility (M) results in the differential between the social origin (O) and social destination (D), such as M = D - O. The proposed solutions to the identification problem by scholars of social mobility dates back to the 1960s, such as the square additive model by Duncan (1966), the halfway/difference model by Hope, 1971, 1975, and more recently, the Diagonal Reference Model (DRM) by Sobel (1981, 1985). Among these models, the Diagonal Reference Model gained particular interest and is used in different fields, including health inequalities (Jonsson et al., 2017; Missinne et al., 2015; Präg & Richards, 2019). However, the increasing literature deploying the DRM seems not capable of clarifying the role of mobility on health, as it evidenced several null or weak findings of mobility effects, in stark contrast to the expectations derived from theory[2].

The dynamic perspective on social mobility and health outcomes might involve using longitudinal (or panel) data. Health and social research extensively use panel data to study mortality rates (Arbeev et al., 2014; Haviland et al., 2011; X. Liu, 2013; X. Liu et al., 2010; Stolz et al., 2018; Zarulli et al., 2013; Zheng, 2020). Even if panel data are a powerful tool to assess causal relationships (Halaby, 2004), this type of data collection suffers from (unavoidable) dropout rate. The dropout rate might severely bias the statistical estimation of an outcome measured over time (such as health and

social mobility) if it correlates with the longitudinal process. In this case, the statistical literature classifies the phenomenon as informative dropout, or Missing Not at Random (MNAR, Diggle and Kenward 1994; Little 1995; Rubin 1976). In particular, MNAR introduces two sources of bias in estimating a longitudinal outcome. Endogeneity occurs as the longitudinal outcome variable might influence and, at the same time, be influenced by the dropout rate. Secondly, progressive MNAR leads to a sample in which the survival of the fittest process dynamically selects individuals, thus homogenizing the observation units according to key characteristics. For example, healthier individuals have the lowest chances of dropout from the study. Suppose the study focuses on trends among individuals on health inequalities. In that case, the risk is to base the statistical model on the healthiest individuals, therefore homogeneous in terms of the health outcome, and the model underestimates the actual effect on the longitudinal outcome.

Research Questions and Analytical Strategies

This Ph.D. thesis has been motivated by the objective of answering three interrelated main research questions. The first research question relates to the substantive and empirical approach to studying inequalities in health. The second and the third research questions are methodological by their nature. The substantive research question regards the assessment of inequalities in health between individuals with different social conditions and differentials among individuals at the same level of SES. The second research question relates to the efficiency of the DRM to capture social mobility effects and distinguish them from the origin and destination effects. Lastly, the third research question concerns the problem of informative missing data in longitudinal design and how it can tackle this methodological problem.

The analytical strategies I have implemented in this Ph.D. are diverse and specifically tailored to the relative research questions aforementioned. To answer the first one, I used the Bayesian approach to assess inequalities in health among and within groups. I capitalized on the opportunity given by the Bayesian approach to set a hierarchical structure of the research question. That means the analytical strategy for the first chapter conceives the levels of SES as higher groups that embody the individuals. Therefore, the effects of SES on health outcomes are population-level effects (or contextual effects), and they relate to the macro-structure of inequality in health. To assess the differentials within the macro-structures, I made use of a particular specification of the Bayesian model. The distributional model differs from the

more traditional statistical approaches because, alongside the group-specific means, I could also specify the standard deviation of the dependent variable as a linear function dependent upon the characteristics of interest. The substantive advantage related to the simultaneous estimation of the mean and the standard deviation relates to the capability to interpret the model as between and within-group differences.

I implemented the second analytical strategy to design a statistical experiment using the Monte Carlo simulation technique. The Monte Carlo simulation had the objective of analyzing the behavior of the DRM in two main scenarios: the first one involves a continuous dependent variable, while the other is a dichotomous dependent variable. The analytical strategy focused on examining the Diagonal Reference Model in these scenarios, particularly on the capability of the model to detect the true magnitude effect of mobility and the correct statistical significance hypothesis for these effects correctly. The focus of the first aspect regards the unbiasedness of the estimates. The second aspect concerns the correct computation of the statistical significance of mobility effects.

The third analytical strategy uses the Monte Carlo simulation technique to compare the behavior of three statistical models in a scenario of missing data that are informative of a longitudinal process of interest. The three statistical approaches are the Linear Mixed effect model, the Weibull regression model, and the Joint Modeling (JM) approach. The aim was to assess the benefits and the limitations of the JM, particularly with a comparative perspective on the other two widely used approaches in the presence of informative dropout. The comparison of the three statistical models involved two scenarios. The first scenario considered the models' behavior under unobserved heterogeneity. The second scenario examines the models' capability to correctly estimate complex longitudinal trends, such as cubic nonlinear outcomes.

Structure of the Thesis

The outline of the Ph.D. thesis incorporates four main parts. The first part, the Preamble, provides the Introduction section and a summary of the chapters. The summary of the chapters highlights the main topics and the key findings of each contribution. The second part, the core of this manuscript, embodies the three main chapters of the Ph.D. Thesis. The third part draws up the conclusive discussion of this manuscript, which sketches the limitations, the contributions to the literature, and the suggestions for future research.

The last part of the Ph.D. provides additional materials for each chapter of the Chapters section.

The organization of the Chapters section provides a temporal order of the chapters. The underlying reasoning that outlines the core part of the Ph.D. thesis depends on the vision of health inequalities that stretches from a static to a more dynamic perspective of the SES-health association. The first chapter outlines a static, bounded picture of health inequalities in terms of time and space. In this chapter, SES corresponds to the social stratification structure that embeds individuals in that particular period and place. Individuals might experience, throughout their life stages, shifts in their current social position. That is to say, to capture one aspect of the dynamic characteristic of the SES-health relationship, it is essential to include, in the theoretical and the empirical model, also individuals' social mobility. Indeed, I dedicated the second chapter to the methodological problems the empirical researcher might incur when including social mobility as a critical aspect of the social gradient of health. The dynamic perspective of health inequalities can be fully understood by considering the changes across the life course of social position and measuring the longitudinal evolution of individuals' health. In this regard, panel data represent an invaluable resource for social scientists. However, panel data introduce at the same time methodological issues of which social scientists should be aware. The third chapter of this manuscript addresses the problems that social scientists might face when looking at the full dynamics (i.e., considering social mobility and longitudinal pattern of health) of the social gradient of health using panel data. The third part, the Conclusions, discusses the results presented in the chapter and their main implications for social scientists. This section then tries to answer the research questions discussed before. The Conclusions lastly elucidate the main contributions and limitations of the contributions and addresses suggestions for further research.
The main goal of this Ph.D. thesis is to innovate the current methodological procedures to understand better the dynamics of health inequalities and the mechanisms underlying this social phenomenon.

Notes

[1]The Whitehall Study consisted of a 10-year study focused on morbidity and mortality among British civil servants.

[2]The identification problem involved here is similar to the age-period-cohort (APC) literature, where one variable is a function of the others.

Bibliography

Adler, N. E., Boyce, T., Chesney, M. A., Cohen, S., Folkman, S., Kahn, R. L., & Syme, L. S. (1994). Socioeconomic Status and Health: The Challenge of the Gradient. *American Psychologist*, *49*(1), 15–24.

Adler, N. E., & Ostrove, J. M. (1999). Socioeconomic status and health: What we know and what we don't. *Annals of the New York Academy of Sciences*, *896*, 3–15.

Arbeev, K. G., Akushevich, I., Kulminski, A. M., Ukraintseva, S. V., & Yashin, A. I. (2014). Joint analyses of longitudinal and time-to-event data in research on aging: Implications for predicting health and survival.

Blane, D., Smith, G. D., & Bartley, M. (1993). Social selection: what does it contribute to social class differences in health? *Sociology of Health & Illness*, *15*(1), 1–15.

Corna, L. M. (2013). A life course perspective on socioeconomic inequalities in health: A critical review of conceptual frameworks. *Advances in Life Course Research*, *18*(2), 150–159.

Crimmins, E., Kim, J. K. I., & Vasunilashorn, S. (2010). Biodemography: New approaches to understanding trends and differences in population health and mortality. *Demography*, *47*, 41–64.

Davillas, A., Jones, A. M., & Benzeval, M. (2019). The Income-Health Gradient: Evidence From Self-Reported Health and Biomarkers in Understanding Society. In *Panel data econometrics: Empirical applications* (pp. 709–741). Elsevier Inc.

Diggle, P., & Kenward, M. G. (1994). Informative Drop-Out in Longitudinal Data Analysis. *Applied Statistics*, *43*(1), 49.

Duncan, O. D. (1966). Methodological issues in the analysis of social mobility. In *Social structure and mobility in economic development* (pp. 51–97). Chicago, Aldine Publishing Company.

Elo, I. T. (2009). Social class differentials in health and mortality: Patterns and explanations in comparative perspective. *Annual Review of Sociology*, *35*, 553–572.

Feinstein, J. S. (1993). The Relationship between Socioeconomic Status and Health: A Review of the Literature. *The Milbank Quarterly*, *71*(2), 279.

Ferraro, K. F., Schafer, M. H., & Wilkinson, L. R. (2016). Childhood Disadvantage and Health Problems in Middle and Later Life: Early Imprints on Physical Health? *American Sociological Review*, *81*(1), 107–133.

Ferraro, K. F., & Shippee, T. P. (2009). Aging and cumulative inequality: How does inequality get under the skin? *Gerontologist*, *49*(3), 333–343.

Freese, J., Li, J.-C. A., & Wade, L. D. (2003). The Potential Relevances of Biology to Social Inquiry. *Annual Review of Sociology*, *29*(1), 233–256.

Glei, D. A., Goldman, N., Rodríguez, G., & Weinstein, M. (2014). Beyond self-reports: Changes in biomarkers as predictors of mortality. *Population and Development Review*, *40*(2), 331–360.

Haas, S. A. (2006). Health Selection and the Process of Social Stratification: The Effect of Childhood Health on Socioeconomic Attainment. *Journal of Health and Social Behavior*, *47*, 339–354.

Halaby, C. N. (2004). Panel models in sociological research: Theory into practice. *Annual Review of Sociology*, *30*(Villareal 2002), 507–544.

Harris, K. M. (2010). An integrative approach to health. *Demography*, *47*(1), 1–22.

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

Haviland, A. M., Jones, B. L., & Nagin, D. S. (2011). Group-based trajectory modeling extended to account for nonrandom participant attrition. *Sociological Methods and Research*, *40*(2), 367–390.

Ho, J. Y., & Hendi, A. S. (2018). Recent trends in life expectancy across high income countries: Retrospective observational study. *BMJ (Online)*, *362*.

Hobcraft, J. (2009). Reflections on the incorporation of biomeasures into longitudinal social surveys: An international perspective. *Biodemography and Social Biology*, *55*(2), 252–269.

Hope, K. (1971). Social Mobility and Fertility. *American Sociological Review*, *36*(6), 1019–1032.

Hope, K. (1975). Models of Status Inconsistency and Social Mobility Effects. *American Sociological Review*, *40*(3), 322–343.

Jacobs, J. A., & Frickel, S. (2009). Interdisciplinarity: A critical assessment. *Annual Review of Sociology*, *35*, 43–65.

Jonsson, F., Sebastian, M. S., Hammarström, A., & Gustafsson, P. E. (2017). Intragenerational social mobility and functional somatic symptoms

in a northern Swedish context: analyses of diagonal reference models. *International Journal for Equity in Health*, *16*(1), 1–10.

Karimi, M., Castagné, R., Delpierre, C., Albertus, G., Berger, E., Vineis, P., Kumari, M., Kelly-Irving, M., Chadeau-Hyam, M., Lynch, S. M., & Bartlett, B. (2019). Early-life inequalities and biological ageing: A multisystem Biological Health Score approach in U nderstanding S ociety. *Journal of Epidemiology and Community Health*, *73*(8), 693–702.

Levine, S. (1995). Time for creative integration in medical sociology. *Journal of Health and Social Behavior*, *Spec No*(1995), 1–4.

Link, B., & Phelan, J. (1995). Social conditions as fundamental causes of health inequalities. *Journal of Health and Social Behavior*, *35*, 80–94.

Little, R. J. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, *90*(431), 1112–1121.

Liu, R. S., Aiello, A. E., Mensah, F. K., Gasser, C. E., Rueb, K., Cordell, B., Juonala, M., Wake, M., & Burgner, D. P. (2017). Socioeconomic status in childhood and C reactive protein in adulthood: A systematic review and meta-analysis. *Journal of Epidemiology and Community Health*, *71*(8), 817–826.

Liu, X. (2013). Survival Models on Unobserved Heterogeneity and their Applications in Analyzing Large-scale Survey Data. *Journal of Biometrics & Biostatistics*, *05*(02).

Liu, X., Engel, C. C., Kang, H., & Gore, K. L. (2010). Reducing selection bias in analyzing longitudinal health data with high mortality rates. *Journal of Modern Applied Statistical Methods*, *9*(2), 403–413.

Lundberg, O. (1991). Childhood Living Conditions, Health Status, and Social Mobility: A Contribution to the Health Selection Debate. *European Sociological Review*, *7*(2), 149–162.

MacKenbach, J. P. (2020). Re-thinking health inequalities. *European Journal of Public Health*, *30*(4), 615.

Mackenbach, J. P. (2012). The persistence of health inequalities in modern welfare states: The explanation of a paradox. *Social Science and Medicine*, *75*(4), 761–769.

Manor, O., Matthews, S., & Power, C. (2003). Health selection: The role of inter- and intra-generational mobility on social inequalities in health. *Social Science and Medicine*, *57*(11), 2217–2227.

Marmot, M. G., Stansfeld, S., Patel, C., North, F., Head, J., White, I., Brunner, E., Feeney, A., Marmot, M. G., & Smith, G. D. (1991). Health inequalities among British civil servants: the Whitehall II study. *The Lancet*, *337*(8754), 1387–1393.

McEwen, B. S. (2015). Biomarkers for assessing population and individual health and disease related to stress and adaptation. *Metabolism: Clinical and Experimental*, *64*(3), S2–S10.

Missinne, S., Daenekindt, S., & Bracke, P. (2015). The social gradient in preventive healthcare use: What can we learn from socially mobile individuals? *Sociology of Health and Illness*, *37*(6), 823–838.

National Research Council. (2001). *Cells and surveys: Should biological measures be included in social science research?* (C. E. Finch, J. W. Vaupel, & K. Kinsella, Eds.). The National Academies Press.

National Research Council. (2008). *Biosocial surveys* (M. Weinstein, J. W. Vaupel, & K. W. Wachter, Eds.). The National Academies Press.

Phelan, J. C., Link, B. G., & Tehranifar, P. (2010). Social Conditions as Fundamental Causes of Health Inequalities: Theory, Evidence, and Policy Implications. *Journal of Health and Social Behavior*, *51*, 28–40.

Piazza, J. R., Almeida, D. M., Dmitrieva, N. O., & Klein, L. C. (2010). Frontiers in the use of biomarkers of health in research on stress and aging. *Journals of Gerontology - Series B Psychological Sciences and Social Sciences*, *65 B*(5), 513–525.

Power, C., Matthews, S., & Manor, O. (1996). Inequalities in self rated health in the 1958 birth cohort: Lifetime social circumstances or social mobility? *Bmj*, *313*(7055), 449–453.

Präg, P., & Richards, L. (2019). Intergenerational social mobility and allostatic load in Great Britain. *Journal of Epidemiology and Community Health*, *73*, 100–105.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592.

Sobel, M. E. (1981). Diagonal Mobility Models : A Substantively Motivated Class of Designs for the Analysis of Mobility Effects. *American Sociological Review*, *46*(6), 893–906.

Sobel, M. E. (1985). Social Mobility and Fertility Revisited : Some New Models for the Analysis of the Mobility Effects Hypothesis. *American Sociological Review*, *50*(5), 699–712.

Stolz, E., Mayerl, H., Rásky, V., & Freidl, W. (2018). Does Sample Attrition Affect the Assessment of Frailty Trajectories among Older Adults? A Joint Model Approach. *Gerontology*, *64*(5), 430–439.

Udry, J. R. (1995). Sociology and biology: What biology do sociologists need to know? *Social Forces*, *73*(4), 1267–1278.

Wang, J., & Hulme, C. (2021). Frailty and socioeconomic status: A systematic review. *Journal of Public Health Research*, *10*(3), 553–560.

Yang, Y. C., Schorpp, K., Boen, C., Johnson, M., & Harris, K. M. (2020). Socioeconomic Status and Biological Risks for Health and Illness Across

the Life Course. *The journals of gerontology. Series B, Psychological sciences and social sciences*, *75*(3), 613–624.

Zarulli, V., Marinacci, C., Costa, G., & 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.

Zheng, H. (2020). Unobserved Population Heterogeneity and Dynamics of Health Disparities. *Demographic Research*, *43*, 1009–1048.

2. Summary of the Chapters

In Chapter I, I tried to shed light on one of the potential mechanisms through which health inequalities occur. Indeed, the aim of this study was to analyze the association between SES, risks of mortality due to cardiovascular diseases (CVD) and levels of chronic inflammation, measured through the biomarker C-reactive protein (CRP). This pathway can be considered particularly important as CVD represents the leading cause of death in developed countries (Brummett et al., 2014). In order to assess this mechanism, I used data from the Understanding Society - United Kingdom Household Longitudinal Study (UKHLS). These data have the advantage of having a wide range of collected biomarkers. I have taken into account three measures of SES, in order to capture the multidimensional characteristics of social position: occupational status, educational levels and equalized household income. We used the Bayesian Regression Model (BRM) in order to provide to the scientific community a range of possible effect magnitudes related to SES on CRP. The models I implemented aimed to analyze the differences between the categories of the SES variables, and also the within SES group variance, implementing the distributional model. The results suggest that, among occupational groups, there is a remarkable difference between the highest category in the social hierarchy and the lowest. It is worth to note that also small employers have found to have similar levels of CRP, with respect to the individuals hired in the highest management or large employers. The most determinant SES dimension is educational level, in which we found the largest differences. For what concerns the distributional model, interesting findings suggest a strong similarity between individuals in the highest category of occupational status, but individuals differ greatly within the lowest. Even more interesting is that individuals within the same educational level do not differ significantly, in terms of mortality risks due to CVD.

Chapter II is devoted to fulfill the gap concerning the linear dependence between individuals' social origin, destination and mobility. Through means of Monte Carlo Simulation, I have tested the DRM in two main scenarios. Firstly, I tested the model assuming a continuous dependent variable. Subsequently, I have analyzed the model behavior when the dependent variable is

dichotomous. I have then assessed the performances of the DRM highlighting the estimation bias and the Empirical Coverage Rate (ECR) in both the scenarios. The results showed that the DRM is generally not affected by bias. However, my attention shifted toward worrying levels of ECR. Indeed, in many cases the model failed to statistically infer the mobility parameter, which could explain why the social scientific literature on consequences of social mobility hardly found statistically significant effects of mobility.

The last contribution, presented in Chapter III, proposed a relatively new model to tackle informative dropout, that is the Joint Modeling (JM) of longitudinal and survival data. The model originated in the end of the twentieth century in the biomedical field. In fact, early developments of the JM approach have been used to study HIV/AIDS progression and chances of survival among patients. The model has been applied then in cancer studies, where the main interest was in assessing the relationship between Quality-of-Life (QOL) and survival. The JM approach uses a two-stage estimation procedure: the first submodel grasps the longitudinal pattern under study; secondly, the JM uses the estimated longitudinal pattern as time-varying covariate in a survival regression model. This kind of estimation strategies to account for MNAR is not new in the social sciences (see Heckman 1979). However, this class of models only recently gained attention from the social scientists (Li et al., 2020). In this contribution, I have compared the JM approach against the Linear Mixed Model (LMM) for what concerns the estimation of the longitudinal pattern and against the Weibull regression model (WM) for the survival/dropout part. The scenarios in which I compared the models have been theoretically founded by previous literature. In the first scenario, I introduced unobserved heterogeneity as source of bias; the second scenario concerned the appropriate modeling of more complex (i.e., cubic shaped) longitudinal pattern. Both scenarios have been modeled so as the association between the longitudinal pattern and the dropout rate would be 0, mildly associated (0.25) and strongly associated (0.5). The results suggests that the LMM and the JM are pretty similar to each other. However, when the JM approach takes considerable advantages in modeling the dropout mechanism. In fact, in all scenarios the JM approach performed better than the WM. This means that the JM is particularly useful for the fields of social sciences that make of extensive use of survival analysis such as demography and health research.

Chapters

I. Social Conditions Under the Skin: Socioeconomic Status, C-reactive Protein and Health Inequalities in Bayesian Perspective.

Outline

Bibliographic Information

This chapter is drafted. Expected submission to a Health (Social Sciences) journal.

Author's contribution

I developed the methodology, writing preparation, interpretation of findings, and the data generation and analysis. Robin Samuel conceptualized, supervised, and reviewed the chapter. We jointly collaborated at the conceptualization of the project.

Abstract

In the last two decades, social stratification research on health inequalities has seen a steady increase of studies involving biological factors to measure individuals' health and to further understand the social gradient of health. Concurrently, biomedical, epidemiological, and public health research into the role of socioeconomic status (SES) in shaping health inequalities has been fostering and inspiring sociological investigations.
In particular, the innovative use of biomarkers as measures of respondents' health helped social researchers to address problems of endogeneity inherent to subjective measures such as self-reported health. Additionally, the inclusion of biological measures of health has the potential to identify causal pathways by taking into consideration the intertwined biological and social characteristics that affects individuals' health and well-being.
Contributing to this emergent strand of research, our study investigates how SES inequalities get under the skin. We are interested in the effect of socioeconomic inequality on chronic inflammation, measured through the C-reactive protein (CRP), a widely used biomarker for the immune function of individuals. Drawing on cumulative inequality theory, our contribution aims to elucidate how particular social conditions affect the immune function across the life course and, thus, individuals' health.
We use data from the Understanding Society Health Assessment Panel, wave 2012. We employ two Bayesian Regression Models (BRM) to account for differences in between and within SES groups. The Bayesian framework further allows us to calculate a range of likely parameters (e.g., effect magnitudes) that best fit the data. We measured individuals' SES in terms of level of education, income and occupational status. This strategy allows us to assess the specific effect of each of the measures on chronic inflammation. Our results suggest an association between SES and CRP, indicating that SES is not only related to the immediate social exposure to stressors (such as deprived social conditions), but also to the healthy ageing of individuals. The BRMs further reveal considerable heterogeneity in the likely parameters' distributions relative to our SES measures. Our analysis highlights the benefits of using C-reactive protein as a proxy for individuals' health in a Bayesian framework, especially if taken into account alongside self-reported health. In conclusion, the study of mechanisms through which social conditions influence the functioning of individuals' biological systems appears to be a promising avenue to advance our understanding of health inequalities.

I.1 Introduction

In the biomedical and social scientific literature, a growing number of contributions confirm the graded association between socioeconomic status (SES) and mortality risks among individuals (Adler & Ostrove, 1999; Brummett et al., 2014; K. M. Harris & Schorpp, 2018; House, 2002; Nazmi et al., 2010; Winkleby et al., 1992; Yang et al., 2020). Although social scientists have a well-established corpus of contributions that link individuals' social conditions and health outcomes (Adler et al. 1994; Adler and Ostrove 1999; Marmot et al. 1991; Ferraro and Shippee 2009; Elo 2009; Clark et al. 2009;Pudrovska 2014), only recently scientific community turned the gaze to the mechanisms through which we observe this relationship (Elo, 2009; Freese, 2018; K. M. Harris & Schorpp, 2018). In particular, social scientists and epidemiologists focused on the mediator role of chronic illness and inflammation on the association between individuals' SES and health disparities (Baum et al., 1999; Dowd & Zajacova, 2007; Pudrovska, 2014), in the life course (Ben-Shlomo & Kuh, 2002; Corna, 2013; Ferraro & Shippee, 2009; Hallqvist et al., 2004; Pollitt et al., 2008; Power et al., 1999).

This paper focuses on how individuals' SES potentially shapes the distribution of mortality risks due to cardiovascular diseases (CVDs). CVD is one of the leading causes of mortality and morbidity in high-income countries (Brummett et al., 2014; Clark et al., 2009; Gruenewald et al., 2012; Kavanagh et al., 2010; Liu et al., 2017; Loucks et al., 2010; Mitchell & Aneshensel, 2017). Previous research has found the evidence concerning the socially patterned onset of CVDs, in which individuals from a lower SES encountered higher risks of mortality due to CVD (Goodman et al., 2005; Lubbock et al., 2005; Winkleby et al., 1992), confirming the social causation theory (Link and Phelan 1995; Phelan et al. 2010). Additionally, CVD causes a considerable burden on individuals and public health. Therefore, biomedical and social science research strove to understand and identify the principal physiological changes that could signal the onset of CVD in individuals at the pre-symptom stage of the disease (Davillas et al., 2019; Dowd & Zajacova, 2007; McEwen, 2015; Mitchell & Aneshensel, 2017; Rosvall et al., 2008). In this sense, prevention and identification of the population at most significant risk of CVD has been a primary objective for social scientists, public health policy-makers, and epidemiologists(Davillas et al., 2017; Herd et al., 2007; Mitchell & Aneshensel, 2017). Biomedical studies indicated as a reliable physiological marker of CVD risk is the C-reactive Protein (CRP, T. B. Harris et al. 1999; Laaksonen et al. 2005). Specifically, CRP is an acute-phase protein produced by hepatocytes as a response of the immune system to acute infection or systemic inflammation (Alley et al., 2006; McDade et al., 2011). In the field of social inequalities in

aging and mortality risks, accumulating evidence highlight the association between CRP and individuals' social conditions, where individuals in higher SES position have lower levels of CRP and thus lower risks of CVD onset (Brummett et al., 2014; Davillas et al., 2019; Gimeno et al., 2008; Jousilahti et al., 2003; Karimi et al., 2019; Koster et al., 2006; Lubbock et al., 2005; McDade et al., 2011).

Our study contributes to this emergent research strand. Our paper investigates how SES inequalities get under the skin, addressing the mechanism that links individuals' SES, CVD, and mortality risks and the role of CRP as a mediator factor. The aim of this study is twofold. Firstly, the contribution of our paper concerns the methodological paradigm deployed in this study, which used two types of Bayesian Regression Model (BRM). To the best of our knowledge, this is the first paper that evaluates this mechanism of health inequality from a Bayesian perspective. Using the BRM, our focus of the empirical analysis shifts from the conventional point estimates (and their statistical significance) to a targeted distribution of likely parameters (thus, assessing which coefficient magnitude is more likely than others). Secondly, we did not limit the attention to SES group comparison concerning the mortality risks due to CVD (i.e., comparing the means specific to each category of SES). Still, we also explored the intra-group differences in CRP levels. In doing so, we modeled the standard deviation of CRP as a function of individuals' SES through a Bayesian distributional model (Umlauf & Kneib, 2018). We used data from the wave 2012 (Health Assessment Panel) of the United Kingdom Household Longitudinal Study (UKHLS), a nationally representative longitudinal survey set in the United Kingdom[3]. To take into account the multidimensional characteristics of SES, we included in the empirical analysis three measures of individuals' social conditions: occupational status, educational level, and household income (see Elo 2009; Goldthorpe 2010 for a methodological discussion).

The remainder of the paper is as follows: the following section will review the current state of the biomedical literature on risks of CVD and CRP due to deprived social conditions. Next, we present the UKHLS and the specifications of the Bayesian models we implemented. In the Results section, we offer the empirical findings of the model. Finally, we discuss the potential implications of this study and the limitations to be addressed to the social scientific community.

I.2 Data & Methods

Data

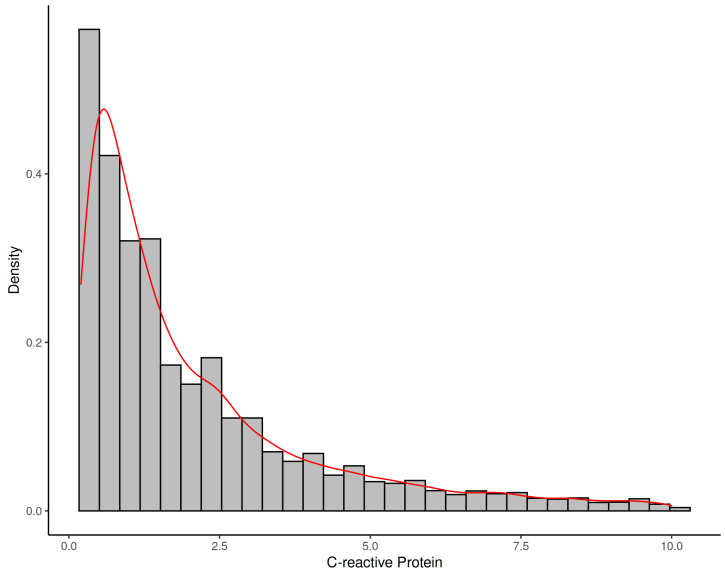
The data we used for the empirical analysis comes from the UK Household Longitudinal Study (UKHLS), wave 2 (Health Assessment Panel, 2012). The Understanding Society – UKHLS is a large and representative survey of households sampled in the United Kingdom (UK), Scotland, Wales, and Northern Ireland. Administration of the main survey interested the General Population survey (GPS)[4]. The sampling procedure for the GPS consisted of a two-stage step: the first primary sampling unit (PSU) consisted of a sample of postcode sectors, within which the addresses were the sampling units. The UKHLS provides a multi-purpose questionnaire to the respondents, covering various topics relevant to social research. In 2010 and 2012, alongside the main questionnaire, the survey design included questions on health and collected blood samples from the respondents that provided their consent to be part of the procedure. The eligibility criteria for the respondents to take part in the nurse health assessment were completing the face-to-face interview, aged 16 or older, living in England, Scotland or Wales, completing the questionnaire in English, and lastly, for not pregnant women (Mcfall et al., 2012). Excluded participants were individuals with HIV, hepatitis A or B, and clotting or bleeding disorders. The aim of the collection of biospecimens by registered nurses was to, on the one hand, collect information on potential health risks. On the other hand, blood sample collection supports genetic analyses and creates a genetic database. The nurse health assessment interested a subsample of the GPS and included anthropometric measures (such as height, weight, percent body fat, and waist circumference), blood pressure, grip strength, lung function, and blood samples. The basis of the biomarkers selection from the blood samples regards the environmental effect (socioeconomic, physical, or psychosocial), the impact on the biospecimen, its importance to essential health conditions, and the proportion of the population affected by the disease. From the 9896 observations of the initial sample, 523 individuals recorded a CRP level higher than 10 mg/L. Thus, we deleted those cases from the study. Individuals with inapplicable values (N=517) have been set to missing values and excluded from the statistical analysis. After deleting inapplicable values on the covariates, the total sample size was N=8514.

Measures

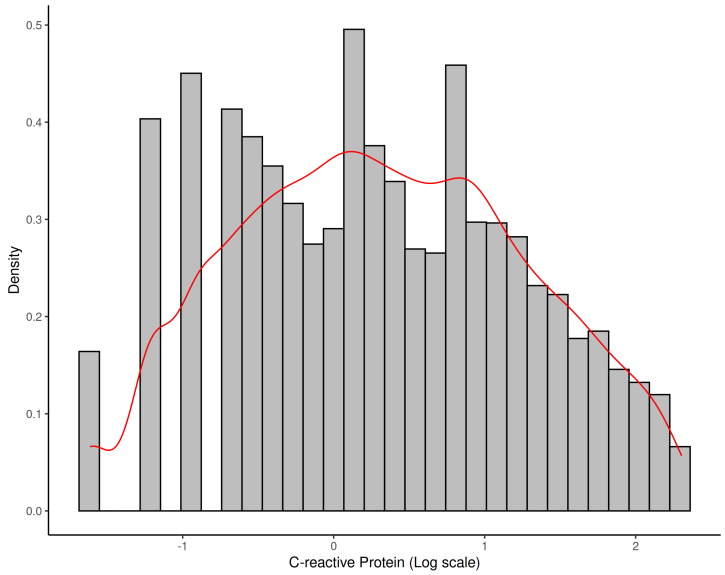
Dependent Variable

The dependent variable of the two BRMs is the recorded CRP in the Wave 2012 of the Understanding Society - UKHLS, measured in mg/L. We consider at high risk of CVD individuals with a CRP level great than 3 mg/L, which is the cut-off point for defining an individual affected by chronic inflammation (Brummett et al., 2014; McDade et al., 2011). We dropped CRP levels higher to 10 mg/L, as such values are indexes of temporary acute inflammation. As the distribution of the recorded values among the individuals in the sample is highly skewed, we log-transformed the variable to normalize the distribution. Figure I.1 shows the distributions of the CRP as in the sample and after log-transformation.

Figure I.1: Distribution of C-reactive Protein in the sample. Panel (a) shows the natural scale. Panel (b) the log-transformed distribution of CRP.



(a) Distribution of C-reactive Protein in the sample



(b) Distribution of the log-transformed CRP distribution

Covariates

SES Measures:
To capture the multidimensionality of SES (Elo, 2009; Goldthorpe, 2010),

we included in the statistical model three measures of SES: occupational status, education, and income. Occupational status comprises eight categories from the National Statistics Socio-economic Classification (NS-SEC): Large employers & higher management (3.1%), Higher professional (4.4%), Lower management & professional (16%), Intermediate (7.8%), Small employers & own account (5.5%), Lower supervisory & technical (4.3%), Semi-routine (9.4%), Routine (5.6%) and Not in LM (44.1%). The last category includes retired individuals, students, and individuals that are not currently working at that moment. The main reason is to compare individuals who have a job (and within those, compare the occupational categories) vs. individuals who do not currently.

Educational level has been measured in five categories: Degree (22.4%), Other higher degree (13.5%), A-level(18.2%), GCSE(19.7%), Other Qualifications(11.6%) and No qualification(14.5%).

We have taken the gross household income registered the month before the interview concerning the third SES measure. We have then equivalized the scale, dividing the income scale by the equivalence scale set by the OECD, returning the equivalized income scale for the number of household members.

Controllers:
The covariates we included to control the relationship between the SES measures and the levels of CRP concern sociodemographic characteristics and health behavior of the individuals.
Among the sociodemographic variables, we have considered in the analysis the age (measured as a continuous scale from 16 to 102 years old) of the individuals, gender (males are the 44.4% of the sample, while females are the 55.5%), and house ownership (see McDade et al. 2011) as an indicator of wealth (77.7% owned the house the household lives, the 21% has a house on rent and the 1.27% has a mixed form of ownership). The health behavior of individuals that might influence CRP levels (Dowd et al., 2009; Yang et al., 2020) has been measured by taking into account: the level of sports activity (scale from 0 - no activity - to 10 - very active), the Body Mass Index (BMI) calculated as weight/(height/100)^2 and smoking history: in the sample, 18.9% reported to be a current smoker, while individuals that used to smoke are the 41.5%; individuals that never tried smoking are the 39.6% of the sample. Lastly, we included a measure of Self Rated Health (as it is associated with CRP levels, see Shanahan et al. 2014, coded into five categories: excellent (15.4%), very good (35.6%), good (28.9%), fair (15.2%), and poor health (4.8%).

Descriptive Statistics

Table I.1 provides an overview of the descriptive statistics (mean and standard deviations) of the covariates included in the models.

Table I.1: Summary Statistics of the dependent variable and the covariates

Variables	Mean	St. Dev.
CRP (log scale)	0.323	0.930
Sport Activity	0.000	1.000
Age	0.000	1.000
Income	−0.000	1.000
BMI	0.000	1.000
Occupation		
Large employers & higher management	0.047	0.211
Lower management & professional	0.171	0.377
Intermediate	0.078	0.268
Small employers & own account	0.059	0.235
Lower supervisory & technical	0.042	0.202
Semi routine	0.089	0.285
Routine	0.052	0.222
Not in LM	0.427	0.495
Education		
Degree	0.233	0.423
Other higher degree	0.143	0.350
A-level etc	0.168	0.373
GCSE etc	0.190	0.392
Other qualification	0.120	0.325
No qualification	0.147	0.354
Gender		
Male	0.445	0.497
Female	0.555	0.497
House Ownership		
Owned	0.777	0.416
Rent	0.210	0.408
Other	0.013	0.112
Self-rated Heath		
Excellent	0.154	0.361
Very good	0.357	0.479
Good	0.288	0.453

Fair	0.152	0.359
Poor	0.048	0.214
Smoking Behavior		
Smoker	0.188	0.391
Ex smoker	0.415	0.493
Non smoker	0.397	0.489

Figure I.2 provides a graphical depiction of the bivariate association between CRP levels and occupational status. The depiction shows the kernel density distribution of CRP given the specific level of occupational status. The vertical lines represent the group-specific means of the distributions. Figure I.2 shows a clear graded association between CRP and occupational status. Indeed, the lower bound of occupational status has a higher CRP mean than individuals at the higher level. It is interesting to note that, from this exploratory relationship, individuals not in the labor market have a similar level of CRP, both nearby 2.5 mg/L.

Figure I.2: Kernel density estimates of CRP distribution according to the levels of occupational status.

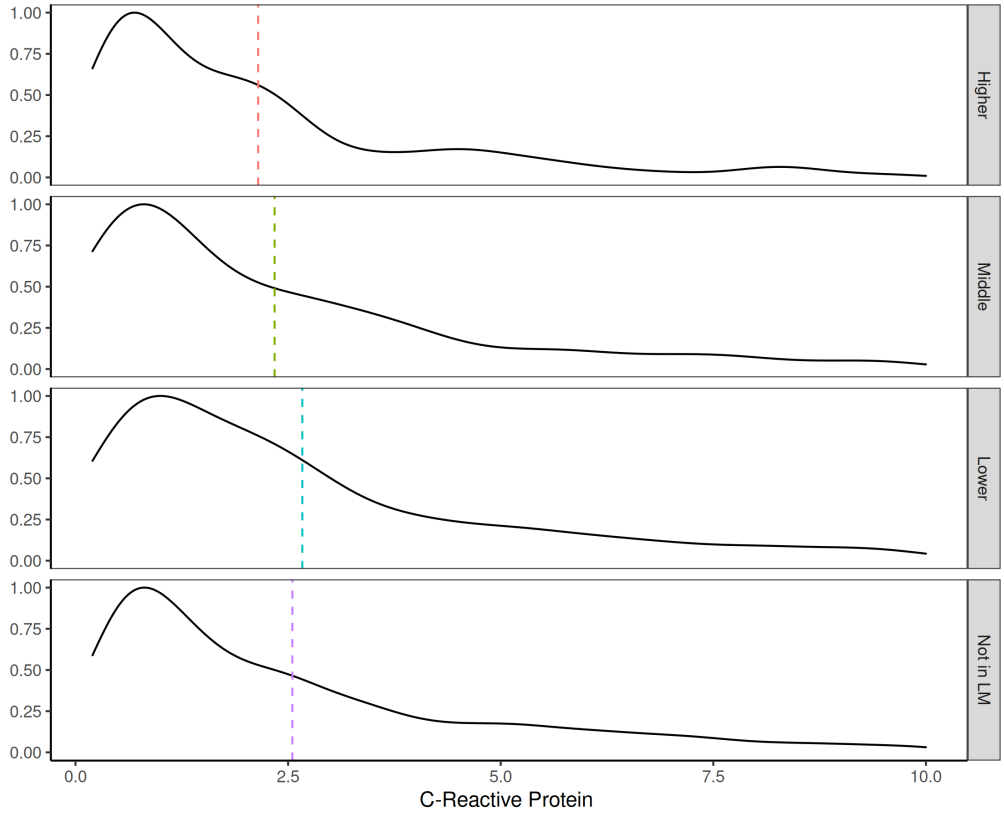
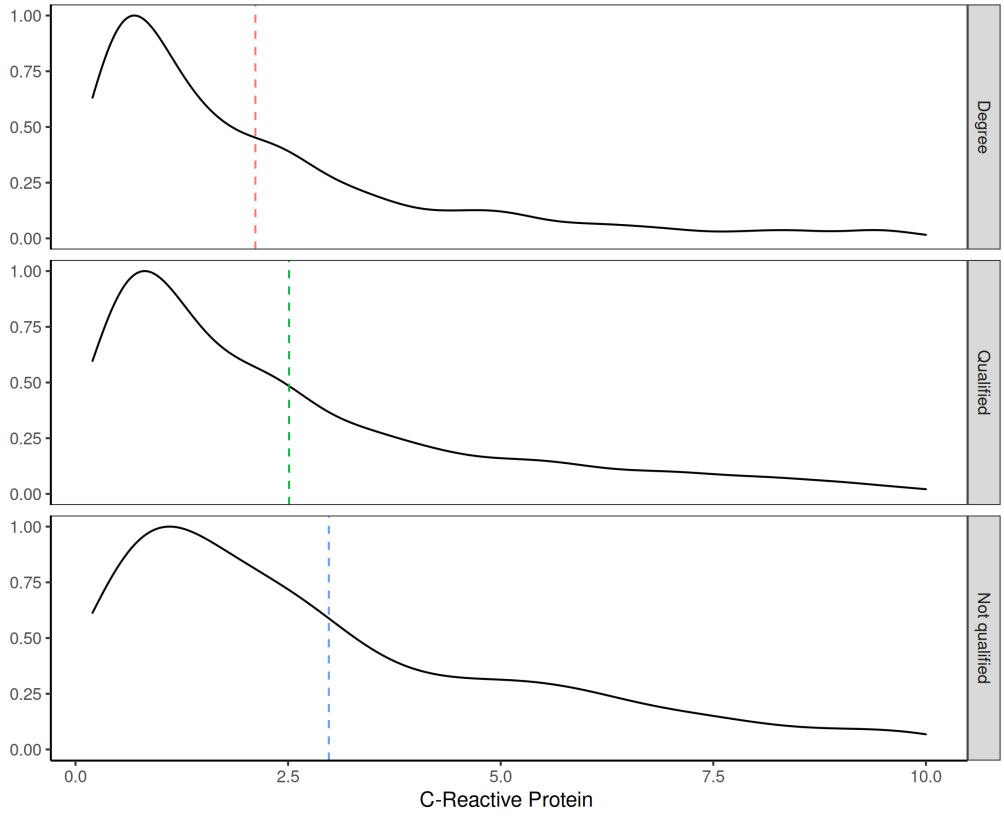


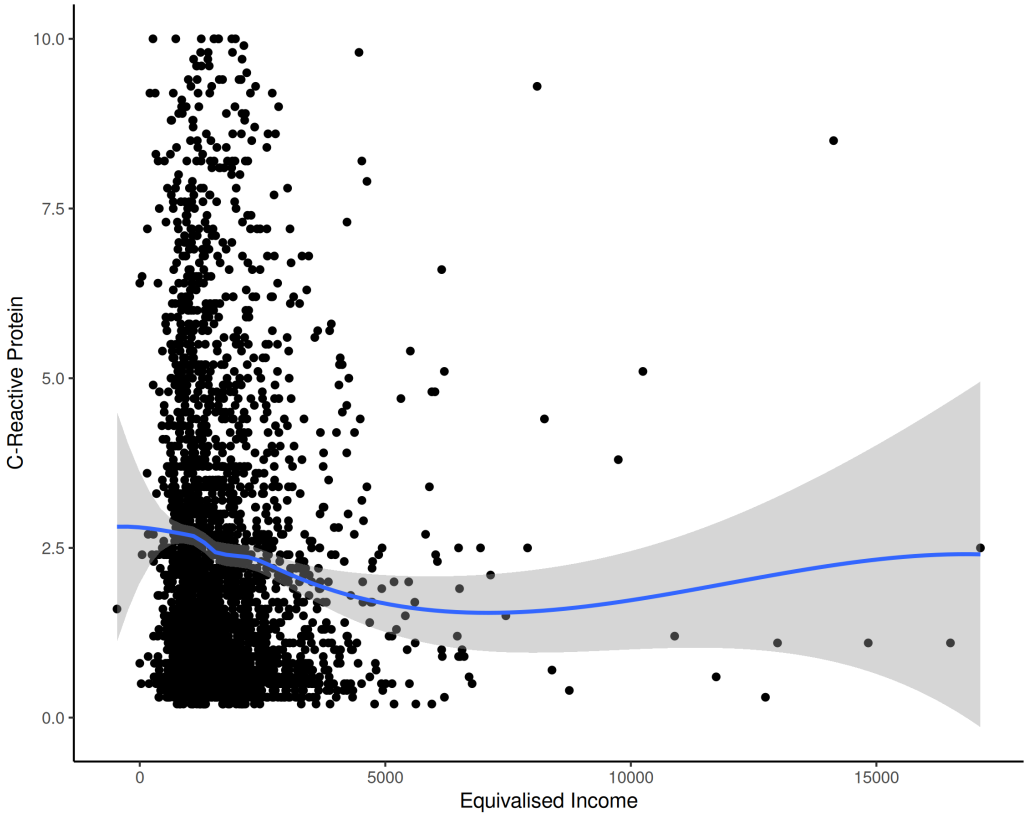
Figure I.3, similarly, depicts the kernel density estimates of CRP levels according to the level of educational attainment. From figure I.3, the bivariate relationship provides a more straightforward depiction of the social gradient of health. The mean of the CRP distribution for individuals with no qualifications is substantially higher than for individuals with an educational degree. Individuals who are not qualified tend to have a CRP level near the cut-off level of 3 mg/L of blood, indicating low-grade systemic inflammation and higher mortality risks due to CVD.

Figure I.3: Kernel density estimates of CRP distribution according to the levels of educational attainment.



Lastly, figure I.4 shows the relationship between CRP (on the y-axis) and equivalized income (on the x-axis). Both variables have been log-transformed in order to ease the depiction of the relationship. Figure I.4 shows the Locally Weighted Scatterplot Smoothing (LOWESS)[5] regression line and (in shaded gray area) the 95% confidence interval. From figure I.4 it is possible to see a slightly negative relationship, so as individuals with higher income tend to have lower levels of CRP. However, outliers might influence the association, in particular at the right tail of the equivalized income distribution.

Figure I.4: Scatter plot of (Logged) CRP on the y axis and (logged) equivalized income on the x axis. Lowess relationship in blue and relative confidence intervals in grey.



Statistical Analysis

The Bayesian Regression Model aims to draw a posterior distribution from all the possible values through Markov Chain Monte Carlo (MCMC) algorithm (Lynch & Bartlett, 2019). Integrating MCMC algorithms into modern Bayesian data analysis has been crucial to expanding such a class of models in empirical research. To see why in the past, the implementation of Bayesian models was almost practically impossible, we start with the fundamental Bayes rule, which states:

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)} \tag{I.1}$$

where $p(D)$ is

$$p(D) = \sum p(D|\theta)p(\theta) \tag{I.2}$$

Given the observed data D, the Bayesian framework aims to infer a likely distribution of a determined parameter θ, given the observed data D. As stated on the right-hand side of the Bayes' rule, the posterior distribution is the product of the data likelihood $p(D|\theta)$ and a prior distribution $p(\theta)$, divided by the marginal likelihood. The prior term $p(\theta$)is the distribution of credibilities that θ could take a particular value without considering the observed data. The denominator of the Bayes' rule defines the data likelihood, and it tells us the probability that the model with parameter $generatesthedatatheta$. The marginal likelihood informs us about the overall probability of the data $p(D|\theta)$, weighting these probabilities by the strength of their prior likelihood $p(\theta)$. In the case the researcher is interested in drawing a posterior distribution from a continuous variable, the marginal likelihood becomes:

$$p(D) = \int p(D|\theta)p(\theta)d\theta \tag{I.3}$$

In the context of the linear regression model, we can substitute the θ parameter with the typical coefficients of Ordinary Least Square (OLS): β_0 (the intercept), β_1 (the slope of the regression line), σ (the variance). In this context, the posterior distribution of the OLS parameters becomes:

$$p(\beta_0, \beta_1.\sigma|D) = \frac{p(D|\beta_0, \beta_1.\sigma)p(\beta_0, \beta_1.\sigma)}{\int\int\int p(D|\beta_0, \beta_1.\sigma)p(\beta_0, \beta_1.\sigma)\mathrm{d}\beta_0\mathrm{d}\beta_1\mathrm{d}\sigma} \tag{I.4}$$

Even when the sample size is small, the resolution of (in the simplest case, as shown) three integrals was cumbersome and time-consuming. The widespread use of MCMC algorithms in modern statistical analyses assisted the increase of empirical studies deploying the Bayesian framework. The Bayesian Regression models were performed through R, using as a backend the Stan program. Specifically, Stan uses a particular algorithm of MCMC, defined as Hamiltonian Markov Chain (HMC)[6].

The empirical analysis provides two types of Bayesian regression models.

Both models conceive a hierarchical structure of the statistical analysis. That means, within the Bayesian framework, the first step is to apply a determined distribution (a gaussian or t-student) that could best fit the dependent variable[7]. Model 1 applies to the (log) CRP distribution a t-student distribution. The choice of the t-student concerns the potential outliers present in the observed data, thus providing a more robust model. Model 1 sets the mean of the dependent variable as a linear function of the covariates shown before. The model exploits the hierarchical features of the Bayesian framework by calculating the deviations from the mean (as population-level effects) of the groups outlined by the two categorical variables representing individuals' SES: occupation and education. The specification of Model 1 takes the form of:

$$\begin{aligned} \mu_y &= \beta_0 + \sum_j \beta_i X_i + \beta_j X_j \qquad \text{(I.5a)} \\ p(\beta_0) &= \mathcal{N}(\mu_y, \sigma_y) \\ p(\beta_i) &= \mathcal{C}(0, \sigma_y) \\ p(\beta_j) &= \mathcal{N}(0, 1) \\ p(\sigma) &= \mathcal{N}(0, \sigma_y) \\ p(\nu) &= \mathcal{E}(1/29) \end{aligned}$$

The prior distributions of Model 1 and Model 2 follow the suggestions from Gelman (2006), and Kruschke (2014). The hierarchical structure of Model 1 allows setting a prior distribution for the deviations from the mean of the SES categorical variables $p(\beta_i)$. The prior distribution for the parameter β_i, as suggested by the previous literature, follows a (half-) Cauchy distribution with shape and scale parameters 0, σ_y. The scale parameter at zero has a twofold function. As the prior distribution should inform the likely parameter regarding the deviations from the mean of the dependent variable of the categories of the SES measures Occupation and Education, the average deviation should be 0. Secondly, for efficiency reasons (i.e., the MCMC would not sample from implausible values), we want the values sampled from the MCMC to be not too far from the mean of the dependent variable. The prior distribution for the equivalized income (and the other continuous covariates) $p(\beta_j)$ informs model 1 which parameters (among all the possible in the hyperparameter space) are more likely for the slope coefficients. As we have centered the continuous variables as $X_i(c) = X_i - \bar{X}_i$, the prior distribution for the slopes is normally shaped with mean 0 and a standard

deviation of 1. Finally, $p(\nu)$ is the exponentially distributed prior for the ν parameter of the t-student distribution. The ν parameter shapes the thickness of the tails of the distributions, thus accommodating the outliers. Model 2 exploits the unique feature of the Bayesian framework, that is, the possibility to model also the variance across groups through distributional models. The difference between model 1 and model 2 is, thus, in model 1 only the mean parameter can depend on predictors while σ_y is assumed to be constant across observations. Conversely, in model 2 both μ_y and σ_y can be objective of the statistical modeling, relaxing thus the assumption of homogeneity of variance. Therefore, the specification of model 2 is:

$$\begin{aligned}
\mu_y &= \beta_0 + \sum_j \beta_i X_i + \beta_j X_j && \text{(I.6a)} \\
\ln(\sigma_y) &= \beta_0 + \sum_j \beta_i X_i + \beta_j X_j \\
p(\sigma_\sigma) &= \mathcal{N}(0, \sigma_y) \\
p(\beta_0) &= \mathcal{N}(\mu_y, \sigma_y) \\
p(\sigma_{\beta_i}) &= \mathcal{N}(0, 1) \\
p(\beta_i) &= \mathcal{C}(0, \sigma_y) \\
p(\beta_j) &= \mathcal{N}(0, 1) \\
p(\nu) &= \mathcal{E}(1/29)
\end{aligned}$$

The difference between Model 1 and Model 2 is that in the latter, we apply the same linear function also for the log-transformed[8] standard deviation of the dependent variable. The two new prior distributions are $p(\sigma_\sigma)$ and $p(\sigma_{\beta_i})$. The prior distribution $p(\sigma_\sigma)$ defines the our expectations for the standard deviation of the σ distribution of the dependent variable. The prior $p(\sigma_{\beta_i})$ defines the *a priori* distribution for the deviations relative to the categories of the occupational status and education. Lastly, we have performed 4 MCMC chains, which included 2000 iterations. We set the burn-in (i.e., the number of initial iterations not considered due to strong autocorrelation) as the first 1000 iterations.

I.3 Results

Model 1

Beginning with the results provided by Model 1, table I.3 shows the summary of the findings by Model 1. We focus on the results concerning the effects of Occupation, Education, and income on the distribution of log-CRP. Figure I.5 shows the posterior distributions of contrasts against the reference category (i.e., Large employers & higher management) drawn by the MCMC for what concerns the occupational status. Under each distribution, the dot represents the expected value (i.e., the mean) while the thicker lines around the 65% of the posteriors' probability density function (PDF). The dashed black line serves as a reference for null-divergence of the distribution toward the grand mean of the dependent variable. Figure I.5 shows that small employers, in comparison with the other categories, have a lower concentration of CRP, thus a less risk of CVD. Surprisingly, the lower categories of the occupational status (Lower supervisory and technical staff, Semi-routine, Routine, and individuals not in the labor market) do not show remarkable differences.

Figure I.5: Posterior Distributions of the likely deviationsσ_{β_i} from the mean according to occupational status.

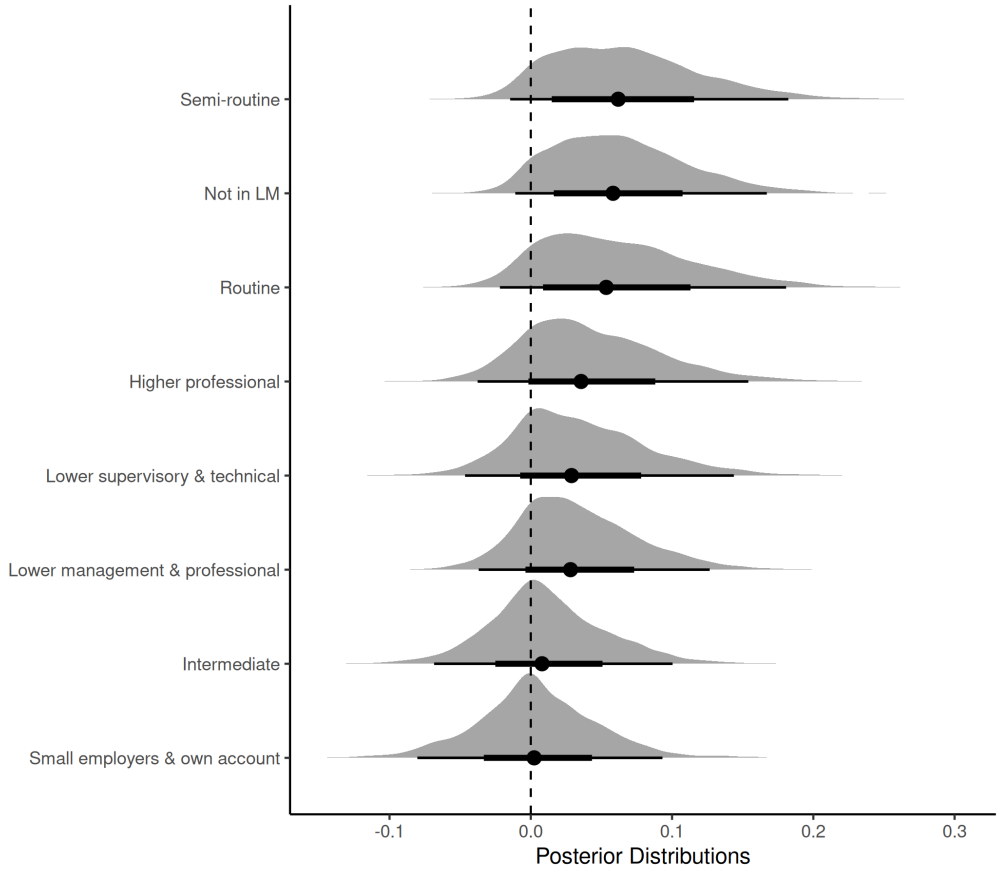


Table I.2: Bayesian Regression results, relaxing the assumption of homogeneity of variance

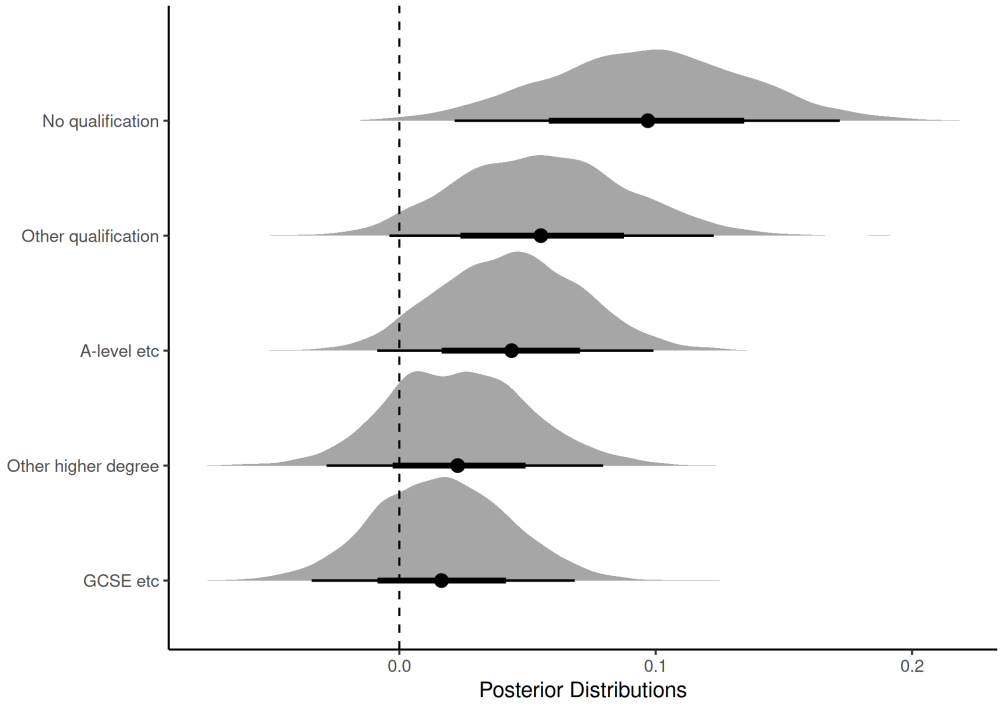
Parameter	Rhat	mean	sd	2.5%	50%	97.5%
Intercept	1.0027	0.2890	0.0460	0.1987	0.2894	0.3766
σ Intercept	1.0000	-0.1890	0.0331	-0.2556	-0.1883	-0.1246
Gender: Female						
Gender (female)	0.9997	0.1293	0.0189	0.0925	0.1295	0.1663
σ Gender (female)	0.9995	0.0096	0.0165	-0.0226	0.0096	0.0418
House Ownership: Owner						
Rent	1.0004	0.0584	0.0248	0.0115	0.0588	0.1071
Other	1.0000	0.0647	0.0762	-0.0866	0.0643	0.2141
σRent	1.0001	0.0141	0.0213	-0.0267	0.0140	0.0571
σOther	0.9992	-0.0452	0.0717	-0.1798	-0.0481	0.0978
SRH: Excellent						
Very good	1.0031	0.0661	0.0278	0.0138	0.0659	0.1213
Good	1.0017	0.0841	0.0298	0.0269	0.0845	0.1404
Fair	1.0016	0.1527	0.0351	0.0842	0.1520	0.2230
Poor	1.0015	0.2452	0.0501	0.1469	0.2459	0.3401
σVery good	1.0016	0.0139	0.0238	-0.0323	0.0142	0.0611
σGood	1.0021	0.0432	0.0249	-0.0051	0.0435	0.0911
σFair	1.0018	0.0402	0.0300	-0.0179	0.0396	0.0977
σPoor	1.0009	0.0097	0.0438	-0.0772	0.0104	0.0964
Smoking: Smoker						
Ex smoker	1.0012	-0.1782	0.0278	-0.2321	-0.1781	-0.1226
Non smoker	1.0011	-0.2138	0.0274	-0.2675	-0.2137	-0.1609
σEx smoker	1.0006	-0.0321	0.0227	-0.0756	-0.0323	0.0130

σNon smoker	1.0009	-0.0576	0.0226	-0.1025	-0.0575	-0.0144
Sport Activity	0.9996	-0.0625	0.0105	-0.0828	-0.0625	-0.0419
σSport Activity	0.9995	0.0046	0.0091	-0.0131	0.0047	0.0225
Age	0.9993	0.0805	0.0118	0.0573	0.0807	0.1034
σ Age	1.0000	-0.0228	0.0099	-0.0423	-0.0228	-0.0036
Income	0.9996	-0.0148	0.0104	-0.0349	-0.0148	0.0057
σ Income	0.9997	0.0031	0.0090	-0.0139	0.0030	0.0209
BMI	0.9997	0.3326	0.0094	0.3136	0.3327	0.3507
σBMI	0.9994	-0.0443	0.0083	-0.0603	-0.0443	-0.0281
Std Education (Intercept)	1.0015	0.0510	0.0316	0.0108	0.0446	0.1298
Std Occupation (Intercept)	1.0022	0.0404	0.0236	0.0060	0.0368	0.0971
σStd Education (Intercept)	1.0028	0.0171	0.0162	0.0007	0.0130	0.0586
σStd Occupation (Intercept)	1.0021	0.0312	0.0165	0.0044	0.0289	0.0705
ν	0.9992	177.1761	52.8198	97.2784	170.0995	296.5170
Education Intercepts						
Degree	1.0025	-0.0370	0.0316	-0.1065	-0.0341	0.0161
Other higher degree	1.0006	-0.0143	0.0312	-0.0824	-0.0124	0.0429
A-level	1.0006	0.0040	0.0300	-0.0603	0.0039	0.0636
GCSE	1.0017	-0.0208	0.0308	-0.0886	-0.0186	0.0356
Other qualification	1.0009	0.0102	0.0318	-0.0515	0.0085	0.0770
No qualification	1.0019	0.0524	0.0349	-0.0049	0.0501	0.1283
Occupation Intercepts						
Large employers & higher management	1.0004	-0.0325	0.0377	-0.1232	-0.0265	0.0268
Higher.professional	0.9999	0.0015	0.0306	-0.0592	0.0006	0.0663
Lower management & professional	0.9994	0.0005	0.0244	-0.0507	0.0003	0.0505
Intermediate	1.0006	-0.0208	0.0280	-0.0836	-0.0181	0.0274

Small employers & own account	1.0006	-0.0291	0.0322	-0.1020	-0.0253	0.0239
Lower supervisory & technical	0.9993	-0.0007	0.0309	-0.0645	-0.0010	0.0647
Semi-routine	0.9998	0.0271	0.0305	-0.0224	0.0228	0.0962
Routine	1.0000	0.0247	0.0317	-0.0287	0.0205	0.0958
Not in LM	0.9997	0.0266	0.0255	-0.0166	0.0242	0.0812
σ **Education, Intercepts**						
Degree	1.0003	0.0009	0.0137	-0.0260	0.0001	0.0312
Other higher degreee	1.0002	0.0037	0.0147	-0.0227	0.0016	0.0383
A-level	1.0013	0.0095	0.0163	-0.0155	0.0054	0.0510
GCSE	1.0002	-0.0090	0.0157	-0.0481	-0.0054	0.0151
Other qualification	1.0012	0.0032	0.0152	-0.0247	0.0011	0.0388
No qualification	0.9996	-0.0048	0.0154	-0.0411	-0.0021	0.0242
σ **Occupation, Intercepts**						
Large employers.&.higher management	0.9995	-0.0038	0.0262	-0.0578	-0.0024	0.0487
Higher professional	1.0010	0.0194	0.0268	-0.0266	0.0155	0.0811
Lower management & professional	1.0000	-0.0257	0.0214	-0.0724	-0.0244	0.0113
Intermediate	0.9996	-0.0069	0.0220	-0.0521	-0.0054	0.0366
Small employers & own account	1.0002	-0.0188	0.0252	-0.0747	-0.0156	0.0252
Lower supervisory & technical	1.0005	-0.0073	0.0245	-0.0618	-0.0055	0.0388
Semi-routine	1.0010	0.0183	0.0233	-0.0223	0.0157	0.0704
Routine	0.9996	-0.0021	0.0236	-0.0506	-0.0016	0.0472
Not in LM	1.0003	0.0261	0.0203	-0.0082	0.0247	0.0708
log-posterior	1.0038	-10613.8337	7.0487	-10627.9000	-10613.7000	-10600.7000

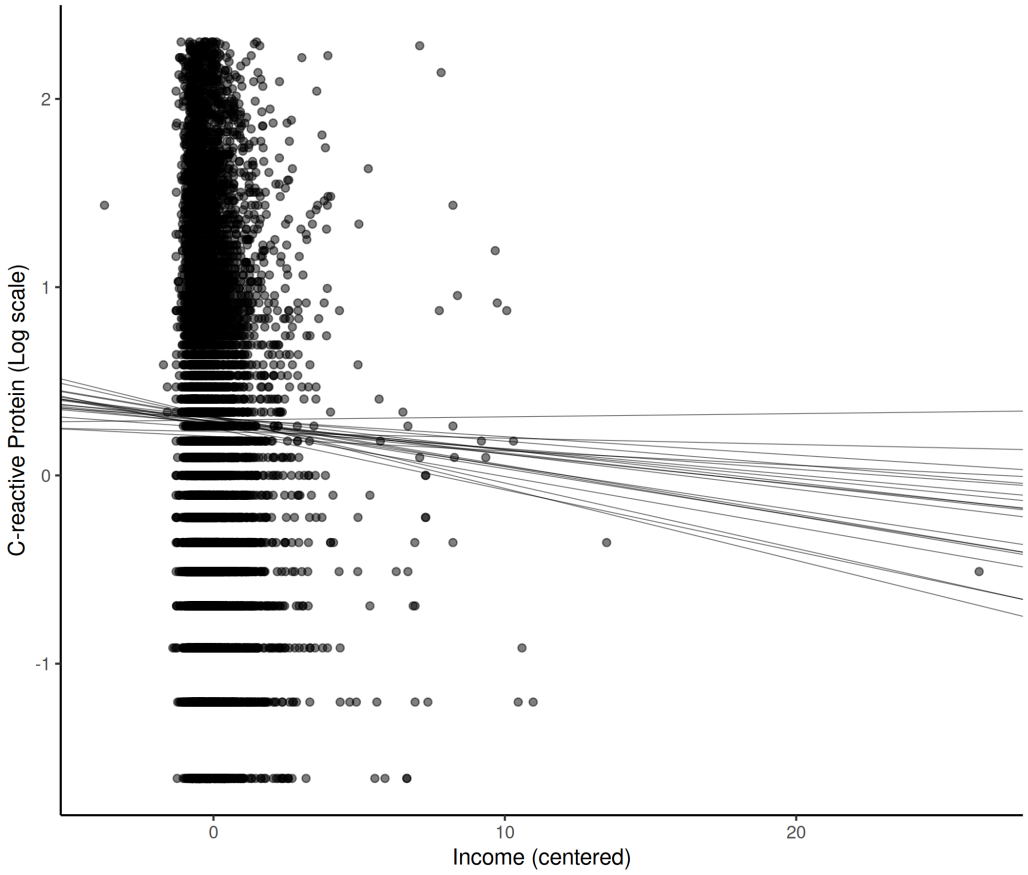
For what concerns the educational level and CRP concentration, the second SES measure shows a clear higher level of CRP for individuals with no educational qualification concerning the referent category (i.e., individuals with educational degree), as depicted in figure I.6. On the opposite, individuals with higher educational degrees (e.g., individuals with GCSE and higher degrees) are more likely to have lower CRP concentrations in the blood, thus fewer risks of CVD mortality.

Figure I.6: Posterior Distributions of the likely deviations from the mean according to Educational levels. σ_{β_i}.



To depict the relationship between household income (equivalized) and levels of CRP, we make use of scatterplots where the x-axis represents the levels of household income (in the centered scale) and the levels of (log) CRP on the y-axis. Figure I.7, shows this relationship. Conversely, to the traditional frequentist approach, it is possible to visualize different likely regression lines in the Bayesian framework, as the MCMC samples from the posterior distribution of likely regression slopes. In figure I.7 we show 20 possible regression lines assessing the relationship between income and CRP.

Figure I.7: Plot of Income distribution (on the x-axis) and log-CRP (on the y-axis) and model fit of 20 possible regression lines sampled from the posterior distribution β Income of Model 1.



From figure I.7, it is possible to recognize a general negative relationship between equivalized income and levels of CRP. Thus, economic resources positively impact the risks of CVD measured through CRP. However, the slopes of the regression lines are not relatively steep, suggesting a mild relationship. Even more, from figure I.7, it seems that Model 1 is affected by the leverage effect.

Model 2

In model 2, we relaxed the homogeneity of variance assumption among the categories of occupational status and educational levels, thus predicting the mean of the dependent variable μ and its standard deviation σ. This

subsection shows the results from modeling the σ parameter. As in the previous section, we firstly provide a tabular format of the results provided by model 2 as in table I.2. The table makes it possible to find the results in modeling the parameter μ and the σ parameters.

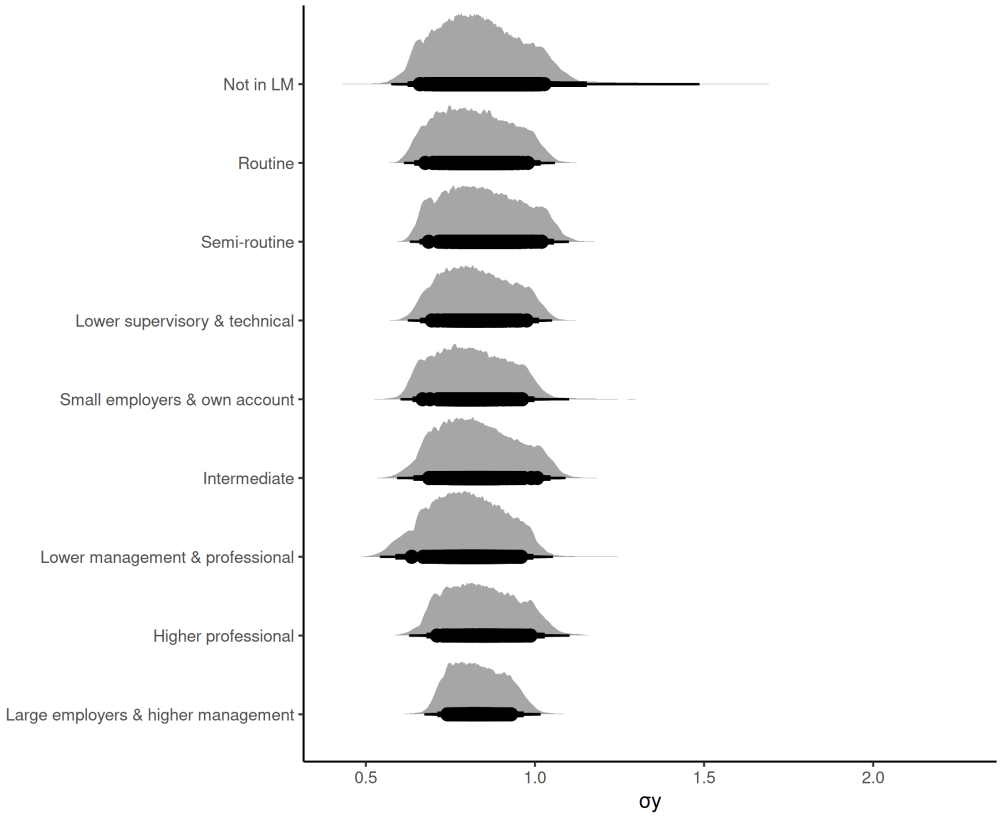
Table I.3: Bayesian Regression results, assuming homogeneity of variance

Parameter	Rhat	mean	sd	2.5%	50%	97.5%
Intercept	1.0024	0.2916	0.0470	0.1991	0.2918	0.3846
Gender: Female						
Gender (Female)	1.0001	0.1226	0.0184	0.0875	0.1230	0.1581
House Ownership: Owner						
Rent	0.9996	0.0594	0.0246	0.0123	0.0597	0.1065
Other	0.9995	0.0615	0.0810	-0.0984	0.0620	0.2232
SRH: Excellent						
Very good	0.9997	0.0649	0.0281	0.0093	0.0647	0.1198
Good	1.0002	0.0828	0.0299	0.0216	0.0835	0.1408
Fair	0.9997	0.1544	0.0357	0.0834	0.1544	0.2246
Poor	1.0002	0.2478	0.0522	0.1485	0.2468	0.3527
Smoking: Smoker						
Ex smoker	0.9997	-0.1796	0.0267	-0.2325	-0.1797	-0.1248
Non smoker	0.9998	-0.2081	0.0269	-0.2618	-0.2080	-0.1551
Sport Activity	0.9993	-0.0617	0.0102	-0.0817	-0.0617	-0.0418
Age	0.9996	0.0814	0.0118	0.0586	0.0813	0.1047
Income	0.9999	-0.0147	0.0101	-0.0342	-0.0150	0.0050
BMI	0.9998	0.3354	0.0095	0.3163	0.3355	0.3542
Std Education (Intercept)	1.0029	0.0566	0.0362	0.0135	0.0485	0.1528
Std Occupation (Intercept)	1.0007	0.0426	0.0241	0.0061	0.0392	0.0980
σ	0.9997	0.8330	0.0066	0.8201	0.8329	0.8461
ν	0.9996	176.4664	52.4283	96.4933	168.2630	297.1096
Education, Intercepts						

Degree	1.0069	-0.0401	0.0340	-0.1121	-0.0376	0.0168
Other higher degree	1.0060	-0.0167	0.0335	-0.0846	-0.0147	0.0454
A-level	1.0060	0.0030	0.0328	-0.0643	0.0039	0.0660
GCSE	1.0048	-0.0230	0.0325	-0.0915	-0.0215	0.0395
Other qualification	1.0050	0.0160	0.0353	-0.0496	0.0142	0.0932
No qualification	1.0030	0.0558	0.0376	-0.0061	0.0533	0.1380
Occupation, Intercepts						
Large employers.& higher management	0.9999	-0.0348	0.0374	-0.1223	-0.0291	0.0239
Higher professional	0.9996	0.0068	0.0317	-0.0559	0.0053	0.0751
Lower management & professional	1.0006	-0.0010	0.0249	-0.0508	-0.0012	0.0509
Intermediate	0.9991	-0.0237	0.0291	-0.0868	-0.0209	0.0291
Small employers & own account	0.9995	-0.0319	0.0334	-0.1055	-0.0287	0.0243
Lower supervisory &.technical	0.9994	-0.0003	0.0317	-0.0671	-0.0007	0.0665
Semi-routine	1.0001	0.0301	0.0310	-0.0198	0.0268	0.0999
Routine	0.9999	0.0254	0.0327	-0.0278	0.0207	0.1004
Not in LM	0.9998	0.0273	0.0249	-0.0162	0.0251	0.0814
log-posterior	1.0004	-10611.3258	5.0972	-10622.2000	-10611.0000	-10602.5000

Here we focus on the results produced by fitting the model in the σ of the dependent variable. Similarly to the section dedicated to the results related to model 1, we start with presenting the posterior distributions related to the occupational status of individuals.
Figure I.8 shows the posterior distributions for each category of occupational status and how they deviate according to the scale of the standard deviation of the CRP levels.

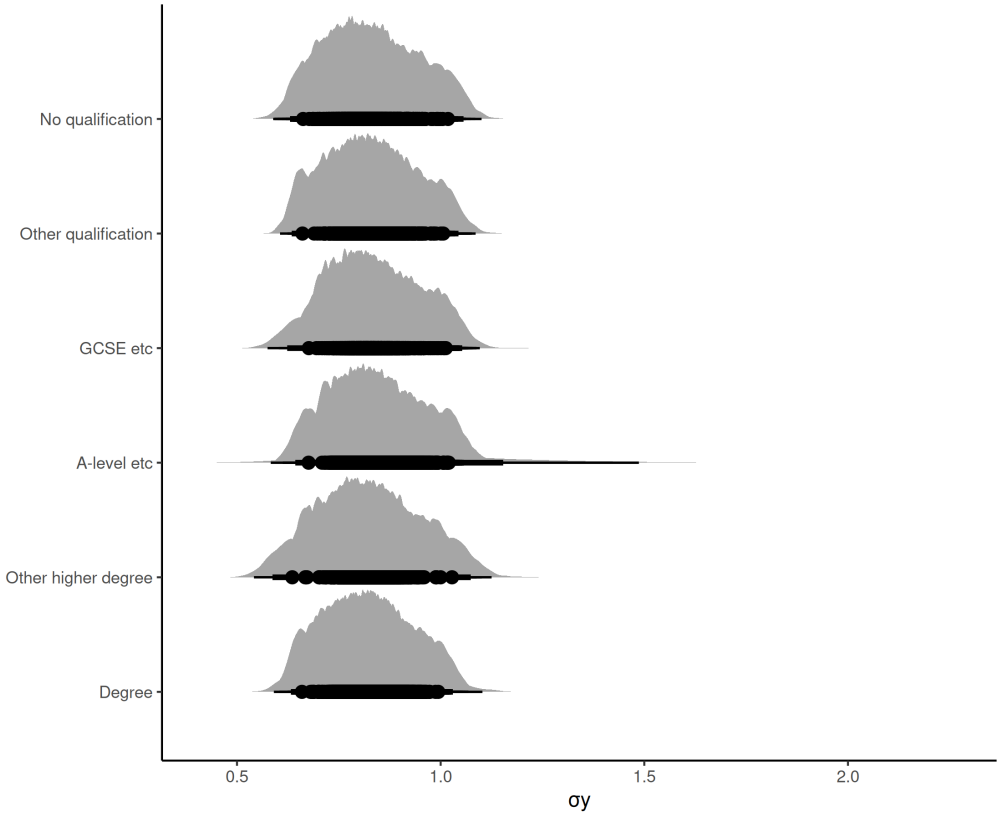
Figure I.8: Posterior Distributions of the likely deviations from the mean according to Occupational status. σ Occupation on the σ_y.



When the model relaxes the assumption of homogeneity of variance, the findings suggest a rather clear difference of variance between the Large employers and higher management category and individuals not in the labor market. The in-between categories, however, show a similar pattern.
Moving to the effects of educational status on the variance between categories on the variance of CRP observed through the data, figure I.9 shows the findings of Model 2 focused on individuals' education.
From figure I.9, the findings of Model 2 shows a surprising similarity between

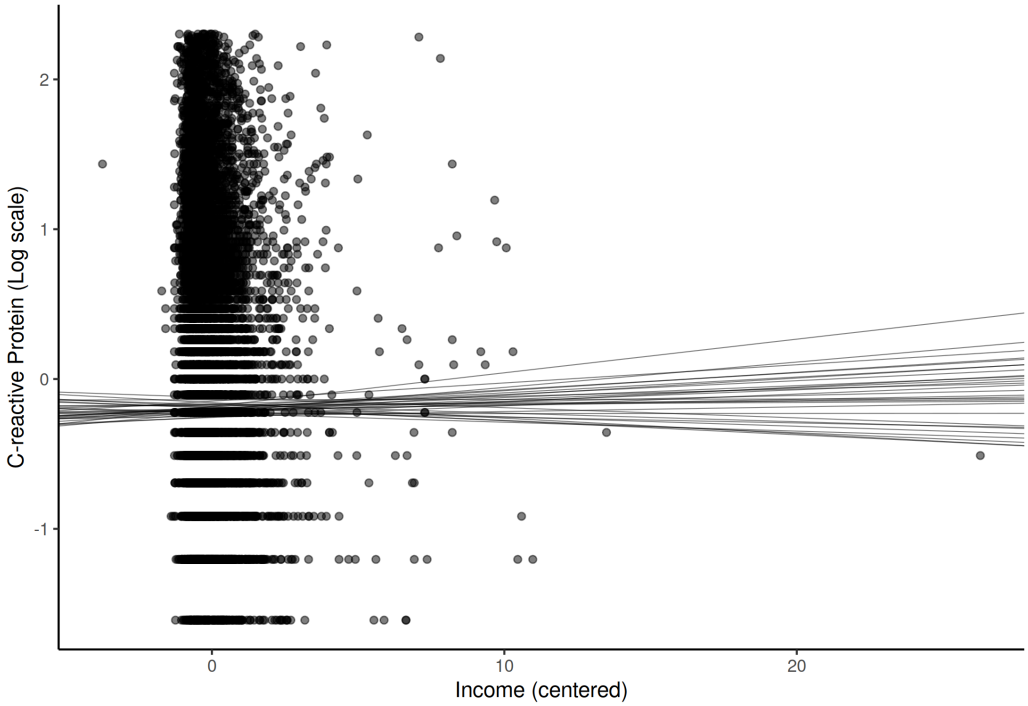
the educational levels, with the notable exception of the individuals with an A-level. In fact, from Model 2, the variance for the individuals with an A-level educational degree is stretched toward the right.

Figure I.9: Posterior Distributions of the likely deviations from the mean according to Educational levels. σ Education on the σ_y.



Similar to Model 1, to present the findings from model 2 for what concerns the relationship between equivalized income and logged CRP, figure I.10 depicts 20 sampled regression lines from the posterior distribution of the regression slopes. However, in this case, the intercepts and slopes are computed according to σ_y. From figure I.10, findings suggested an even weaker relationship between equivalized income and logged CRP, with the homogeneity of variance assumption relaxed in the model.

Figure I.10: Plot of Income distribution (on the x-axis) and log-CRP (on the y-axis) and model fit of 20 possible regression lines sampled from the posterior distribution β Income of Model 2.



In the Supplementary Materials, we provide the trace and the autocorrelation plots. These are valuable tools to assess the validity of the Bayesian inference and the correct sampling of the algorithm from the posterior distributions of the specific parameters.

I.4 Discussion & Conclusions

One of the main challenges currently in social stratification and health inequalities is to assess the mechanisms through which the well-documented link between deprived social conditions and health worsening takes form. The rising availability of social data that includes biological information of individuals is rising as a propitious track of research, and it has a possible double effect. On the one hand, social scientists acquire additional information potentially helpful to studying the mechanisms through which the social gradient of health occurs and socioeconomic status gets under the skin. Secondly, sociologists inform the biomedical literature on the importance of the social and economic environment individuals live in for health levels. This

paper investigated one of the potential mechanisms that health inequalities generate. Indeed, this study aimed to shed light on the connection between socioeconomic status (SES), risks of mortality due to cardiovascular diseases (CVD), and levels of chronic inflammation, measured through the biomarker C-reactive protein (CRP). We consider this pathway particularly important as mortality due to CVD represents the first cause of death among individuals in developed countries (Brummett et al., 2014). We used data from the Understanding Society - United Kingdom Household Longitudinal Study (UKHLS), a representative sample of the population living in England, Scotland, and Wales. In 2012, the UKHLS collected voluntary individuals' blood samples to collect markers of socially relevant health risks factors and diseases. To capture the multifaceted characteristics of individuals' SES, we included three measures of social conditions in the statistical analysis: occupational status, educational levels, and equivalized household income. This paper provides a Bayesian framework to assess the pathway that links individuals' SES, mortality risks to CVD, and levels of CRP. The first Bayesian regression model (BRM) provides posterior distributions of likely effect magnitude parameters assuming homogeneity of variance across occupational and educational groups. The second step of the statistical analysis deploys a distributional model, which allows for relaxing the assumption of homogeneity of variance through modeling the standard deviation of CRP alongside its mean. Generally, both Model 1 and Model 2 show no inference problems, meaning that the posterior distributions computed by both models represent all possible distributions. For what concerns findings provided by model 1, the educational gradient is the most vital determinant of the risk of chronic inflammation, while equivalized income is the weakest among the three SES measures. According to Davillas et al. (2017), one potential explanation relies on the fact that better-educated individuals tend to pursue a healthier lifestyle and be more aware of health risks. The relationship between occupational status and levels of CRP seems to be polarized. Indeed, the findings suggest a homogeneity of CRP levels and lowest levels (and thus less risk of CVD) between individuals with the highest occupational status and small employers. That means they do not deviate remarkably from the grand mean of the dependent variable among all the other categories. While the first findings are coherent with previous literature (Marmot et al. 1991), we strongly suggest that future research deepen scientific knowledge of what concerns small employers. For what concerns the equivalized income measure, some lessons can be from both the methodological and substantial perspective: the methodological conclusion is that even in the Bayesian framework, the regression model could suffer from leverage effects. The significant lesson is that economic resources are somewhat the weakest determinants to drive health inequalities

and mortality risks due to CVD.
Model 2 provided interesting results for what concerns how the individuals differ from within each category of the SES measures. Indeed, the findings suggest a strong cohesion in the highest class of occupational status, meaning that inequalities in health are evident even when we consider the within variance. Surprisingly, individuals with the same educational level are not very dissimilar, except for individuals with an A-level degree.
It is worth to note that the meaning of income may change during the different stages of the life course we considered in the analysis. We therefore suggest to investigate further on this aspect by, for example, considering a more homogenous sample in terms of age.
We would like to address some limitations of this study as inspiration for future researchers. The first limitation is the lack of direct comparison between the posterior distributions explicitly drawn for the categories of income, education, and occupational status. Further research could address this problem ideally from a Bayesian perspective. The second limitation concerns the model specification. Due to the already complex Bayesian models, we did not test the models for nonlinearities in age patterns and CVD risks through CRP. We believe it could be interesting to see whether the aging process could take other, more complex ways. The last limitation concerns robustness checks. Indeed, we have specified the models only with this set of prior distributions. In this sense, we conclude with an invitation to the social scientific community to use our results for better-refined models coherently with the Bayesian philosophy. The Understanding Society data provides to the empirical researcher a wide range of biological markers of (ab-) normal physiological functioning. Future research could potentially exploits the biological information collected by empirically testing the social gradient of health by using composite measures such as Allostatic Load, for instance. We believe that future researchers could take advantage of this first Bayesian implementation and use the results we provided as a starting point to define a theoretically guided model to advance our knowledge concerning the social gradient of mortality risks due to CVD.

Notes

[3]The UKHLS is a continuation and further evolution of the older British Household Panel Study (BHPS).

[4]Alongside the GPS, the main survey target sample consisted of three additional components: the Ethnic Minority Boost sample, the former BHPS sample, and the Immigrant and Ethnic Minority Boost sample. See Lynn (2009) and Lynn et al. (2018) for additional details.

[5]LOWESS is a type non-parametric regression estimation

[6]More precisely, Stan's specification of the HMC is the No-U-Turn Shape (NUTS) algorithm.

[7]The choice of the distribution is dependent upon the type of variable the outcome is: for instance, the Poisson distribution applies with count data; for instance, if the analysis aims to model a dichotomous variable, then the bayesian framework will start with a binomial distribution.

[8]We have transformed the standard deviation as recommended by the Stan Development Team.

Bibliography

Adler, N. E., Boyce, T., Chesney, M. A., Cohen, S., Folkman, S., Kahn, R. L., & Syme, L. S. (1994). Socioeconomic Status and Health: The Challenge of the Gradient. *American Psychologist*, *49*(1), 15–24.

Adler, N. E., & Ostrove, J. M. (1999). Socioeconomic status and health: What we know and what we don't. *Annals of the New York Academy of Sciences*, *896*, 3–15.

Alley, D. E., Seeman, T. E., Ki Kim, J., Karlamangla, A., Hu, P., & Crimmins, E. M. (2006). Socioeconomic status and C-reactive protein levels in the US population: NHANES IV. *Brain, Behavior, and Immunity*, *20*(5), 498–504.

Baum, A., Garofalo, J. P., & Yali, A. M. (1999). Socioeconomic status and chronic stress. Does stress account for SES effects on health? *Annals of the New York Academy of Sciences*, *896*, 131–144.

Ben-Shlomo, Y., & Kuh, D. (2002). A life course approach to chronic disease epidemiology: Conceptual models, empirical challenges and interdisciplinary perspectives. *International Journal of Epidemiology*, *31*(2), 285–293.

Brummett, B. H., Babyak, M. A., Singh, A., Jiang, R., Williams, R. B., Harris, K. M., & Siegler, I. C. (2014). Socioeconomic indices as independent correlates of C-reactive protein in the National Longitudinal Study of Adolescent Health. *Psychosomatic Medicine*, *75*(9), 882–893.

Clark, A. M., DesMeules, M., Luo, W., Duncan, A. S., & Wielgosz, A. (2009). Socioeconomic status and cardiovascular disease: Risks and implications for care. *Nature Reviews Cardiology*, *6*(11), 712–722.

Corna, L. M. (2013). A life course perspective on socioeconomic inequalities in health: A critical review of conceptual frameworks. *Advances in Life Course Research*, *18*(2), 150–159.

Davillas, A., Benzeval, M., & Kumari, M. (2017). Socio-economic inequalities in C-reactive protein and fibrinogen across the adult age span: Findings from Understanding Society. *Scientific Reports*, *7*(1), 1–13.

Davillas, A., Jones, A. M., & Benzeval, M. (2019). The Income-Health Gradient: Evidence From Self-Reported Health and Biomarkers in Understanding Society. In *Panel data econometrics: Empirical applications* (pp. 709–741). Elsevier Inc.

Dowd, J. B., & Zajacova, A. (2007). Does the predictive power of self-rated health for subsequent mortality risk vary by socioeconomic status in the US? *International Journal of Epidemiology*, *36*(6), 1214–1221.

Dowd, J. B., Zajacova, A., & Aiello, A. (2009). Early origins of health disparities: Burden of infection, health, and socioeconomic status in U.S. children. *Social Science and Medicine*, *68*(4), 699–707.

Elo, I. T. (2009). Social class differentials in health and mortality: Patterns and explanations in comparative perspective. *Annual Review of Sociology*, *35*, 553–572.

Ferraro, K. F., & Shippee, T. P. (2009). Aging and cumulative inequality: How does inequality get under the skin? *Gerontologist*, *49*(3), 333–343.

Freese, J. (2018). The Arrival of Social Science Genomics. *Contemporary Sociology*, *47*(5), 524–536.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian analysis*, *1*(3), 515–534.

Gimeno, D., Ferrie, J. E., Elovainio, M., Pulkki-Raback, L., Keltikangas-Jarvinen, L., Eklund, C., Hurme, M., Lehtimäki, T., Marniemi, J., Viikari, J. S., Raitakari, O. T., & Kivimäki, M. (2008). When do social inequalities in C-reactive protein start? A life course perspective from conception to adulthood in the Cardiovascular Risk in Young Finns Study. *International Journal of Epidemiology*, *37*(2), 290–298.

Goldthorpe, J. H. (2010). Analysing Social Inequality : A Critique of Two Recent Contributions from Economics and Epidemiology. *European Sociological Review*, *26*(6), 731–744.

Goodman, E., McEwen, B. S., Huang, B., Dolan, L. M., & Adler, N. E. (2005). Social inequalities in biomarkers of cardiovascular risk in adolescence. *Psychosomatic Medicine*, *67*(1), 9–15.

Gruenewald, T. L., Karlamangla, A., Hu, P., Stein-Merkin, S., Crandall, C., Koretz, B., & Seeman, T. E. (2012). History of Socioeconomic Disadvantage and Allostatic Load in Later Life. *Social Science and Medicine*, *23*(1), 1–7.

Hallqvist, J., Lynch, J., Bartley, M., Lang, T., & Blane, D. (2004). Can we disentangle life course processes of accumulation, critical period and social mobility? An analysis of disadvantaged socio-economic positions and myocardial infarction in the Stockholm Heart Epidemiology Program. *Social Science and Medicine*, *58*(8), 1555–1562.

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

Harris, T. B., Ferrucci, L., Tracy, R. P., Corti, M. C., Wacholder, S., Ettinger, W. H., Heimovitz, H., Cohen, H. J., & Wallace, R. (1999). Associations of elevated interleukin-6 and C-reactive protein levels with mortality in the elderly. *American Journal of Medicine*, *106*(5), 506–512.

Herd, P., Goesling, B., & House, J. S. (2007). Socioeconomic position and health: The differential effects of education versus income on the onset versus progression of health problems. *Journal of Health and Social Behavior*, *48*(3), 223–238.

House, J. S. (2002). Understanding social factors and inequalities in health: 20th century progress and 21st century prospects. *Journal of Health and Social Behavior*, *43*(2), 125–142.

Jousilahti, P., Salomaa, V., Rasi, V., Vahtera, E., & Palosuo, T. (2003). Association of markers of systemic inflammation, C reactive protein, serum amyloid A, and fibrinogen, with socioeconomic status. *Journal of Epidemiology and Community Health*, *57*, 730–733.

Karimi, M., Castagné, R., Delpierre, C., Albertus, G., Berger, E., Vineis, P., Kumari, M., Kelly-Irving, M., Chadeau-Hyam, M., Lynch, S. M., & Bartlett, B. (2019). Early-life inequalities and biological ageing: A multisystem Biological Health Score approach in U nderstanding S ociety. *Journal of Epidemiology and Community Health*, *73*(8), 693–702.

Kavanagh, A., Bentley, R. J., Turrell, G., Shaw, J., Dunstan, D., & Subramanian, S. V. (2010). Socioeconomic position, gender, health behaviours and biomarkers of cardiovascular disease and diabetes. *Social Science and Medicine*, *71*(6), 1150–1160.

Koster, A., Bosma, H., Penninx, B. W. J. H., Newman, A. B., Harris, T. B., Eijk, J. T. M. V., Kempen, G. I. J. M., Simonsick, E. M., Johnson, K. C., Rooks, R. N., Ayonayon, H. N., Rubin, S. M., & Kritchevsky, S. B. (2006). Association of Inflammatory Markers With Socioeconomic Status. *Journal of Gerontology*, *61*(3), 284–290.

Kruschke, J. (2014). *Doing bayesian data analysis: A tutorial with r, jags, and stan*. Academic Press.

Laaksonen, D. E., Niskanen, L., Nyyssönen, K., Punnonen, K., Tuomainen, T. P., & Salonen, J. T. (2005). C-reactive protein in the prediction of cardiovascular and overall mortality in middle-aged men: A population-based cohort study. *European Heart Journal*, *26*(17), 1783–1789.

Link, B., & Phelan, J. (1995). Social conditions as fundamental causes of health inequalities. *Journal of Health and Social Behavior*, *35*, 80–94.

Liu, R. S., Aiello, A. E., Mensah, F. K., Gasser, C. E., Rueb, K., Cordell, B., Juonala, M., Wake, M., & Burgner, D. P. (2017). Socioeconomic status in childhood and C reactive protein in adulthood: A systematic review and meta-analysis. *Journal of Epidemiology and Community Health*, *71*(8), 817–826.

Loucks, E. B., Pilote, L., Lynch, J. W., Richard, H., Almeida, N. D., Benjamin, E. J., & Murabito, J. M. (2010). Life course socioeconomic position is associated with inflammatory markers: The Framingham Offspring Study. *Social Science and Medicine*, *71*(1), 187–195.

Lubbock, L. A., Goh, A., Ali, S., Ritchie, J., & Whooley, M. A. (2005). Relation of low socioeconomic status to C-reactive protein in patients with coronary heart disease (from the Heart and Soul study). *American Journal of Cardiology*, *96*(11), 1506–1511.

Lynch, S. M., & Bartlett, B. (2019). Bayesian Statistics in Sociology: Past, Present, and Future. *Annual Review of Sociology*, *45*, 47–68.

Lynn, P. (2009). *Sample Design for Understanding Society - Understanding Society Working Paper 2009-01.*

Lynn, P., Nandi, A., Parutis, V., & Platt, L. (2018). Design and implementation of a high-quality probability sample of immigrants and ethnic minorities: Lessons learnt. *Demographic Research*, *38*(1), 513–548.

Marmot, M. G., Stansfeld, S., Patel, C., North, F., Head, J., White, I., Brunner, E., Feeney, A., Marmot, M. G., & Smith, G. D. (1991). Health inequalities among British civil servants: the Whitehall II study. *The Lancet*, *337*(8754), 1387–1393.

McDade, T. W., Lindau, S. T., & Wroblewski, K. (2011). Predictors of C-reactive protein in the National Social Life, Health, and Aging Project. *Journals of Gerontology - Series B Psychological Sciences and Social Sciences*, *66 B*(1), 129–136.

McEwen, B. S. (2015). Biomarkers for assessing population and individual health and disease related to stress and adaptation. *Metabolism: Clinical and Experimental*, *64*(3), S2–S10.

Mcfall, S. L., Booker, C., Burton, J., & Conolly, A. (2012). *Implementing the Biosocial Component of Understanding Society – Nurse Collection of Biomeasures and Anne Conolly Nurse Collection of Biomeasures.*

Mitchell, U. A., & Aneshensel, C. S. (2017). Social Inequalities in Inflammation: Age Variations in Older Persons. *Journal of Aging and Health*, *29*(5), 769–787.

Nazmi, A., Oliveira, I. O., Horta, B. L., Gigante, D. P., & Victora, C. G. (2010). Lifecourse socioeconomic trajectories and C-reactive protein levels in young adults: Findings from a Brazilian birth cohort. *Social Science and Medicine*, *70*(8), 1229–1236.

Phelan, J. C., Link, B. G., & Tehranifar, P. (2010). Social Conditions as Fundamental Causes of Health Inequalities: Theory, Evidence, and Policy Implications. *Journal of Health and Social Behavior*, *51*, 28–40.

Pollitt, R. A., Kaufman, J. S., Rose, K. M., Diez-Roux, A. V., Zeng, D., & Heiss, G. (2008). Cumulative life course and adult socioeconomic status and markers of inflammation in adulthood. *Journal of Epidemiology and Community Health*, *62*(6), 484–491.

Power, C., Manor, O., & Matthews, S. (1999). The duration and timing of exposure: Effects of socioeconomic environment on adult health. *American Journal of Public Health*, *89*(7), 1059–1065.

Pudrovska, T. (2014). Early-Life Socioeconomic Status and Mortality at Three Life Course Stages: An Increasing Within-Cohort Inequality. *Journal of Health and Social Behavior*, *55*(2), 181–195.

Rosvall, M., Engström, G., Berglund, G., & Hedblad, B. (2008). C-reactive protein, established risk factors and social inequalities in cardiovascular disease - The significance of absolute versus relative measures of disease. *BMC Public Health*, *8*, 1–10.

Shanahan, L., Freeman, J., & Bauldry, S. (2014). Is very high C-reactive protein in young adults associated with indicators of chronic disease risk? *Psychoneuroendocrinology*, *40*(1), 76–85.

Umlauf, N., & Kneib, T. (2018). A primer on Bayesian distributional regression. *Statistical Modelling*, *18*(3-4), 219–247.

Winkleby, M. A., Jatulis, D. E., Frank, E., & Fortmann, S. P. (1992). Socioeconomic status and health: How education, income, and occupation contribute to risk factors for cardiovascular disease. *American Journal of Public Health*, *82*(6), 816–820.

Yang, Y. C., Schorpp, K., Boen, C., Johnson, M., & Harris, K. M. (2020). Socioeconomic Status and Biological Risks for Health and Illness Across the Life Course. *The journals of gerontology. Series B, Psychological sciences and social sciences*, *75*(3), 613–624.

II. Origin, Destination or Mobility? A Monte Carlo Simulation of the Diagonal Reference Model.

Outline

Bibliographic Information

This chapter is work in progress.

Author's contribution

I developed the methodology, writing preparation, interpretation of findings, and the data generation and analysis. Robin Samuel conceptualized, supervised, and reviewed the chapter.

Abstract

Statistical modeling on the net effects of socioeconomic origin, destination, and mobility on sociologically highly relevant topics is affected by an identification problem, which cannot be solved with traditional statistical techniques. In current empirical research, Sobel's classic diagonal reference model (DRM) has (re)emerged as the most popular statistical tool to address this problem. The appeal of DRM is twofold. First, the model is solidly built based on theoretical considerations. Second, it is easy to interpret and provides meaningful parametric weights to assess the salience of origin and destination over the outcome variable. We attempt to contribute to a better understanding of the model, using a Monte Carlo simulation. Our data generation process employs a theoretically guided approach to generate a mobility table. The design explores two different scenarios: a) when the dependent variable is continuous and b) when the dependent variable is dichotomous. A particular focus is on bias and coverage assessment of mobility estimates. Our findings suggest that the DRM does not yield substantially biased estimates under the generic scenarios studied here. However, the computation of the confidence interval is problematic. We call for further research on the model's behavior in different sets, especially in longitudinal data.

II.1 Introduction

The intuition that experiencing social mobility might influence the attitudes, behaviors, and psycho-social conditions of individuals is firmly entrenched in the social sciences. Over the last few decades, the literature has documented the debate on the potential benefits, harms, and significance of social mobility on the life chances and opportunities for individuals (see, e.g., Sorokin, 1927; Tumin, 1957; Lipset, 1959; Goldthorpe, 1980; Simandan, 2018).

Many theorists associate upward social mobility with an increase in access to resources resulting from occupying a more privileged position in society (Goldthorpe, 1980; Ormel, Lindenberg, Steverink, and Verbrugge, 1999). Others hold that the experience of social mobility can have hidden costs (Friedman, 2014). Specifically, the costs of social mobility lie in the progressive detachment from a set of values and beliefs that are characteristic of the social class into which individuals were born, and an attachment to a new set of values and beliefs embedded in the destination social class (Sorokin 1927; see Houle and Martin 2011, and Chan 2018 for recent empirical tests of Sorokin's dissociative theory). The empirical strategy employed by applied researchers sees the socio-cultural identities of individuals as inherently affected by the primary imprinting to social norms that individuals have learned in their specific class of origin, and the acquisition of values and beliefs that characterize their current class of destination, that is, the new social environment to which they adapt. Social mobility can then be conceptualized as the specific trajectory between a given class of origin and a given class of destination.

However, sociologists struggled to translate their theoretical and conceptual frameworks into statistical models that facilitate an analysis of the consequences of social mobility. This is due to the identification problem that arises when indicators of origin, destination, and mobility are simultaneously entered into a regression model.

Since the 1960s at least, scholars of social mobility have suggested tools to overcome the identification problem: for example, the square additive model by Duncan, 1966, the halfway/difference model by Hope, 1971, 1975, and the Diagonal Reference Model by Sobel, 1981, 1985; for a comparison of the three, see Hendrickx et al., 1993.

Sobel's Diagonal Reference Model (DRM) was widely applied in a first phase in the 1980s and 1990s (see, e.g., Sorenson, 1989; Weakliem, 1992; Dirk de Graaf and Heath, 1992; Clifford and Heath, 1993; Hendrickx et al., 1993; Van Berkel and de Graaf, 1995; De Graaf, Nieuwbeerta, and Heath, 1995 Marshall and Firth, 1999) and has regained interest an on-going second phase, starting in the late 2000s (Breen, 2001; Van der Slik, De Graaf, and

Gerris, 2002; Tolsma, De Graaf, and Quillian, 2009; Houle and Martin, 2011; Eeckhaut et al., 2013; Willekens, Daenekindt, and Lievens, 2014; Missinne, Daenekindt, and Bracke, 2015; Daenekindt, 2017; Jonsson et al., 2017; Van der Waal, Daenekindt, and de Koster, 2017; Billingsley, Drefahl, and Ghilagaber, 2018; Chan, 2018; Schuck and Steiber, 2018; Gugushvili, Zhao, and Bukodi, 2019; Präg and Richards, 2019). The DRM is now considered to be the first choice in statistical methodology to study the potential consequences of social mobility (e.g., Van der Waal et al., 2017). In comparison studies, the DRM has been found to be more parsimonious than other models, mostly owing to its capability to meaningfully disentangle the effects of social mobility, from the part attributable to origin and destination (Hendrickx et al., 1993; Eeckhaut et al., 2013; Van der Waal et al., 2017; Billingsley et al., 2018).

Although the model has been regaining interest, the DRM does not seem to be able to clarify the role of mobility on the many facets that compose the individual sphere, as the accumulation of null or weak evidence of mobility effects is in stark contrast to the expectations derived from theory (Lipset, 1959; Ellis and Lane, 1967; Kessin, 1971; Bean, Bonjean, and Burton, 1973; Friedman, 2014; Friedman, 2016) on consequences of social mobility (Weakliem, 1992; De Graaf et al., 1995; Breen, 2001; Daenekindt, 2017; Houle and Martin, 2011; Van der Waal et al., 2017; Schuck and Steiber, 2018; Gugushvili et al., 2019; Präg and Richards, 2019). From a methodological perspective, the null findings may suggest that the DRM is still unable to solve the identification problem. Should the DRM, in principle, allow an unbiased estimation of the effects of interest, the null and weak findings may be taken as evidence against some of the expectations derived from theory. However, there is a lack of systematic studies on model behavior to advance our understanding of the statistical characteristics of the DRM and its ability to disentangle origin, destination, and social mobility effects. Against this backdrop, it is of specific importance to evaluate, first, whether the estimates of the mobility coefficients are biased, and if so to what extent. Second, we must evaluate the overall capability of the DRM to detect effects of various sizes present in the population.

In our paper, we address these gaps and assess the potential benefits and limitations of the DRM using Monte Carlo simulation. Our experimental design is divided into two main sets: one scenario tested the DRM with a continuous dependent variable, and the other included a model with a dichotomous dependent variable. Our work has potential implications for a wide range of applied sociological research: the first field relates to the recent literature on social stratification and mobility on health inequalities (Missinne et al., 2015; Jonsson et al., 2017; Van der Waal et al., 2017; Präg and Richards, 2019; Gugushvili et al., 2019). In this type of study,

the dependent variable may be measured as continuous, such as functional somatic symptoms (Jonsson et al., 2017), depressive symptoms (Gugushvili et al., 2019), biomarkers (Präg & Richards, 2019), or dummy/dichotomous response variables, such as regular mammography screening (Missinne et al., 2015) or at-risk/not-at-risk for obesity (Van der Waal et al., 2017). A second application can be found in electoral studies, where a researcher can operationalize political behavior as a continuous scale (De Graaf & Ultee, 1990) or as a dichotomous indicator of left- or right-wing vote (Clifford & Heath, 1993). A third application is in demographic studies, where the dependent variable is continuous on fertility (Sorenson, 1989), while mortality is often operationalized as a dichotomy (e.g., survived/not survived) (Billingsley et al., 2018). A fourth example can be found in research on status inconsistency and attitudes: the applied researcher can think of a continuous dependent variable as attitudes toward ethnic minorities (Tolsma et al., 2009) or as a dichotomous indicator concerning the demand for redistribution (Jaime-Castillo & Marqués-Perales, 2019).

Our contribution is organized as follows. First, we provide a review of the roots of the identification problem and how the DRM addresses this methodological issue. Second, we show our experimental design, specifically the random data generating process (DGP), which is composed of the random occupational table generation and the steps required to simulate the dependent variables. Third, we extract the results of our study and discuss their implications at an empirical level. We conclude with suggestions for furthering the assessment and development of the DRM.

II.2 The Roots of the Identification Problem

Let us assume that we are interested in evaluating and disentangling the effects of the social class of origin and destination, as well as the specific effect of mobility between these classes on a given outcome, such as self-rated health. The researcher may then think that the most intuitive approach to empirically assess the effects of social origin (O), destination (D), and mobility (M) is to incorporate these measurements into a linear model (Blalock 1966, 1967; Duncan 1966; Mason et al. 1973) of the form:

$$Y = \mu + \alpha_i O + \beta_j D + \gamma_k M + \epsilon_{ij} \tag{II.1}$$

where μ is the grand mean of the dependent variable Y, α_i is the effect of the i^{th} origin class, β_j is the effect of being currently in the j^{th} destination class, and γ_k is the effect of mobility. To avoid the overparameterization of the model,

the parameters of the regression equation can be constrained to $\sum_{i=1}^{I} \alpha = \sum_{j=1}^{J} \beta = \sum_{k=1}^{K} \gamma = 0$. Even if we can deal with overparameterization, it is evident that social mobility is a linear combination of origin and destination as $M = D - O$ in this case. Linear dependency can be considered as perfect collinearity between explanatory covariates. This type of linear transformation is familiar to many social scientists, as it relates to a similar problem that occurs when trying to isolate age, period, and cohort effects (see, e.g., Yang, Fu, and Land, 2004; for a more technical discussion, see Carstensen, 2007 and Kuang, Nielsen, and Nielsen, 2008). The use of conventional methods, such as the generalized linear regression, will typically fail to yield a set of coefficients that will uniquely identify the effects α_i, β_j, and γ_k.

To see this, we can rewrite the formula in matrix form as $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b}$, where the dependent variable is $\boldsymbol{y}_{n\times 1} = (y_1, y_2, y_3, \dots y_n)^T$, the regressor matrix is $\boldsymbol{X}_{n\times p}$, and the matrix of coefficients is $\boldsymbol{b}_{p\times 1} = (\alpha, \beta, \gamma)^T$. Recalling the Gauss-Markov theorem, the best linear unbiased estimator (BLUE) in ordinary least squares (OLS) is $\hat{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$. However, owing to the linear dependency between O, D, and M, the rank (the number of linearly independent rows/columns) of $\boldsymbol{X}_{n\times p}$ is less than p (see Searle and Khuri 2017). If $\boldsymbol{X}$ is rank deficient, we cannot find the inverse of the square matrix:

$$\boldsymbol{X}^T\boldsymbol{X} = \begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix}$$

In this case, the square matrix $\boldsymbol{X}^T\boldsymbol{X}$ is said to be singular or rank deficient (less than full column rank). Consequently, the matrix of regressors can have infinite possible solutions (Yang et al., 2004). In the next section, we will demonstrate how the DRM addresses this problem.

II.3 The Diagonal Reference Model

The DRM is a theoretically founded model (Cox 1990; Yamaguchi 2002) in that it draws on acculturation process theory, which states that individuals tend to adopt the behaviors, beliefs, and values typical of the reference aggregate when a status inconsistency occurs. In terms of social mobility, individuals tend to conform to the social behavior typical of the class of destination: that is, the referent aggregate. The difference between the equation II.1 and the DRM lies in the decomposition of y_{ijk}. In the former, y_{ijk} can be decomposed into two additive effects, α_i for origin and β_j for the destination. The DRM decomposes y_{ijk} in μ_{ii}, the population means specific to the i^{th} origin category, and μ_{jj}, the population means specific to the j^{th}

destination category. For immobile individuals, in the diagonal of the mobility table, the equality $\mu_{ii} = \mu_{jj}$ holds. To construct a meaningful mobility model, it is essential to quantify the acculturation process. This is achieved by the choice of the referent group to which status inconsistents are assigned in a mobility table $i * j$, in which row i represents the origin status and column j represents their destination classes. This can be defined as:

$$\hat{\mu}_{ij} = \rho\mu_{ii} + (1-\rho)\mu_{jj} + \sum_{w=1}^{W} \gamma_w M_{ijw} + \epsilon_{ijk} \tag{II.2}$$

$$\rho = \frac{e^{\delta_i}}{e^{\delta_i} + e^{\delta_j}} \tag{II.2a}$$

$$(1-\rho) = \frac{e^{\delta_j}}{e^{\delta_i} + e^{\delta_j}} \tag{II.2b}$$

The DRM is said to be a parametrically weighted regression model (Yamaguchi, 2002) as the means μ_{ii} and μ_{jj} are weighted by ρ and $(1-\rho)$. These quantify the relative salience of origin, and the destination on off-diagonal cells mean values μ_{ij}. In model II.2, the referent values are taken as the population means, μ_{ii} and μ_{jj} , and estimated along the diagonal of the mobility table. For people who are off the diagonal, there are two referent values: the first represents the value found in the i^{th} origin category μ_{ii}; the second is the value of the j^{th} destination class, μ_{jj}. The last block of equation II.1 refers to the inclusion of mobility covariates $w = 1, 2, 3, \ldots W$ that take the value M_{ijw}, and have effects quantified in γ_w. The mobility effects related to the k^{th} observation are not indexed, as they are constant across individuals, clustering them over their origin or destination class. The parameter ϵ_{ijk} represents the stochastic error term and follows $\epsilon_{ijk} \sim \mathcal{N}(0, \sigma^2)$.

As mentioned before, the weight parameters ρ in equations II.2a and $(1-\rho)$ in equation II.2b quantify the relative salience of origin to the current destination. The parameters are computed as the ratio of the weighted effects of origin δ_i and destination δ_j (to be estimated by the model) to the off-diagonal means and ranges in the interval $0 < \rho < 1$.

When $\rho = 0.5$, the origin and destination have the same weight on the dependent variable. If $\rho > 0.5$, then the effects of the i^{th} class of origin, for example, owing to socialization, are of greater importance than those of the j^{th} class of destination, for example, owing to the set of values dominant there. If $\rho < 0.5$, the weighted parameter should be interpreted in the opposite direction. The equality constraint ensures that the weights sum up to 1 and enable a meaningful interpretation of the parameters.

As Sobel, 1981 noted, the DRM uses non-linear least square estimation (which is equivalent to maximum likelihood estimation; see the following section for further details) because the model includes multiplicative effects, specifically for the product of the weight parameters and the off-diagonal averages. Table II.1 shows the parameterization applied by the DRM in mobility contingency tables using an example where stratification variables are cross-classified as the origin of individuals in the rows and their destination in the columns, in four possible classes.[9]

Table II.1: Cells-Generating Mechanism of the DRM when there are Four Classes of Origin and Destination (I—IV)

	Destination				
Origin	I	II	III	IV	all
I	μ_1	$\rho\mu_1 + r\mu_2$	$\rho\mu_1 + r\mu_3$	$\rho\mu_1 + r\mu_4$	$\rho\mu_1$
II	$\rho\mu_2 + r\mu_1$	μ_2	$\rho\mu_2 + r\mu_3$	$\rho\mu_2 + r\mu_4$	$\rho\mu_2$
III	$\rho\mu_3 + r\mu_1$	$\rho\mu_3 + r\mu_2$	μ_3	$\rho\mu_3 + r\mu_4$	$\rho\mu_3$
IV	$\rho\mu_4 + r\mu_1$	$\rho\mu_4 + r\mu_2$	$\rho\mu_4 + r\mu_3$	μ_4	$\rho\mu_4$
all	$r\mu_1$	$r\mu_2$	$r\mu_3$	$r\mu_4$	μ

$r = 1 - \rho$

The cells shaded in gray along the diagonal of Table II.1 contain the immobile individuals, who serve as the referent groups. In the off-diagonal cells of Table II.1, the decomposition of the values of the dependent variable is shown: for the i^{th} origin category, the rows, and for the j^{th} destination category, the columns. The cells shaded in gray along the diagonal of the contingency table are the intercepts, or main effects, while the off-diagonal cell values are a weighted average (by the salience parameters ρ for the origin and r for the destination) of the specific i^{ith} row and j^{th} column main effects[10].

Maximum Likelihood and Nonlinear Least Squares Estimations

Following Sobel, 1981, the maximum likelihood estimation and the nonlinear least square estimation of the parameters in the DRM model yielded equivalent results. This section retakes the steps shown in Sobel, 1981 that demonstrate this equivalence to present how the DRM derives the parameters of interest, and to provide the computational methodology implemented in statistical software such as R —with the `gnm` package —or in Stata —with the `Diagref` package.

Recall the formulation of the baseline DRM, in equation II.1, with the mobility variables included. Following Sobel, 1981, equation II.1 can be converted as a function of μ_1 and μ_2, which are the values of the dependent variable along the diagonal of the squared table of means:

$$\mu_1 = \sum_{\mathrm{i}=1}^{\mathrm{I}} \mu_{\mathrm{ii}} \mathrm{X}_{\mathrm{.i}} \qquad\qquad \mu_2 = \sum_{\mathrm{j}=1}^{\mathrm{J}} \mu_{\mathrm{jj}} \mathrm{X}_{\mathrm{.j}} \tag{II.3}$$

where $x_{.i}$ and $x_{.j}$ are a set of dummy variables for each factor of origin and destination, where $X = 1$ if $i = i$, and 0 otherwise. Substituting equation II.3 into II.1, we obtain

$$Y_{ijk} = \sum_{\mathrm{i}=1}^{\mathrm{I}} \rho \mu_{\mathrm{ii}} \mathrm{X}_{\mathrm{.i}} + (1-\rho) \sum_{\mathrm{j}=1}^{\mathrm{J}} \mu_{\mathrm{jj}} \mathrm{X}_{\mathrm{.j}} + \sum_{w=1}^{W} \gamma_w M_{ijw} \tag{II.4}$$

Equation II.4 can be rewritten using matrix notation:

$$y = (\mu_1, \mu_2, X) \begin{pmatrix} \rho \\ 1-\rho \\ \gamma \end{pmatrix} \implies (y - \mu_2) = (\mu_1 - \mu_2, X) \begin{pmatrix} \rho \\ \gamma \end{pmatrix} \tag{II.5}$$

where y, μ_1, and μ_2 are the column vectors of the dependent variable, and the diagonal values (indicated by the vector X composed of 0s and 1s) of origin and destination, respectively. Sobel, 1981 specifies that, as the vectors μ_1 and μ_2 are unobserved, using the means $\bar{y}_{ii}$ and $\bar{y}_{jj}$ as proxies for μ_1 and μ_2 would yield biased but consistent parameter estimates. The alternative approach proposed by Sobel, 1981 addresses the likelihood function under random sampling of the form of:

Maximum Likelihood Estimation:

$$\mathcal{L}_{(\mu,\rho,\gamma,\sigma)} = \prod_{ijk} \Bigg((2\pi\sigma^2)^{-\frac{1}{2}} \exp\Bigg\{ -(2\sigma^2)^{-1} \times, \tag{II.6}$$

$$\left(y_{ijk} - \sum_{i=1}^{I} \rho \mu_{\mathrm{ii}} \mathrm{X}_{.i} - \sum_{j=1}^{J} (1-\rho) \mu_{\mathrm{jj}} \mathrm{X}_{.j} - \sum_{w=1}^{W} \gamma_w \mathrm{X}_{ijw} \right)^2 \Bigg\} \Bigg)$$

Nonlinear Least Squares:

$$\underset{f(\rho,\mu,\gamma,\sigma)}{\arg\min} = \sum_{ijk} \left(y_{ijk} - \sum_{i=1}^{I} \rho\mu_{ii} X_{i.} - \sum_{j=1}^{J} (1-\rho)\mu_{jj} X_{.j} - \sum_{w=1}^{W} \gamma_w \mathrm{X}_{ijw} \right)^2 \tag{II.7}$$

The maximization of the likelihood function in equation II.6 with respect to the parameters ρ, μ_{ii}, γ_w, and σ^2 are estimated to be equal to the choice of these parameters to minimize the nonlinear least square parameters. This condition ensures, following Sobel's specification, that the estimates are unbiased, consistent, and efficient.

II.4 Simulation Design

Following the above steps, our Monte Carlo simulation will have to consider two main elements: a) the random generation of an occupational mobility table of the form IxJ, in which rows I indicate the origin classes and columns J the destination classes[11]; b) generation of the dependent variables. The total sample sizes are N = {500, 750, 1000}. The performance measures we are interested in are the bias of the estimators and their coverage. We performed 2,000 repetitions to reduce the Monte Carlo standard error of these performance measures, relying on Burton et al. (2006) and Morris, White, and Crowther (2019), using the formula: $N_{sim} = \left(\frac{Z_{1-(\alpha/2)}\sigma}{\delta} \right)^2$, where N_{sim} represents the total number of simulations, δ is the specified level of accuracy (in empirical studies, it is usually set at 95%), σ^2 is the variance of the parameter to be simulated, and $Z_{1-(\alpha/2)}$ is the $1-(\alpha/2)$ quantile in a standard normal distribution. To perform the Monte Carlo simulation, we used R and the packages `rTableICC` (Demirhan, 2016) for the generation of the mobility table and the `gnm` package for the computation of the DRM estimates (Turner & Firth, 2015). The organization of the Monte Carlo coding structure relied on the `SimDesign` package (Sigal & Chalmers, 2016).

The Data Generation Process

Generation of a random social mobility table

The Monte Carlo simulation relies on a square contingency table in which $R, C \in \{1, 2, 3, 4\}$ are the levels of the categorical responses. In the context of social stratification research, one can consider the categorical responses as, for example, the contracted Goldthorpe's 4-classes schema (Erikson & Goldthorpe, 1992). The random generation process of the counts within each cell relied on sampling from the multinomial distribution[12] $X \sim Mult(n, \pi)$, where X is

the random variable, n the number of trials (in our case, observations), and π the within-cell probability. The crucial step, thus, was to create a matrix of probabilities that would guide the counts, given π and n. To allow for reasonably realistic and sociologically relevant scenarios, we draw on work by Erikson and Goldthorpe, 1992. In particular, we used the topological model as a reference occupational table *hierarchy 1* (HI1). We chose four classes for both the origin and destination categories. The first social class, the white collars, comprises of classes I and II (higher technical, professional, managerial, and administrative occupations) that are mainly in a service relationship. The second social class, the intermediate class, comprises routine non-manual workers (IIIa, IIIb) and lower technical occupations (V). The third social class, the manual class, comprises skilled and unskilled manual workers (class VI and VIIa) and unskilled manual workers in agriculture (class VIIb). The bourgeoisie comprises the fourth social class in our simulation; these are large and small employers (classes I, IVa, IVc) as well as the self-employed workers, classes IVb and IVc (for further details, see Erikson and Goldthorpe, 1992).

Table II.2 shows the cell probabilities used to construct the counts within the cells of the contingency table:

Table II.2: Square Matrix of Probabilities Used to Generate the Contingency Table

	I	II	III	IV
I	0.08	0.02	0.01	0.01
II	0.02	0.26	0.04	0.01
III	0.01	0.04	0.35	0.04
IV	0.01	0.01	0.04	0.05

The joint multinomial sampling of count distribution within each cell has been modeled to establish the probabilities π_{rc}. The sum of all probabilities in Table II.2 has the constraint to sum up to 1 as $\sum_{r=1}^{r} \pi_{rc} = \sum_{c=1}^{c} \pi_{rc} = 1$. That is, the probabilities stated in Table II.2 determine the number of individuals falling within a particular cell, given the total number of cases.

The Dependent Variables

We constructed two dependent variables. Reflecting typical applications of the DRM, the first outcome variable was modeled as continuous, and the second outcome variable was dichotomous. Substantive examples for the former could, for example, include a specific biomarker (see Harris and Schorpp 2018 for a general discussion and Präg and Richards 2019 for a practical

Table II.3: Population True Values for the Data Generating Process

ρ	$(1-\rho)$	γ_{Up}	γ_{Down}
0.70	0.30	$\{-0.1, -0.5\}$	$\{0.1, 0.5\}$
0.50	0.50	$\{-0.1, -0.5\}$	$\{0.1, 0.5\}$
0.30	0.70	$\{-0.1, -0.5\}$	$\{0.1, 0.5\}$

application), a political preferences scale, or a life satisfaction scale. Examples of the latter could be voting vs. non-voting for a determined political party or agreement vs. disagreement with a determined attitude or value. For our purposes, we used an 11-point scale of self-rated health and a dichotomous mortality variable (survived vs. not survived). [13] The variables were modeled as follows:

Continuous Dependent Variable

$$y_{ijk} = \rho\mu_{ii} + r\mu_{jj} + \gamma_{up}Upward + \gamma_{down}Downward \tag{II.8}$$

Dichotomous Dependent Variable

$$\pi_{ij} = \frac{exp(\rho\mu_{ii} + r\mu_{jj} + \gamma_{up}Upward + \gamma_{down}Downward)}{1 + exp(\rho\mu_{ii} + r\mu_{jj} + \gamma_{up}Upward + \gamma_{down}Downward)} \tag{II.9}$$

The parameters of interest are γ_{up} and γ_{down}. To keep the experimental design manageable, we set the mobility variables as contrasts between immobiles vs. upwardly mobiles and immobiles vs. downwardly mobiles, where mobile individuals are coded as 1, and otherwise as 0. The interpretation of the coefficients is the same as in ordinary least square (OLS) regression. The diagonal means for the continuous dependent variable are $\mu_{ii} \in \{2, 4, 6, 8\}$. As a practical example, one can imagine that the diagonal means are the different levels of self-rated health, where we observe the population means of a 11-point scale, 0 for excellent health and 10 for very poor health, so as the means are lower for the higher class and higher for the lower class.

The diagonal ratios we set for the logistic dependent variable are $\pi_{ii} \in \{-2, -1.5, .5, 1\}$. One might think of, as a practical example, the risk of surviving or not surviving. In this example, the diagonal ratios observed are the probabilities of not surviving, which are then lower for the higher classes and higher for the lower classes. Table II.3 summarizes the population values we have set to fix the parameters ρ, r, γ_{Up} and γ_{Down}:

We set negative effects for people experiencing upward mobility, one moderate (-0.1) and one strong (-0.5). Conversely, experiencing downward

mobility has negative effects with the same magnitudes but reversed sign-positive effects on the simulated outcome. One might think, recalling the preceding example, that experiencing upward social mobility may reduce the risk of death, while experiencing downward mobility may increase it. Concerning the class weights, we set the weighting parameters in the first row of Table II.3 such that the origin of individuals has greater salience than the destination class. The second row assumes that both the origin and destination have equal salience. Lastly, the third row attributes greater importance to the destination class than to the origin class. To keep the experiment's design focused on the mobility coefficients, we did not allow the weighting parameters to vary across the categories of origin and destination.

We reported the standardized bias computed as $\left(\frac{\bar{\hat{\beta}}-\beta}{SE(\hat{\beta})}\right)$, where $\bar{\hat{\beta}}$ is the average value of the simulated estimates and β is the true population value, and the empirical coverage rate (ECR), computed as $\frac{1}{N_{sim}}\sum_{i=1}^{N_{sim}} 1(\hat{\beta}_{low} \leq \beta \leq \hat{\beta}_{up})$, which is the average number of times the true population values are within the lower and upper boundaries of the estimated confidence interval (CI). Following Burton et al., 2006 and Collins et al., 2001, as a rule of thumb, a cut-off point of .40 was assigned to consider an estimate severely biased, as "it has been shown to have a noticeable impact on the efficiency, coverage, and error rate" (Burton et al. 2006, :4287). The ECR indicates the proportion the $100(1-\alpha)$ CIs contain the true population values, where the α chosen is the canonical 0.05. The optimal value of ECR is 0.95; if it is lower, the confidence intervals are said to be too permissive and the model will incorrectly detect a significant result, leading to higher type I errors: that is, the rejection of a true null hypothesis. If the ECR is higher than 0.95, the CIs are said to be conservative, leading to a loss of statistical power and higher chances of type II errors: that is, the non-rejection of a false null hypothesis. Following Burton et al., 2006, to define the boundaries within which the ECR can be considered acceptable, the estimate should not fall outside two standard errors (SEs) of the nominal coverage probability $p = 0.95$. The SE is calculated as $SE(p) = \sqrt{p(1-p)/N_{sim}}$, where p is defined as the nominal coverage probability (i. e., $p = 0.95$) and N_{sim} represents the total number of simulations performed. In our case, $SE(p) * 2$ is equal to 9.747×10^{-3}. Adding and subtracting this quantity to and from .95 gives us (rounded) lower and upper boundaries of 0.940 and 0.960, respectively. Tables in the Appendix A and Appendix B show the standardized bias (γ_{Up} and γ_{Down}) and ECR ($100(1-\alpha)Up$ and $100(1-\alpha)Down$) summary measures of the 2,000 repetitions for each scenario.

Continuous Dependent Variable

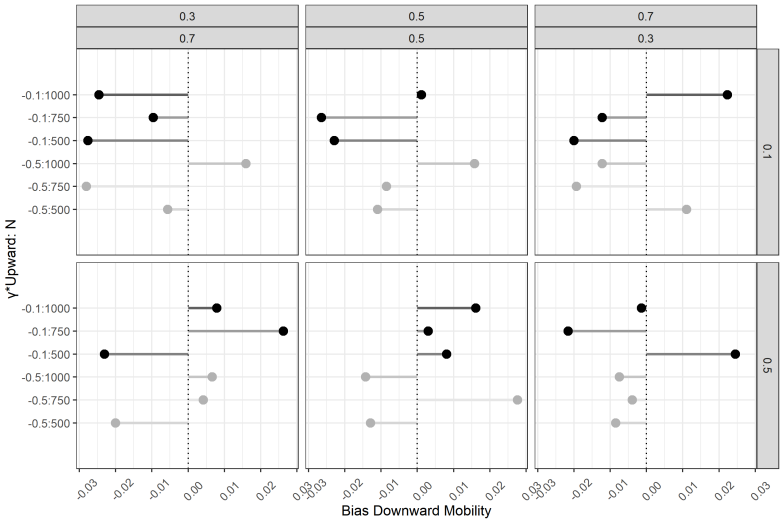
Considering the continuous dependent variable scenario, we focused our analysis of the performance measure on the standardized bias and the ECR of the upward and downward mobility coefficients. The first aspect to be highlighted is that there is no substantive bias δ affecting the parameters under test. For what concerns the capability of the model to detect statistically significant effects at the 95% level, the findings suggest that the DRM is affected by under- and over-coverage in specific cases. Appendix A contains a table with a complete summary of the performance measures.

Bias

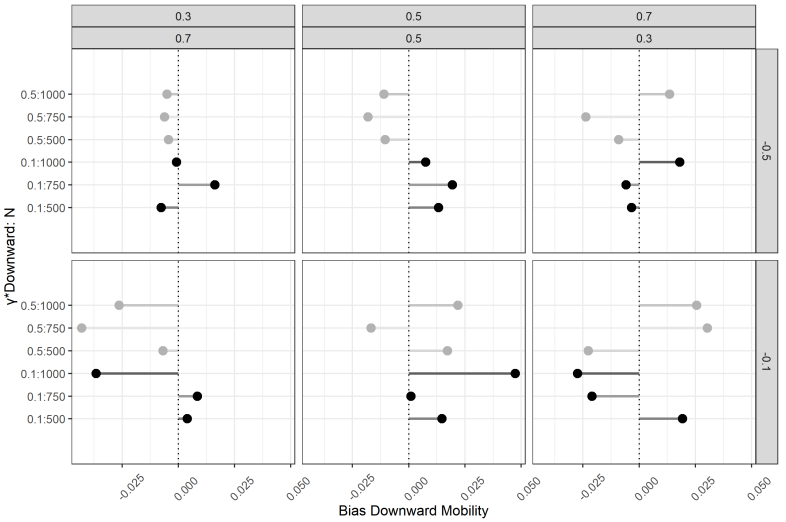
Figure II.1 provides a visual representation of the bias of the estimates (full results can be found in Appendix A). Panel II.1a shows the standardized bias affecting the indicator of upward mobility. Panel II.1b depicts the standardized bias affecting the dummy variable for downward mobility, where the subgraphs are generated by combining the salience parameters that sum up to 1 for the columns, and the true population values of the indicator for upward mobility for the rows. The subgraphs are generated according to the initial conditions of the experimental design: the columns are the true weighting parameters combined, to sum up to 1, and the rows are the values of the downward mobility indicator. The vertical axis shows the combinations of the true upward mobility parameter value set and the total sample size used for the DGP; the horizontal axis shows the degree of standardized bias.

Figure II.1: Upward and Downward Mobility Bias Lollipop Plot for Linear Scenario

(a) Bias Upward Mobility



(b) Bias Downward Mobility



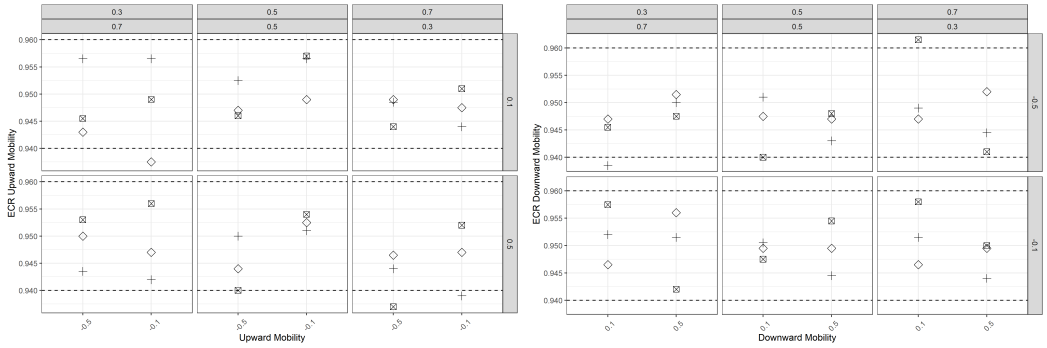
For the linear dependent variable scenario, panels II.1a and II.1b show that the bias affecting the mobility estimates is well below the .40 threshold. Considering the bias direction of the mobility indicators, panels II.1a and II.1b show a tendency of the bias to move in opposite directions. For instance, in the upper- and bottom-left and in the upper- and bottom-center subgraphs of panels II.1a and II.1b, we see that the direction of the bias is negative for the upward mobility indicator and vice versa.

Empirical Coverage Rates

Focusing on the capability of the model to detect the true population effect, figures II.2a and II.2b depict the empirical coverage rates (ECR), according to each true population value. The two panels are divided as before: the columns are formed by the salience parameters (summing up to 1), and the rows represent the true population values concerning the upward or downward mobility variables. The dashed lines indicate the minimum and maximum boundaries within which the ECR is considered acceptable.

Figure II.2: Upward and Downward Mobility ECR Plot for Linear Scenario

(a) ECR Upward Mobility (b) ECR Downward Mobility



Panels II.2a and II.2b show that almost all of the ECR computed are within the lower and upper boundaries represented by the red dashed lines. However, it is worth noting that in panel II.2a, when the salience of origin is weaker and the mobility coefficients are weak (i.e., -0.1) for both upward and downward mobility variables (upper-left subgraph of panel II.2b), the DRM underestimated the true population value within the CI when the sample size was set to 750 and 500 individuals. This means that, as the computed ECR is lower than 0.95, the DRM yields CI that are too permissive when both the true upward and downward mobility effects are weak, thereby losing statistical power and increasing chances of type I errors. The same result

can be found in the bottom-right subgraph of panel II.2a, in the case of $N = 1,000$ and $\gamma_{Up} = -0.5$, and N = 500 and $\gamma_{Up} = -0.1$.

Logistic Dependent Variable

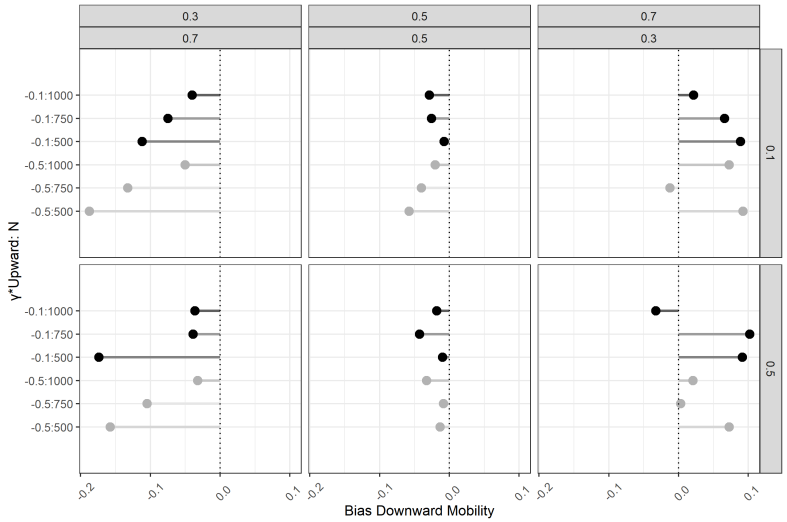
As with the continuous dependent variable scenario, we focused our analysis of the performance measure on the standardized bias and the ECR of the upward and downward mobility coefficients. Appendix B contains a complete overview of all results. We provide a visual representation of the results in the next section.

Bias

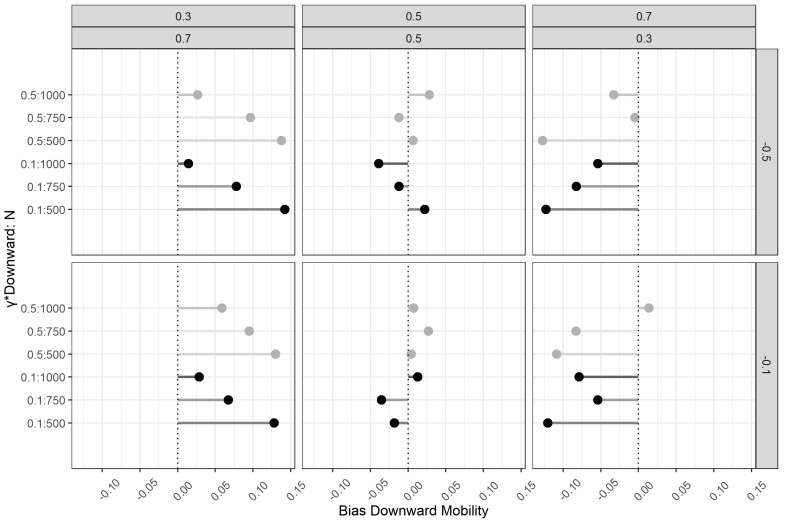
Figure II.3 shows that, despite the increase in magnitude compared with the continuous dependent variable scenario, the bias estimation should not affect the efficiency, coverage, and error rate as it is below the 0.40 threshold.

Figure II.3: Upward and Downward Mobility Bias Lollipop Plots for Logistic Scenario

(a) Bias Upward Mobility



(b) Bias Downward Mobility



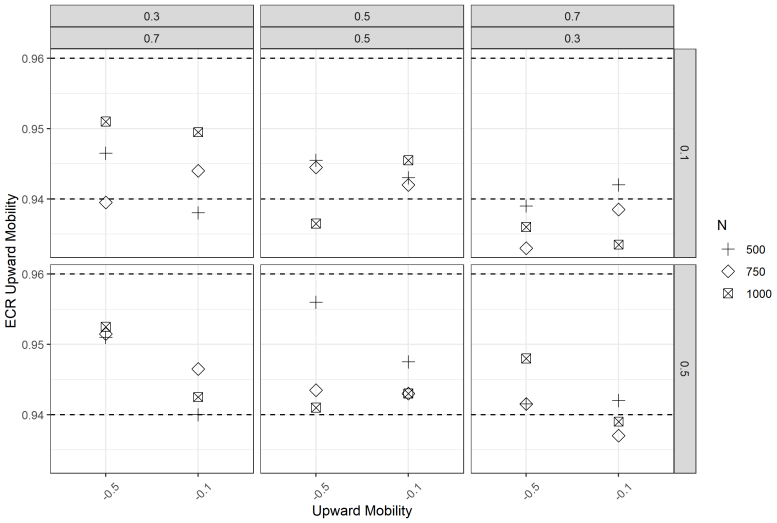
As in the linear dependent variable scenario, panels II.3a and II.3b show opposite directions of bias when we consider upward and downward mobility variables concurrently. The only exception can be found in both columns where the salience parameters are of equal importance: that is, when both are set to 0.5.

Empirical Coverage Rate

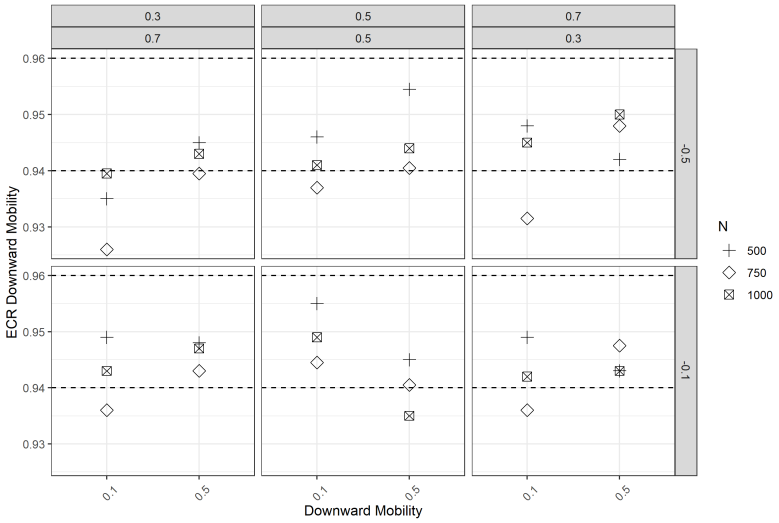
We now consider the capability of the DRM to capture the true population effect when the dependent variable is dichotomous. Figure II.4 depicts the ECR computed for each true parameter. Panel II.4a shows the ECR for the parameters set to model the effect of upward mobility, while II.4b shows the ECR for the parameters related to the downward mobility variable included in the true model.

Figure II.4: Upward and Downward Mobility ECR Plot for Logistic Scenario

(a) ECR Upward Mobility



(b) ECR Downward Mobility



Panel II.4a shows under-coverage for both weak and strong upward mobility effects, especially when the concomitant downward mobility coefficient is weak (upper row of panel II.4a). This means that the DRM computes CI that is too permissive, leading to higher chances of type I errors (rejection of true null hypothesis). Panel II.4a shows further that when the weighting parameters are of equal salience, the DRM correctly computes the confidence intervals (center column of panel II.4a), although the degree is less severe. When the origin of individuals has greater salience (upper and lower right subgraphs), the ECR computed are under the lower boundary, especially when the concomitant downward mobility effect is weak.

Panel II.4b shows that when the weighting parameter is weaker for social origin, we found severe under-coverage, especially when the upward mobility effect was strong and the downward mobility effect was weak. Surprisingly, panel II.4b shows under-coverage when the sample size becomes larger (specifically, N = 750 and N = 1,000) when the upward mobility coefficient is weaker.

II.5 Discussion & Conclusions

In empirical research on the effects of social mobility, social scientists strive to design a statistical model capable of disentangling the effects of origin, destination, and mobility. The main methodological challenge has been the identification problem: that is, the linear dependency of social origin, destination, and mobility. In recent literature, applied researchers rely increasingly on Sobel's Diagonal Reference Model, owing to its desirable property of distinguishing the origin, destination, and mobility parsimoniously and with interpretable parameters. However, the collection of weak or null findings calls into question the supposed capability of the DRM to overcome the identification problem. In this paper, we have tried to obtain a deeper understanding of the model through Monte Carlo simulation, testing how the DRM behaves when dealing with the identification problem under common conditions. The experimental design considered two types of scenarios: the first tests the behavior of the DRM assuming the dependent variable is continuous, while the second assumes a dependent variable that follows an inverse-logistic function: that is, generating a dichotomous dependent variable. We also tested the model for different sample sizes, specifically at 500, 750, and 1,000 individuals. Our measures of interest were the standardized bias, to see how far the computed estimates departed from the population value, and the empirical coverage rate, to assess how many times the computed confidence intervals include the true parameter value.

Overall, our results show that the DRM is not affected by severe bias, although the standardized bias is greater: that is, the estimates depart further from the true population value in the logistic scenario, but still at acceptable levels. In this sense, we expected a degree of bias in the mobility estimates, as we constructed the differences among categories in the main diagonal using a determined set of means. This is in accordance with Sobel, 1981, who warned that estimates obtained by the DRM would be biased but consistent as the distribution of our computed estimates confirm (see Appendix C).

Although the magnitude of the standardized bias is not at a critical level to affect efficiency, coverage, and error rate, our findings show that the DRM tends to underestimate the effects of mobility. Specifically, as the relative charts show (see Figures B.1, B.2, B.3 and B.4), the direction of the standardized bias seems to go in opposite directions when compared to the bias affecting upward and downward mobility indicators simultaneously. This behavior of the DRM might partly explain why many empirical studies have failed to detect mobility effects, even if they were to be expected by theory. For applied researchers, this would translate to a bias toward zero that hides the true effects of mobility in the outcome variables. For what concerns the capability of the model to correctly capture the true population value, our findings suggest that in both the linear and logistic scenarios the computed ECR are affected by under- or over-coverage. The DRM failed to detect the true population value correctly quite evidently in the logistic scenario, and in specific cases when we set the dependent variable as continuous. Applied researchers might hence want to exercise special care when the dependent variable is dichotomous.

Our study has some limitations. First, as this paper relies on an experimental design, the sets try to simplify the overarching complexity of the real world. Although our random occupational mobility table was theory-driven and we constructed the scenarios according to common types of data empirical researchers usually encounter, the generalization of the results may be limited in some cases. Second, our Monte Carlo simulation did not introduce any source of "disturbance" into the model. This is for two reasons: we aimed to assess the behavior of the DRM in general, and its capability to detect mobility effects in particular. That is to say, we wanted to test whether the DRM is capable of simultaneously detecting origin, destination, and mobility effects when they are present in the population, without other potential confounding factors. The other reason focuses on the secondary aim of this contribution: to provide input for further developments concerning the identification problem in social stratification and mobility research, particularly on the DRM. The limitations of this study may serve to encourage future contributions on the DRM. More specifically, future research should attempt

to include a full-range mobility variable, for example, a variable that takes into account the full range of social mobility trajectories to build up the population model. To reduce the discrepancy between simulated datasets and real-world data, the inclusion of unobserved heterogeneity would result in a more realistic dataset. Another aspect that we did not cover in this study is the behavior of the DRM to other types of non-normal distributions. We think this is of particular interest in light of the findings related to the dichotomous dependent variable. To enable causal inference in further applications, we suggest that future research should address the implementation of the DRM to simulated longitudinal data. In addition, recent research has started to apply the DRM to panel data (Billingsley et al., 2018), specifically comparing the efficiency of the DRM with traditional statistical models. The last point is of particular importance, as current research has applied the DRM mainly to cross-sectional data, with a few exceptions. In conclusion, we hope that our contribution will help advance the study of the consequences of social mobility. Our results show that the DRM is a helpful tool, but no silver bullet. It is best used with careful consideration of the particular application context.

Notes

[9]See Hendrickx et al., 1993 for a comparison between the Square Additive model, the halfway/difference model, and the DRM.

[10]See also De Graaf and Ultee, 1990 and Van der Waal et al., 2017 for an explanation of the parameterization of the DRM in the occupational mobility table.

[11]We follow the consolidated notation as introduced by Goodman, 1979.

[12]The random generation of the counts within each cell is $p(y) = \frac{n!}{y!(n-y)!}\pi y(1-\pi)^{n-y}$ (Agresti, 2018)

[13]However, this choice is arbitrary, and readers are free to think of alternative outcomes more related to their areas of interest.

Bibliography

Agresti, A. (2018). *An introduction to categorical data analysis.* John Wiley & Sons.

Bean, F. D., Bonjean, C. M., & Burton, M. G. (1973). Intergenerational occupational mobility and alienation. *Social Forces*, *52*(1), 62–73.

Billingsley, S., Drefahl, S., & Ghilagaber, G. (2018). An application of diagonal reference models and time-varying covariates in social mobility research on mortality and fertility. *Social Science Research*, *75*, 73–82.

Blalock, H. M. (1966). The Identification Problem and Theory Building: The Case of Status Inconsistency. *American Sociological Review*, *31*(1), 52–61.

Blalock, H. M. (1967). Status inconsistency, social mobility, status integration and structural effects. *American Sociological Review*, 790–801.

Breen, R. (2001). Social mobility and constitutional and political preferences in Northern Ireland. *British Journal of Sociology*, *52*(4), 621–645.

Burton, A., Altman, D. G., Royston, P., & Holder, R. L. (2006). The design of simulation studies in medical statistics. *Statistics in medicine*, *25*(August), 4279–4292.

Carstensen, B. (2007). Age–period–cohort models for the Lexis diagram. *Statistics in Medicine*, *26*.

Chan, T. W. (2018). Social mobility and the well-being of individuals. *The British Journal of Sociology*, *69*(1), 183–206.

Clifford, P., & Heath, A. F. (1993). The Political Consequences of Social Mobility. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *156*(1), 51–61.

Collins, L. M., Schafer, J. L., & Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological methods*, *6*(4), 330.

Cox, D. R. (1990). Role of Models in Statistical Analysis. *Statistical Science*, *5*(2), 169–174.

Daenekindt, S. (2017). The Experience of Social Mobility: Social Isolation, Utilitarian Individualism, and Social Disorientation. *Social Indicators Research*, *133*(1), 15–30.

De Graaf, N. D., Nieuwbeerta, P., & Heath, A. (1995). Class Mobility and Political Preferences: Individual and Contextual Effects. *American Journal of Sociology*, *100*(4), 997–1027.

De Graaf, N. D., & Ultee, W. (1990). Individual preferences, social mobility and electoral outcomes. *Electoral Studies*, *9*(2), 109–132.

Demirhan, H. (2016). rTableICC: An R Package for Random Generation of 22K and RC Contingency Tables. *The R Journal*, *8*(1), 48–63.

Dirk de Graaf, N., & Heath, A. (1992). Husbands' and Wives' Voting Behaviour in Britain: Class-dependent Mutual Influence of Spouses. *Acta Sociologica*, *35*(4), 311–322.

Duncan, O. D. (1966). Methodological issues in the analysis of social mobility. In *Social structure and mobility in economic development* (pp. 51–97). Chicago, Aldine Publishing Company.

Eeckhaut, M. C., Van De Putte, B., Gerris, J. R., & Vermulst, A. A. (2013). Analysing the effect of educational differences between partners: A methodological/theoretical comparison. *European Sociological Review*, *29*(1), 60–73.

Ellis, R. A., & Lane, W. C. (1967). Social Mobility and Social Isolation: A Test of Sorokin's Dissociative Hypothesis. *American Sociological Review*, *32*(2), 237.

Erikson, R., & Goldthorpe, J. H. (1992). *The constant flux: A study of class mobility in industrial societies*. Oxford University Press, USA.

Friedman, S. (2014). The Price of the Ticket: Rethinking the Experience of Social Mobility. *Sociology*, *48*(2), 352–368.

Friedman, S. (2016). Habitus clivé and the emotional imprint of social mobility. *Sociological Review*, *64*(1), 129–147.

Goldthorpe, J. (1980). *Social mobility and class structure in modern britain*. Oxford: Clarendon Press.

Goodman, L. A. (1979). Simple models for the analysis of association in cross-classifications having ordered categories. *Journal of the American Statistical Association*, *74*(367), 537–552.

Gugushvili, A., Zhao, Y., & Bukodi, E. (2019). 'Falling from grace' and 'rising from rags': Intergenerational educational mobility and depressive symptoms. *Social Science and Medicine*, *222*, 294–304.

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

Hendrickx, J., De Graaf, N. D., Lammers, J., & Ultee, W. (1993). Models for status inconsistency and mobility: a comparison of the approaches by Hope and Sobel with the mainstream square additive model. *Quality & Quantity*, *27*, 335–352.

Hope, K. (1971). Social Mobility and Fertility. *American Sociological Review*, *36*(6), 1019–1032.

Hope, K. (1975). Models of Status Inconsistency and Social Mobility Effects. *American Sociological Review*, *40*(3), 322–343.

Houle, J. N., & Martin, M. A. (2011). Does intergenerational mobility shape psychological distress? Sorokin revisited. *Research in Social Stratification and Mobility*, *1*(29(2)), 193–203.

Jaime-Castillo, A. M., & Marqués-Perales, I. (2019). Social mobility and demand for redistribution in Europe: a comparative analysis. *British Journal of Sociology*, *70*(1), 138–165.

Jonsson, F., Sebastian, M. S., Hammarström, A., & Gustafsson, P. E. (2017). Intragenerational social mobility and functional somatic symptoms in a northern Swedish context: analyses of diagonal reference models. *International Journal for Equity in Health*, *16*(1), 1–10.

Kessin, K. (1971). Social and psychological consequences of intergenerational occupational mobility. *American Journal of Sociology*, *77*(1), 1–18.

Kuang, D., Nielsen, B., & Nielsen, J. P. (2008). Forecasting with the age-period-cohort model and the extended chain-ladder model. *Biometrika*, *95*(4), 987–991.

Lipset, R., S.M. & Benedix. (1959). *Social mobility in indusatrial society.* Transaction Publishers.

Marshall, G., & Firth, D. (1999). Social mobility and personal satisfaction: evidence from ten countries. *The British Journal of Sociology*, *50*(1), 28–48.

Mason, K. O., Mason, W. M., Winsborough, H. H., & Poole, W. K. (1973). Some methodological issues in cohort analysis of archival data. *American sociological review*, 242–258.

Missinne, S., Daenekindt, S., & Bracke, P. (2015). The social gradient in preventive healthcare use: What can we learn from socially mobile individuals? *Sociology of Health and Illness*, *37*(6), 823–838.

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, (38(11)), 1–29.

Ormel, J., Lindenberg, S., Steverink, N., & Verbrugge, L. M. (1999). Subjective well-being and social production functions. *Social Indicators Research*, *46*(1), 61–90.

Präg, P., & Richards, L. (2019). Intergenerational social mobility and allostatic load in Great Britain. *Journal of Epidemiology and Community Health*, *73*, 100–105.

Schuck, B., & Steiber, N. (2018). Does Intergenerational Educational Mobility Shape the Well-Being of Young Europeans? Evidence from the European Social Survey. *Social Indicators Research*, *139*(3), 1237–1255.

Searle, S. R., & Khuri, A. I. (2017). *Matrix algebra useful for statistics*. John Wiley & Sons.

Sigal, M. J., & Chalmers, R. P. (2016). Play it again: Teaching statistics with monte carlo simulation. *Journal of Statistics Education*, *24*(3), 136–156.

Simandan, D. (2018). Rethinking the health consequences of social class and social mobility. *Social Science & Medicine*, *200*, 258–261.

Sobel, M. E. (1981). Diagonal Mobility Models : A Substantively Motivated Class of Designs for the Analysis of Mobility Effects. *American Sociological Review*, *46*(6), 893–906.

Sobel, M. E. (1985). Social Mobility and Fertility Revisited : Some New Models for the Analysis of the Mobility Effects Hypothesis. *American Sociological Review*, *50*(5), 699–712.

Sorenson, A. M. (1989). Husbands ' and Wives ' Characteristics and Fertility Decisions : A Diagonal Mobility Model. *Demography*, *26*(1), 125–135.

Sorokin, P. (1927). *Social mobility*. New York: Harper & Brothers.

Tolsma, J., De Graaf, N. D., & Quillian, L. (2009). Does intergenerational social mobility affect antagonistic attitudes towards ethnic minorities? *British Journal of Sociology*, *60*(2), 247–277.

Tumin, M. M. (1957). Some unapplauded consequences of social mobility in a mass society. *Social Forces*, *36*(1), 32–37.

Turner, H., & Firth, D. (2015). Generalized nonlinear models in R: An overview of the gnm package. *R Journal*, 1–61.

Van Berkel, M., & de Graaf, N. D. (1995). Husband's and Wife's Culture Participation and their Levels of Education: A Case of Male Dominance? *Acta Sociologica*, *38*(2), 131–149.

Van der Slik, F. W., De Graaf, N. D., & Gerris, J. R. (2002). Conformity to Parental Rules: Asymmetric Influences of Father's and Mother's Levels of Education. *European Sociological Review*, *18*(4), 489–502+i.

Van der Waal, J., Daenekindt, S., & de Koster, W. (2017). Statistical challenges in modelling the health consequences of social mobility: the need for diagonal reference models. *International Journal of Public Health*, *62*(9), 1029–1037.

Weakliem, D. L. (1992). Does Social Mobility Affect Political Behaviour? *European Sociological Review*, *8*(2), 153–165.

Willekens, M., Daenekindt, S., & Lievens, J. (2014). Whose Education Matters More? Mothers' and Fathers' Education and the Cultural Participation of Adolescents. *Cultural Sociology*, *8*(3), 291–309.

Yamaguchi, K. (2002). Regression Models with Parametrically Weighted Explanatory Variables. *Sociological Methodology*, *32*, 219–245.

Yang, Y., Fu, W. J., & Land, K. C. (2004). A methodological comparison of age-period-cohort models: The intrinsic estimator and conventional generalized linear models. *Sociological methodology*, *34*(1), 75–110.

III. When Attrition Affects Causal Interpretation in Panel Data Analysis: The Potential of the Joint Modeling Approach.

Outline

Bibliographic Information

This chapter is under revision at *Survey Research Methods*.

Author's contribution

I developed the methodology, writing preparation, interpretation of findings, and the data generation and analysis. Robin Samuel conceptualized, supervised, and reviewed the chapter.

Abstract

Missing data due to panel attrition poses a serious challenge in longitudinal survey analysis. Even if an unavoidable phenomenon, progressive dropout affects sample representativeness over time and can limit causal inference severely. The threats to causal inference due to attrition will intensify in case the missing data mechanism is related to a longitudinal process, leading to Missing Not at Random (MNAR). In this paper, we assess the Joint Modeling (JM) approach as a promising tool to tackle threats to causal inference due to MNAR dropout. The JM approach has been developed in the late 1990s in the biomedical sciences, but has not received much attention in the social sciences. Using simulated data, we examined how the JM performs under conditions of dynamic informative dropout. We compared it to the Linear Mixed Model (LMM) and the time-varying parametric (Weibull) survival regression model, testing these models against two theoretically based scenarios. Our results suggest that the JM performs better in terms of bias and efficiency of the estimates, than the parametric time-to-event regression. Compared with the LMM, the estimates of the longitudinal trend are similar. Concerning a scenario investigating unobservable time trend bias, results show that the JM approach provides correct estimates of the assumed longitudinal trend. The comparison between the JM, the LMM, and the time-varying covariate Weibull regression shows that the JM and LMM yield good approximations of the true population values, while the Weibull clearly fails to do so. We conclude that the JM approach can be of great use for social scientific research with, for example, a focus on dynamics of social change and the life course, where problems of attrition and MNAR are rampant in many of the widely used panel datasets.

Missing data due to panel attrition poses a severe challenge in longitudinal survey analysis. Even if an unavoidable phenomenon, progressive dropout affects sample representativeness over time and may limit causal inference severely (Laird, 1988; Lugtig, 2014; Schifeling et al., 2015; Vandecasteele & Debels, 2007). The threats to causal inference due to attrition might be significantly exacerbated in the case the missing data mechanism is related to a longitudinal process, leading to Missing Not at Random (MNAR, see Appendix A for a formal review of missing data processes; Billingham and Abrams, 2002; Diggle and Kenward, 1994; Stolz et al., 2018). In this scenario, panel data analysis of the longitudinal outcome variable would return biased and inconsistent estimates (Lugtig, 2014; Marini et al., 1980; Trappmann et al., 2015; Vandecasteele & Debels, 2007). From a statistical perspective, MNAR introduces two potential sources of bias: endogeneity and progressive homogeneity of the sample. Firstly, endogeneity occurs as the pattern of the longitudinal outcome variable is dependent upon and - at the same time - may be influenced by the probability of dropout of the observation unit. Ignoring the endogenous selection process leads to a problem of unobservables in the causal estimation, as it is possible to observe the evolution of the longitudinal variable only among units of observations that have not dropped out from the study[14] (Halaby, 2004; Little, 1995; Papageorgiou et al., 2019; Rubin, 1976). In terms of sample representativeness, informative dropout leads to a "survival of the fittest" process. That is, the observation units within the sample tend to be more homogeneous over time according to determining key characteristics. Abbring and Van Den Berg, 2007; Balan and Putter, 2019; Haviland et al., 2011; Liu et al., 2010. This paper assesses the Joint Modeling (JM) for longitudinal, and survival data approach as a promising tool to tackle threats to causal inference due to MNAR dropout. Early developments of the JM approach have been within the field of biomedical studies on HIV/AIDS[15]. The main purpose that drove the development of the JM was to assess the association between the survival of patients and the longitudinal trajectory of the CD4 biomarker, which is related to AIDS progression (Wulfsohn and Tsiatis 1997; Wang and Taylor 2001). The model was then further developed in cancer studies and quality-of-life (QOL) studies (Chi and Ibrahim 2006 and in clinical trials in other fields (for instance, Xu and Zeger 2001 and Henderson and Oman 1999; for a review, see Tsiatis and Davidian 2004). Although the increasing interest and empirical applications of the JM in the biomedical sciences, this approach has not yet received much attention in the social sciences (only very recently did sociological literature start to provide attention to the JM. See Li et al., 2020 for an example)[16]. The structure of the JM approach combines two underlying submodels: the longitudinal component (the outcomes measured over time) and the survival component (the time

until an event occurs). In practical terms, the JM estimates simultaneously a given longitudinal pattern employing a linear mixed model (LMM) and a time-to-event regression model [17], thereby allowing to accommodate a given MNAR process. The most common strategy to link the two sub-models is the shared parameter framework, which joins the two components through shared random effects. The main advantage of the JM approach lies in the possibility of assessing the degree of association between the longitudinal and the dropout pattern. These components then share part of the parameter distribution through shared random effects (for details, see Rizopoulos 2011). Additionally, the model allows for either fixed (such as class of origin or gender) or time-varying (such as the age of individuals) covariates in the longitudinal and the survival sub-models. This association makes it possible to assess direct and indirect effects on the overall dropout mechanism (Ibrahim et al., 2010). In this sense, the JM model allows the empirical researcher to gain additional insights by modeling and quantifying the influence of the missing data process on the longitudinal outcome of interest. Despite the apparent advantages of the JM, its potential in a social scientific setting and performance compared to other models remains unclear.

Addressing this research gap, we set out to examine how the JM performs under conditions of dynamic informative dropout. To help contextualize our findings, we compared them to the Linear Mixed Model and the time-varying parametric (Weibull) survival regression model, both models widely used in the social sciences. We tested all models against two theoretically-based scenarios: first, we examined the models' behavior in the case of omitted variable bias; second, we focused on the specification of the assumed time change. The problem of unobserved heterogeneity has been widely recognized within both the biomedical (Liu, 2013; Liu et al., 2010; Zheng, 2020) and the sociological literature (Blossfeld & Hamerle, 1989). We simulated the longitudinal pattern as a linear trend adjusted for a binary and a time-varying continuous covariate in the first scenario. The design of the dropout process depended upon the same set of covariates plus the endogenous longitudinal outcome. We then compared the models with and without the time-varying regressor to evaluate the performance of the estimators. In the second scenario, the simulated longitudinal data and the dropout follow a cubic spline approximation probability function. This scenario examines the characteristics of the models with and without the nonlinear terms (quadratic and cubic parameters). Crowther et al. (2016) investigated the JM approach behavior in the presence of time misspecification, comparing the efficiency of different computational characteristics and association parameters. We generated three simulation studies within each scenario with varying strengths of the association between the longitudinal outcome variable and the dropout process.

We performed 200 replications for each scenario and 1000 observation units for each repetition. The remainder of the paper is as follows: the next section provides an overview of the computational characteristics of the JM approach and additional details on the estimation method deployed by the model. Subsequently, we depict the Monte Carlo simulation design and a descriptive overview of the simulated datasets. The following section describes the results of the Monte Carlo simulation. The section will cover both scenarios by the different association degrees set in the true population model. Finally, we conclude with the implications for applied social researchers of this study.

III.1 The Joint Modeling Approach

In order to see how the JM models the association between longitudinal measurements and chances of event occurrence (such as drop-out from the study), assume that T_i is the observed drop-out time of the i^{th} individual, taken as the minimum of the time of drop-out T_i^*, and C_i is then the censoring time such that $T_i = \min(T_i^*, C_i)$. Define $\delta_i = I(T_i^* \leq C_i)$ as the function of the event indicator, which takes the value 1 if the drop-out occurs (that is, $T_i^* \leq C_i$ is true), 0 otherwise. The observed longitudinal outcome for the individual can be conceived as a vector $Y_{ij} = \{y_i(t_{ij}), j = 1, \ldots j, n_i\}$ where t_{ij} denotes the outcome at time measurement j (Rizopoulos 2011). The true but unobserved longitudinal outcome can be denoted as m_i. The standard option to quantify the unobserved $m_i(t)$ and its effect on the risk that the event occurs is to use a time-to-event regression model such as:

$$\begin{aligned} h_i(t|\ \mathcal{M}_i(t), w_i) = & \\ & = \lim_{dt \to 0} \Pr\{t \leq T_i^* \min t + dt|\ T_i^* \geq t, \mathcal{M}_i(t), w_i\}/dt \\ & = h_0(t)\exp\{\gamma^T w_i + \alpha m_i(t)\}, \end{aligned} \tag{III.1}$$

where $h_0(t)$ in equation III.1 represents the baseline hazard as a function of time and $\mathcal{M}_i(t)$ denotes the history of the longitudinal unobserved outcome. The baseline hazard function can be left unspecified, as in a semi-parametric Cox regression, or be assumed to take a specific distribution, such as the exponential, Weibull, or Gompertz distributions; γ^T represents the vector of coefficients for the set of individual covariates w_i. The parameter α quantifies the association between the risk the event occurs and the true longitudinal history of the covariate. In the biostatistical literature, there are many ways in which this parameter is quantified (see Crowther et al. 2016) for a brief description). In our application, we will use the current value association, as it is the most used in biostatistical and biomedical

research. This parameter returns the hazard change due to a 1-unit change of the longitudinal outcome variable. The term m_i represents the longitudinal trajectory of the i^{th} individual as a function of time. To reconstruct the unobserved m_i trajectory, the observed y_{it} can be used[18]. Assuming that the longitudinal covariate is normally distributed, it is possible to model $\mathcal{M}_i(t)$ using y_{it} for each individual estimating a linear mixed effect model of the form

$$\begin{aligned} y_i(t) &= m_i(t) + \epsilon_i(t) \\ m_i &= x_i^T(t)\beta + z_i^T(t)b_i + \epsilon_i(t), \end{aligned} \tag{III.2}$$

where y_i is the outcome of interest of random variable with normal distribution $\mathcal{N} \sim (\mu, \sigma^2)$. Equation III.2 shows that the true longitudinal pattern m_i can be decomposed in a vector of fixed effects of β parameters and a b_i vector of random effects. The associated $x_i(t)$ and $z_i(t)$ form the rows of the design matrices for the fixed and random effects, respectively. The random effects are to be distributed as $b_i \sim \mathcal{N}(0, \Sigma)$, where Σ is the covariance matrix:

$$\Sigma = \begin{bmatrix} \sigma_{00}^2 & \\ \sigma_{01}^2 & \sigma_{11}^2 \end{bmatrix}. \tag{III.3}$$

The matrix states the covariance between the residuals of levels 2 (individuals) and 1 (repeated measurements or observations). That is, σ_{00}^2 and σ_{11}^2 represent the variance components of the individuals' specific intercepts and the slopes, respectively, while the term σ_{01}^2 represents the correlation between the random intercept and the random slope for each individual in the longitudinal study. As per Papageorgiou et al. (2019), the estimation methods developed for the Joint Modelling approach in the biostatistical literature follow both Bayesian and frequentist theories. Here, we will focus on the frequentist approach (but see Hu et al., 2009 for the Bayesian alternative). The common point of these paradigms is in the full joint likelihood from the joint distribution of the longitudinal and the time-to-event outcomes. The joint likelihood function is:

$$\begin{aligned}\mathcal{L}(\theta \mid \mathcal{D}_n) &= \prod_{i=1}^{n} p(T_i, T_i^U, \delta_i, \mathbf{y}_{1i}, \ldots, \mathbf{y}_{Ki}; \theta) \\ &= \prod_{i=1}^{n} \int p(T_i, T_i^U, \lambda_i, \mathbf{y}_{1i}, \ldots, \mathbf{y}_{Ki}; \theta) d\mathbf{b}_i \\ &= \prod_{i=1}^{n} \int \left\{ \prod_{k=1}^{K} \prod_{l=1}^{n_{ki}} p(y_{kil} \mid \mathbf{b}_{ki}; \theta) \right\} p(T_i, T_i^U, \delta_i \mid \mathbf{b}_i; \theta) p(\mathbf{b}_i; \theta) d\mathbf{b}_i.\end{aligned} \tag{III.4}$$

Equation 4 can be re-expressed as in Crowther et al. (2016):

$$\prod_{i=1}^{N} \left[\int_{-\infty}^{\infty} \left(\prod^{n_i} p(y_i(t_{ij}) \mid b_i) \right) p(b_i \mid \theta) p(T_i, d_i \mid b_i, \theta) db_i \right], \tag{III.5}$$

where the longitudinal outcome is:

$$p(y_i(t_{ij}) \mid b_i) = (2\pi\sigma_e^2)^{-1/2} \exp \left\{ -\frac{[y_i(t_{ij}) - m_i(t_{ij})]^2}{\sigma_e^2} \right\}, \tag{III.6}$$

and the multivariate normally distributed random effects are:

$$p(\mathbf{b}_i \mid \theta) = (2\pi|V|)^{-q/2} \exp \left\{ -\frac{b_i' V^{-1} b_i}{2} \right\}. \tag{III.7}$$

The survival outcome is:

$$\begin{aligned}p(T_i, d_i \mid b_i, \theta) = &[h_0(T_i) \exp(\alpha m_i(t) + \phi v_i)]^{d_i} \\ &\times \exp \left\{ - \int_0^{T_i} h_0(u) \exp(\alpha m_i(t) + \phi v_i) du \right\}.\end{aligned} \tag{III.8}$$

In the frequentist approach, the estimates (represented by the vector θ of equation 4) can be derived through Maximum Likelihood Estimation (MLE), using the expectation-maximization (E-M) algorithm (Dempster, Laird, and Rubin 1977; Wulfsohn and Tsiatis 1997) by maximizing the log-likelihood using the Newton approach (Thiébaut and Bénichou 2004). The estimation, however, is computationally demanding, as the integral over the random effects does not have a closed-form solution. Crowther et al. (2013), Crowther

et al. (2016) propose in their `stjm` package in Stata the Gauss–Hermite quadrature to evaluate intractable integrals[19]. Nowadays, the concepts and principles of equity in scientific health literature have further developed the JM approach to accommodate different data and estimation methods. For instance, the latent class specification Proust-Lima et al., 2014, recurrent events Huang and Liu (2007), and Li et al. (2020) provides a theoretical and empirical exposition. Concerning the longitudinal outcome, Viviani et al. (2014) presented a specification of the JM to include dichotomous and count data; finally, Hu, Li, and Li (2009) and Cremers, Mortensen, and Ekstrøm (2021) developed a Bayesian framework for the specification of the JM approach.

III.2 Simulation Design

The Monte Carlo Simulation consisted of two main scenarios, in which we tested three statistical models: a) the Linear Mixed Model (LMM); b) the time-to-event regression model (with Weibull distributed baseline hazard ratio), and c) the JM approach. The first scenario aims to test the models' behavior in the presence of unobserved heterogeneity, specifically when omitted variable bias is present. The second scenario tests the models' performance in the context of misspecification of the longitudinal outcome variable. The Data Generating Process (DGP) for both scenarios relied on generating two data types: firstly, we generated the dropout mechanism. Secondly, we generated the longitudinal outcome variable, assuming its values as continuous and normally distributed. To generate data in which the longitudinal outcome and the dropout mechanism have different association degrees, we modeled three DGPs within each scenario, where the α parameter (i.e., the parameter quantifying the association) takes values $0, 0.25, 0.5$. This setting is useful to compare the statistical models under no association ($\alpha = 0$, assuming thus Missing at Random dropout mechanism), moderate association ($\alpha = 0.25$), and strong association ($\alpha = 0.5$). The measures of performance through which we compared the models are the bias from the population true value and the coverage rate. The bias measurement is computed as $\left(\bar{\hat{\beta}} - \beta\right)$, where $\bar{\hat{\beta}}$ is the average value of the simulated estimates and β is the true population value(Burton et al., 2006). In Appendix B, the reader can find the bias summary, mean squared error (MSE), and coverage for the parameters of interest. For the unobserved heterogeneity scenario, the parameters of interest are the βX_2 (i.e., the group comparison parameter) and α parameters. For what concerns the time misspecification, the analysis focused on the cubic

term of the longitudinal pattern ($\beta(t)^3$) and the α parameter. The software used for the generation, simulation, and analysis of the results is Stata, version 15.1. The packages used are: for the generation of the survival data, `simsum` (Crowther and Lambert 2013); for the computation of Joint Modeling, `stjm` (Crowther et al. 2013); and for the analysis of the simulated datasets, `simsum` (White 2010). The tables and the graphs shown in the Results section with R.

Data Generating Process

Unobserved heterogeneity

Schumacher, Olschewski, and Schmoor (1987) and Schmoor and Schumacher (1997) have shown (analytically and by simulation) that the omission of relevant explanatory variables could lead to a severe bias toward zero in the estimates. The omission of explanatory variables, which leads to omitted variable bias, can be translated into an unobserved heterogeneity issue. In this data generation process, we want to test the three models over the presence of unobserved heterogeneity in the sub-optimal model. This scenario aims to assess how much we lose if a source of heterogeneity is left unspecified in the model, given the specificity of the JM to exploit frailty in both longitudinal and survival submodels. The true model, which generates the data, and the fitted model (where we did not include the measurement of X_2) can respectively be expressed as:

True Model

$$m_i = 1.5 + .5(t) + .9X_1 + .05X_2 + e_{ij} \tag{III.9}$$

$$e_{ij} = \mathcal{N}(0, \Sigma) = \Sigma = \begin{bmatrix} \sigma^2_{00} & \\ \sigma^2_{01} & \sigma^2_{11} \end{bmatrix} \begin{cases} \sigma^2_{00} = 2.5 \\ \sigma^2_{11} = 1.5 \\ \sigma^2_{01} = \ .3 \end{cases}$$

$$h(t \mid \beta_i) = .05 * 1.25t^{.05-1} + \exp[\alpha * (\beta_0 + \beta_1 + .9X_1 + .05X_2)]$$

Fitted Model

$$m_i = 1.5 + .5(t) + .05X_2 + e_{ij} \tag{III.10}$$

$$e_{ij} = \mathcal{N}(0, \Sigma) = \Sigma = \begin{bmatrix} \sigma^2_{00} & \\ \sigma^2_{01} & \sigma^2_{11} \end{bmatrix} \begin{cases} \sigma^2_{00} = 2.5 \\ \sigma^2_{11} = 1.5 \\ \sigma^2_{01} = \ \ .3 \end{cases}$$

$$h(t \mid \beta_i) = .05 * 1.25t^{.05-1} + \exp[\alpha * (\beta_0 + \beta_1 + .05X_2)]$$

The first equations of both the true and the fitted models represent the DGP for what concerns the fixed effects of the longitudinal pattern. The covariates included in the true population values are one time-varying (X_1) and dichotomous (X_2). Between the categories of X_2, we have let the dropout rate be different to test the models' capability to detect the true effect under different dropout paces. Subsequently, the true random-effects population values in the variance-covariance matrix are the random intercepts (σ^2_{00}), the random slopes (σ^2_{11}), and their correlation to each other (σ^2_{01}). The last equation represents the true dropout mechanism of the simulated sample. Figure III.1 shows graphically the simulated data. Figure III.1a shows the longitudinal pattern across the follow-up years of study among individuals that did not drop out (on the left) vs. the individuals that left prematurely (on the right) when $\alpha = 0.5$.

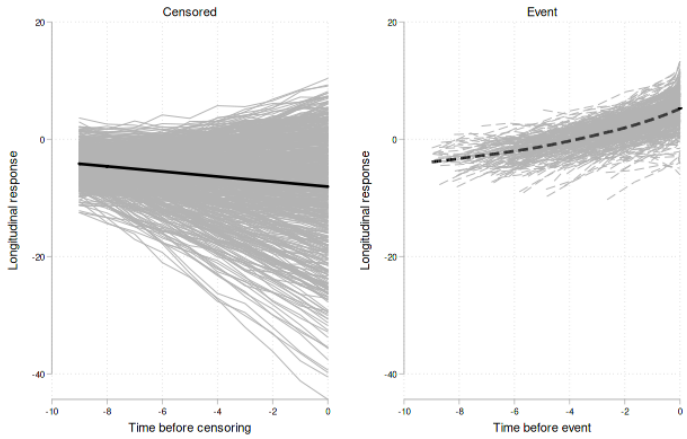
The sub-graph shows clearly that the two longitudinal patterns differ substantially between the two groups due to informative dropout under process. Figures III.1b and III.1c show the longitudinal gradient of the longitudinal outcome and the Kaplan-Meier survival rates, respectively.

Time misspecification

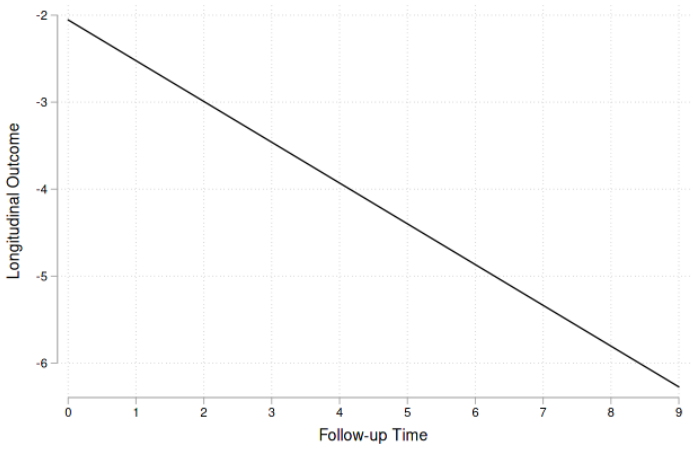
In this data generation process, we want to test the LMM, JM, and Weibull survival models in the presence of misspecification of the longitudinal pattern. Suppose that the true but unobserved longitudinal history $\mathcal{M}_i$ of the longitudinal outcome variable has a cubic shape. However, the researcher (who can compute the model on the observed y_{it}) is unaware of the more complex pattern and estimates the longitudinal coefficient in a more straightforward linear form.

True and Fitted Model

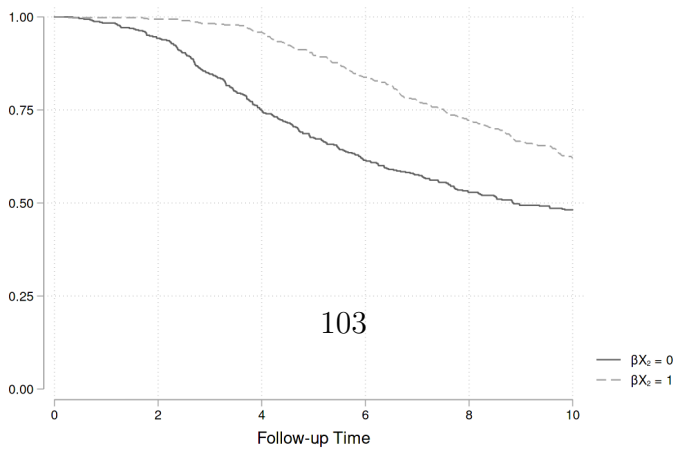
Figure III.1: Descriptive graphs of the DGP. Panel (a) Shows the Differences between the Stayers and the Dropped-out. Panel (b) Shows the Linear Longitudinal Trend. Panel (c) Shows the Kaplan-Meier Estimates to Quantify the Drop-Out Rate.



(a) Longitudinal Pattern by who remained in the study (on the left) vs. who dropped-out (on the right).



(b) Simulated Longitudinal Pattern



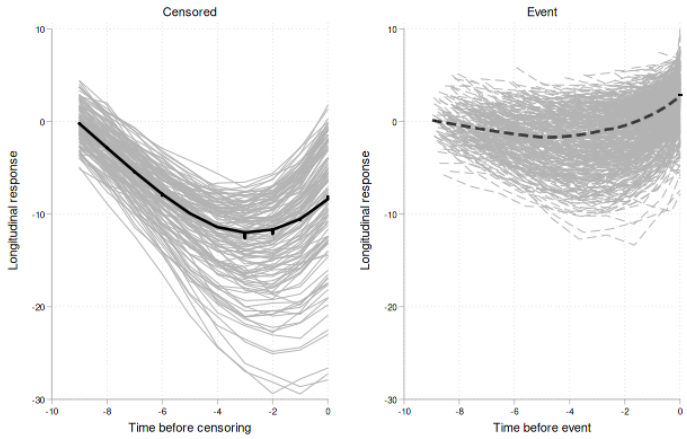
(c) Simulated Survival Curves

$$m_i = 1.5 + .025(t) - .3(t)^2 + .05(t)^3 + e_{ij} \tag{III.11}$$

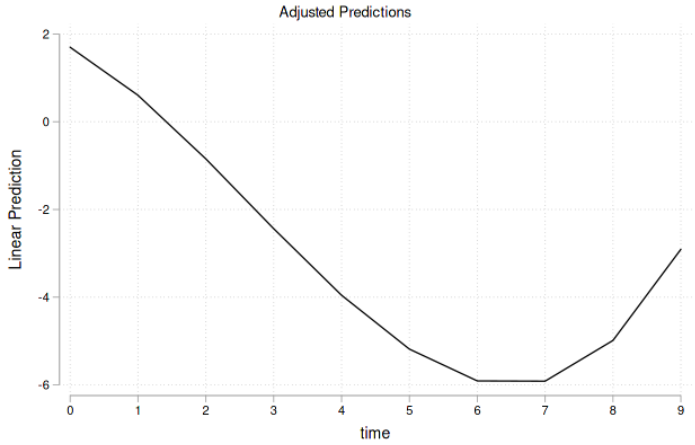
$$e_{ij} = \mathcal{N}(0, \Sigma) = \Sigma = \begin{bmatrix} \sigma^2_{00} & \\ \sigma^2_{01} & \sigma^2_{11} \end{bmatrix} \begin{cases} \sigma^2_{00} = 2.5 \\ \sigma^2_{11} = 1.5 \\ \sigma^2_{01} = \ \ .3 \end{cases}$$

$$h(t \mid \beta_i) = .05 * 1.25t^{.05-1} + \exp[\alpha * (1.5 + ..025(t) - .3(t)^2 + .05(t)^3)]$$

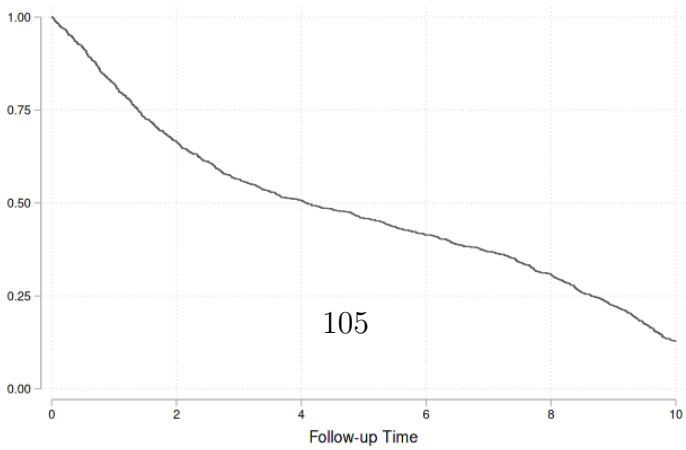
Figure III.2: Descriptive graphs of the DGP. Panel (a) Shows the Differences between the Stayers and the Dropped-out. Panel (b) Shows the Non-Linear Longitudinal Trend. Panel (c) Shows the Kaplan-Meier Estimates to Quantify the Drop-Out Rate.



(a) Longitudinal Pattern by who remained in the study (on the left) vs. who dropped-out (on the right).



(b) Simulated Longitudinal Pattern



(c) Simulated Survival Curves

Similarly to figure III.1, figure III.2 represents the simulated informative dropout, the gradient of the longitudinal outcome, and the Kaplan-Meier estimate of dropout. The subgraph a shows the different longitudinal trends over time between individuals that did not drop out vs. those who had done when $\alpha = 0.5$. Subfigures b and c show the cubic pattern (i.e., the true longitudinal trend) of the longitudinal outcome and the dropout rate from the study.

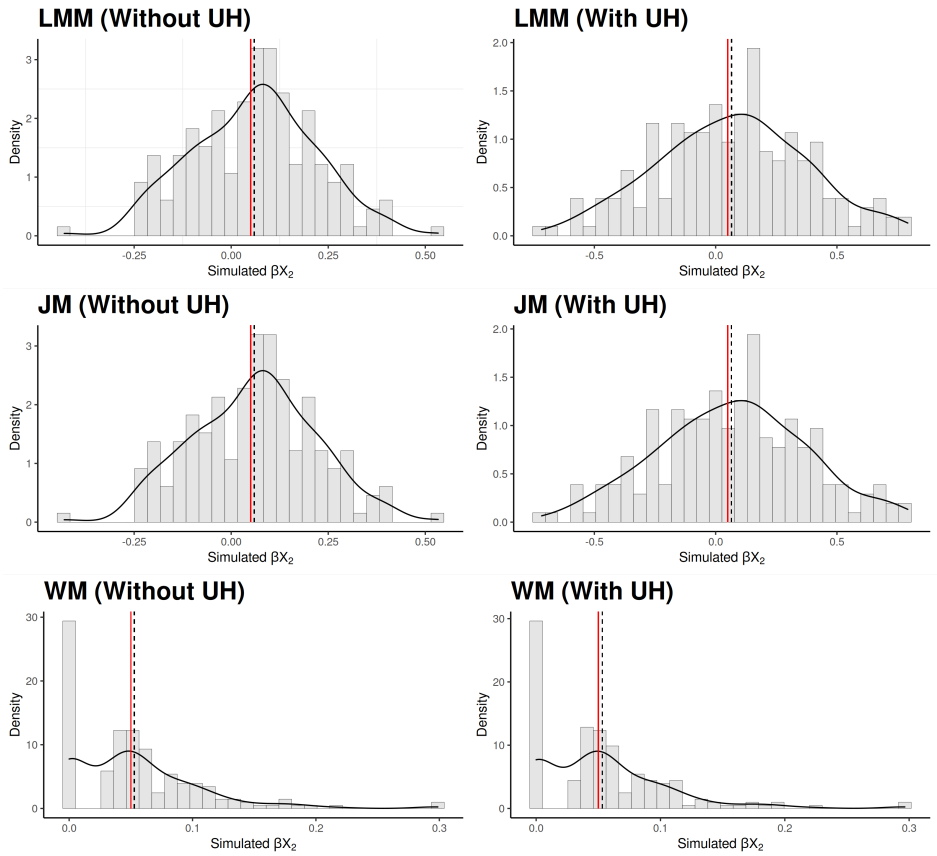
III.3 Results

Considering the unobserved heterogeneity scenario, the figures in this section show the distributions of the 200 parameters simulated through the Monte Carlo technique. The figure shows the histogram and the computed density plot of the parameter distributions. The dashed black line represents the mean of the distribution, while the red line represents the true population value. The degree of bias of the estimators of interest is the distance between the dashed black line and the red line. Appendix B shows the results in tabular format for both scenarios. Specifically, the performance measures we assess the models' behavior are: a) the degree of bias; b) the empirical standard error; c) the mean squared error; d) the Root Mean Squared (RMS) model-based standard error, and e) 95% coverage intervals.

No Association

We begin with the DGP constructed so that there is no association between the longitudinal outcome and the dropout mechanism. Figure III.3 represents the distributions of the group comparison parameter βX_3 estimated by the LMM, the JM, and the WM. In the left column, figure III.3 shows the distributions for the models deployed to analyze the true DGP. In the right column, the figure compares the same estimator but in the presence of unobserved heterogeneity.

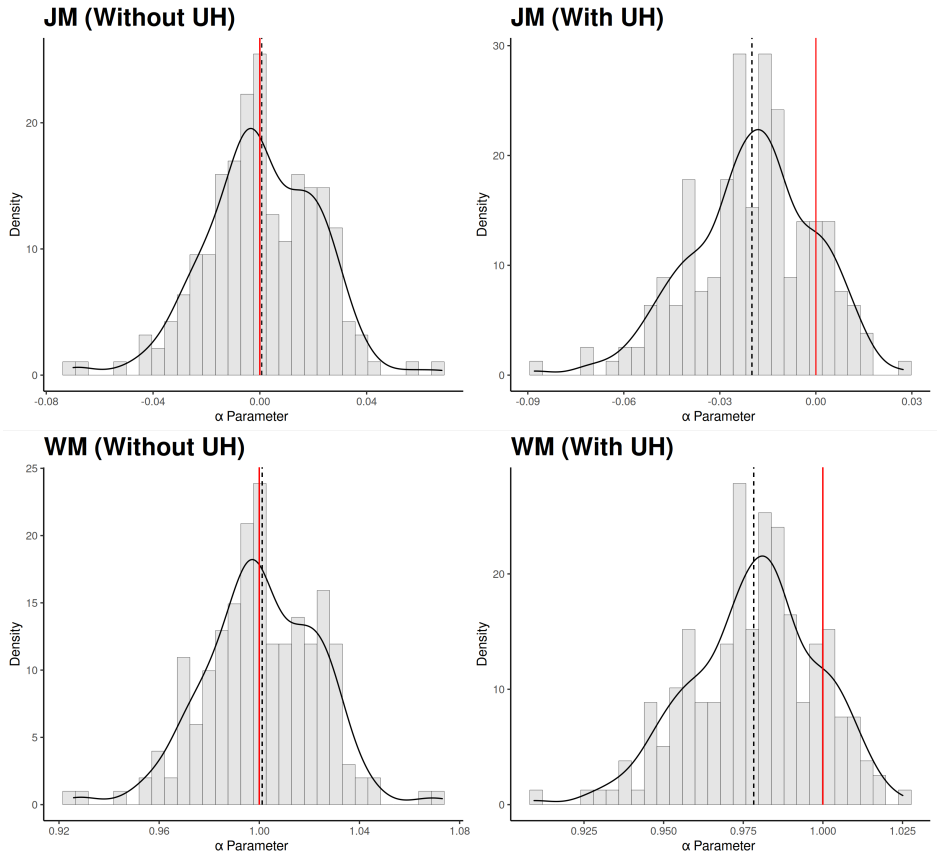
Figure III.3: Distribution of simulated group comparisons βX_3 estimated by the LMM, JM and WM. In the left column, the models are deployed to the true DGP. On the right column, the DGP is affected by unobserved heterogeneity.



It is immediately clear from the graph that the LMM and JM are unbiased estimators for what concerns the group comparison parameter $\beta_3 X_3$, as the true value and the mean of the distribution are very close. For what interests the Weibull regression, also, in this case, the estimator is unbiased. However, the distribution clearly shows the presence of outliers. This finding suggests that, even if the dropout mechanism at this stage should not influence the inference, the model does not return efficient estimations.

Considering the α parameters computed by the WM and the JM, figure III.4 depicts the distributions of the simulated parameters. Figure III.4 shows that, conversely to the βX_3 estimations, the α parameters of both the models are slightly affected by unobserved heterogeneity.

Figure III.4: Distribution of simulated α parameters estimated by the LMM, JM and WM. In the left column, the models are deployed to the true DGP. On the right column, the DGP is affected by unobserved heterogeneity.

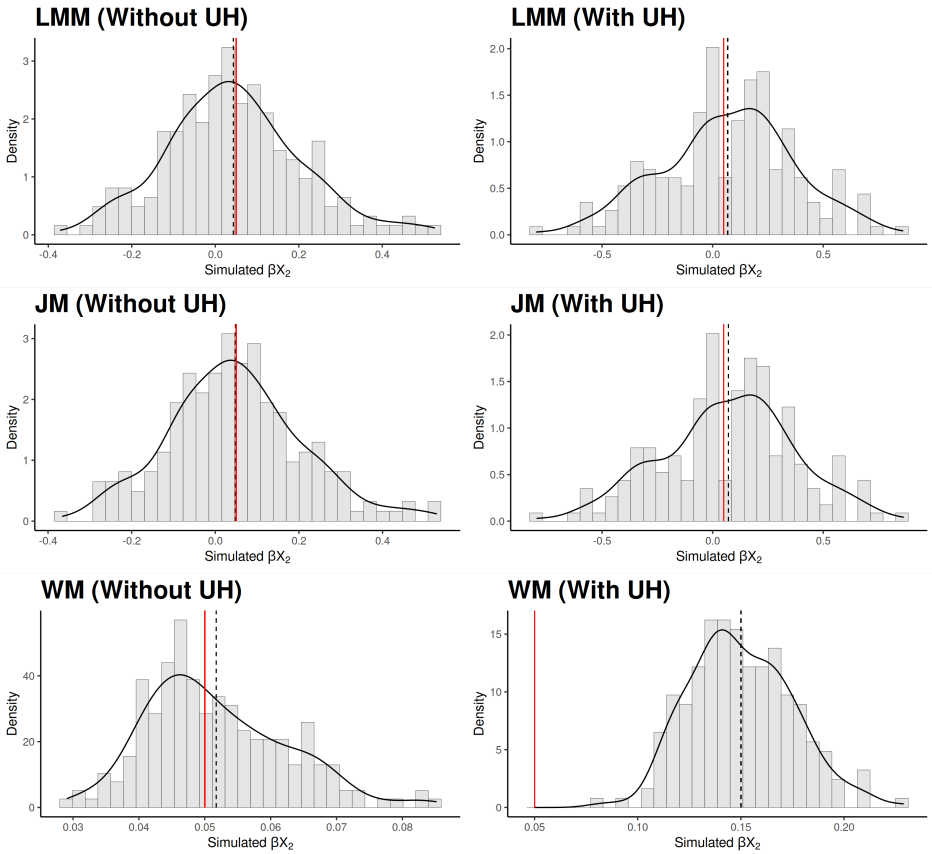


Specifically, the right column of figure III.4 shows evidence of bias toward zero, as the mean of the distribution is on the left concerning the red line, representing the true value set on the simulation scenario. The left column depicted in the figure shows that, under no association between the longitudinal outcome and the dropout process, both the models can detect the true population value. It is worth noting that the WM can capture the true value more efficiently than the previous comparison among the models.

Moderate Association

In this scenario, the association between the longitudinal outcome and the dropout process is moderate. As in the previous part, where the Monte Carlo simulation assumed MAR, the models' comparison focuses on the dichotomous variable coefficient β_3 and the association α parameters.

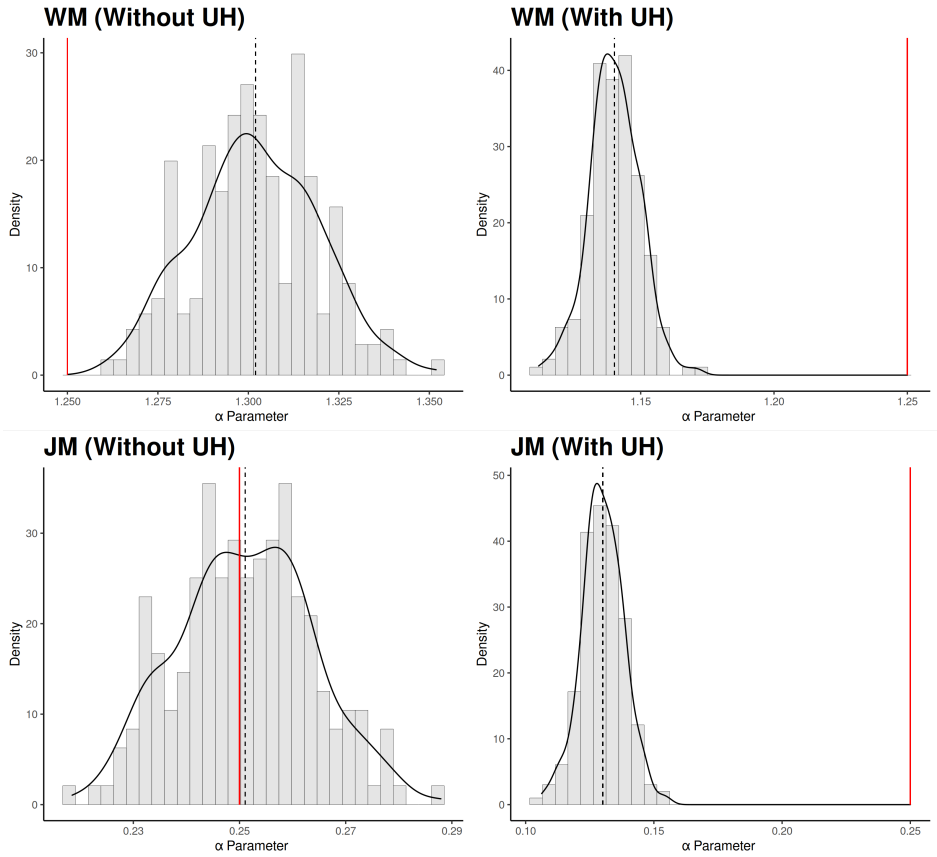
Figure III.5: Distribution of simulated group comparisons βX_3 estimated by the LMM, JM and WM. In the left column, the models are deployed to the true DGP.



The left column of figure III.5 shows that all the models under study return an unbiased estimation of the coefficient βX_3. When unobserved heterogeneity is present, evidence suggests a different behavior between the models. While the LMM and the JM return unbiased estimates and a similar distribution, the WM performs worse in comparison. Indeed, figure III.5 shows that the Weibull regression returns upwardly biased estimations of βX_3 coefficient.

Figure III.6 shows the distribution of the α parameter, comparing the Wm vs. the JM estimations.

Figure III.6: Distribution of simulated α parameters estimated by the LMM, JM and WM. In the left column, the models are deployed to the true DGP. On the right column, the DGP is affected by unobserved heterogeneity.

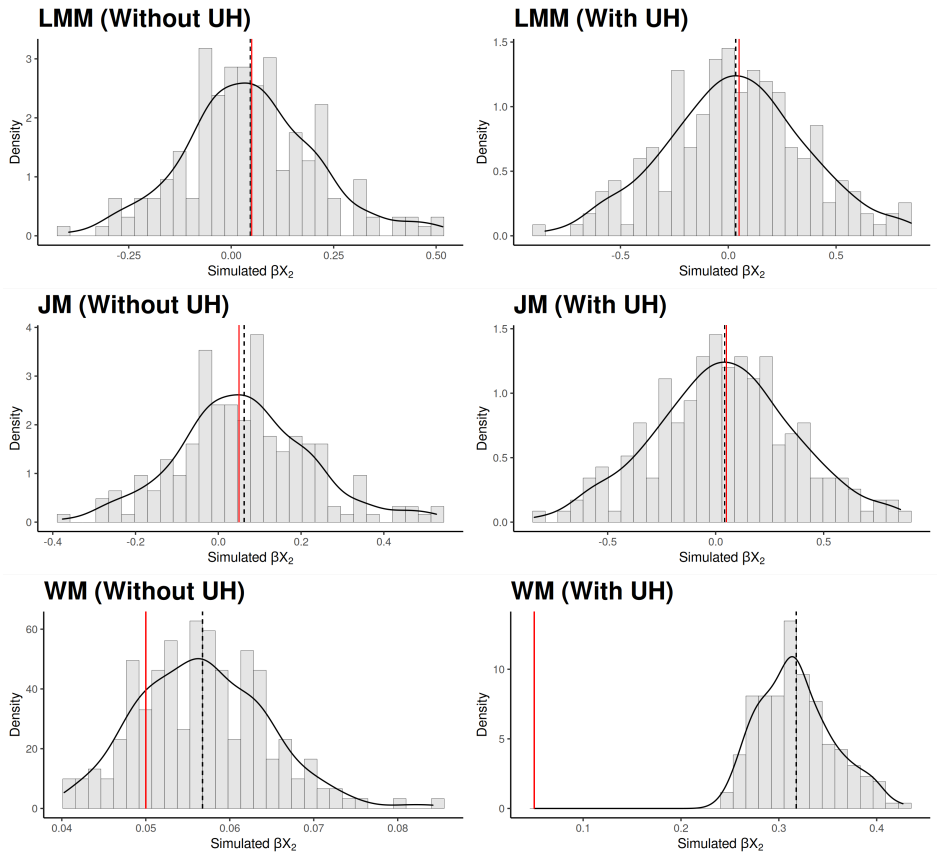


From the left column of the graph, it is possible to note that the WM estimation of the effect of the longitudinal outcome on the dropout rate is upwardly biased when the model is applied to the true DGP, while the JM can correctly estimate this association. In the case of unobserved heterogeneity, the column's right column shows a similar pattern between the two models. Indeed, both distributions of simulated parameters tend to underestimate the association between the longitudinal outcome and the dropout mechanism.

Strong Association

The last DGP computed for the unobserved heterogeneity scenario strongly correlates with the longitudinal pattern and the dropout rate. The results generally show a similar depiction of the models' behavior to the moderate association scenario. Figure III.7 shows the distribution of the coefficients βX_3 estimated by the statistical models.

Figure III.7: Distribution of simulated group comparisons βX_3 estimated by the LMM, JM and WM. In the left column, the models are deployed to the true DGP. On the right column, the DGP is affected by unobserved heterogeneity.

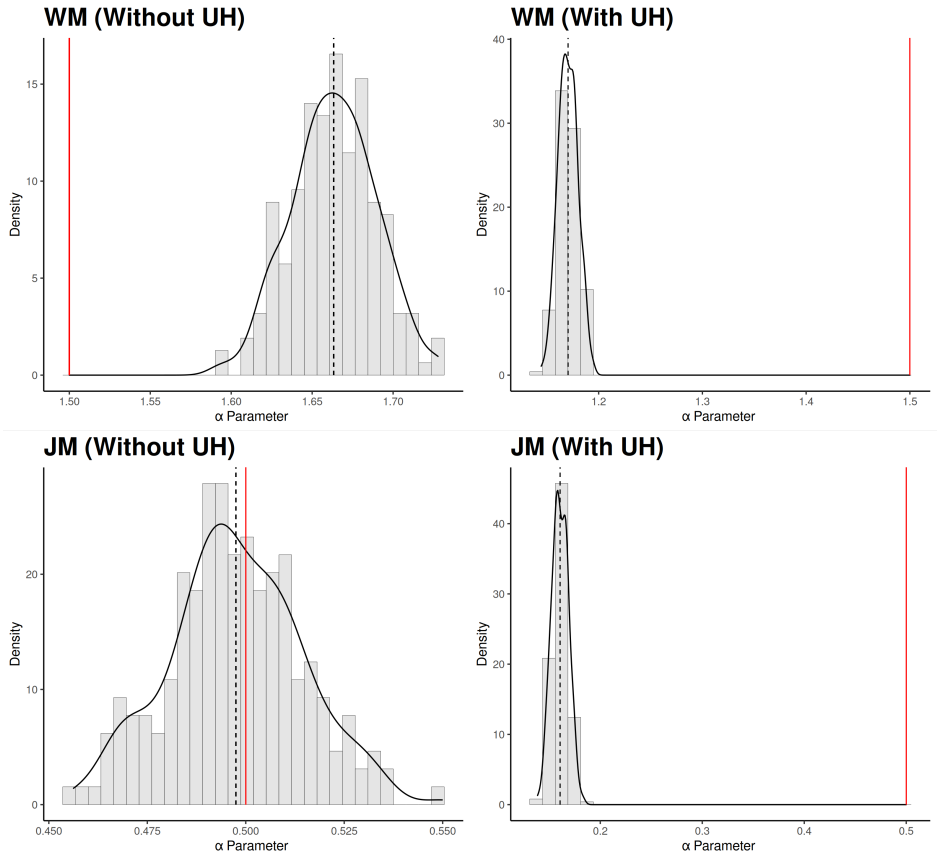


As in the previous DGP, figure III.8 shows that both the JM and the LMM provide unbiased estimates of the dichotomous variable βX_3. Even more, the right column indicates that the two models are quite robust against omitted variable bias.

The figure, however, shows a different depiction of what concerns the estimations computed by the WM. The true and the omitted variable models are upwardly biased (more seriously biased when the omitted variable bias is present).

Moving to the detection of the α parameter among the WM and the JM models, figure III.8 shows the relative distributions of the parameters computed by the models across the Monte Carlo repetitions.

Figure III.8: Distribution of simulated α parameters estimated by the LMM, JM and WM. In the left column, the models are deployed to the true DGP. On the right column, the DGP is affected by unobserved heterogeneity.



The left column of the figure shows substantial differences between the two models regarding the capability of detecting the effect of longitudinal effects on dropout. Indeed, if the JM returns unbiased estimates for the population generating process, the WM fails to do so, returning heavily upward biased estimates. On the contrary, both models show similar behavior in the omitted variable bias case, showing serious bias toward zero.

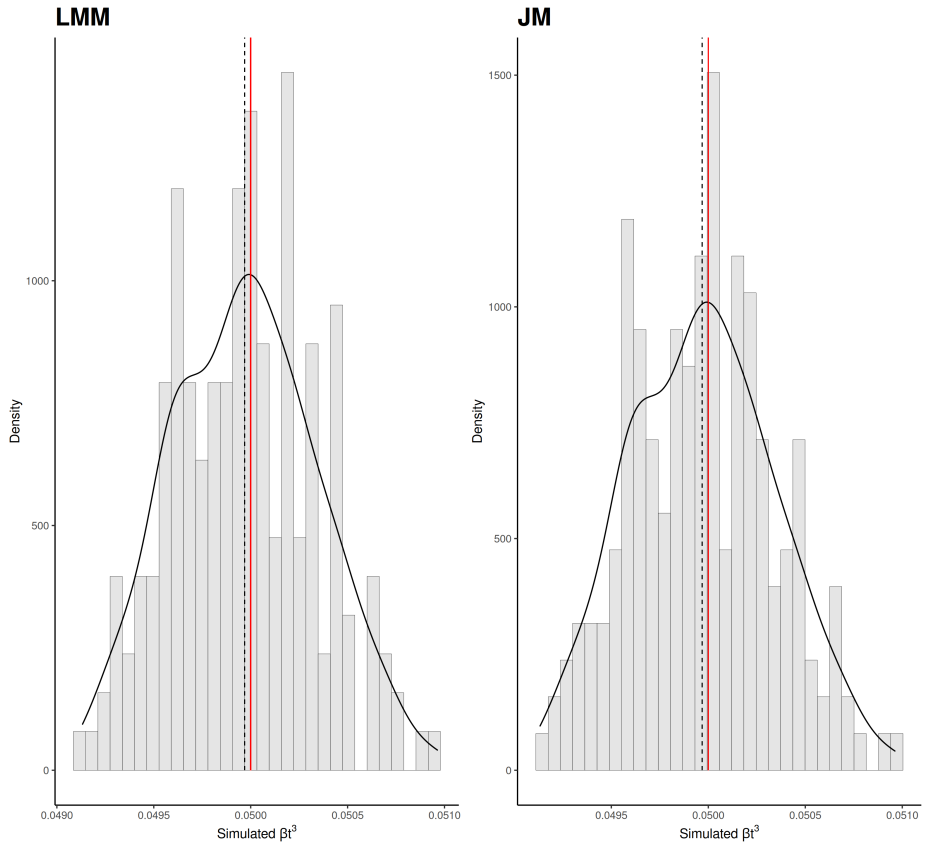
Time misspecification

The second scenario we implemented through Monte Carlo simulation aimed to investigate the capabilities of the models to detect more complex longitudinal patterns. This section shows the distributions of the parameters associated with the cubic term of the longitudinal outcome of interest. We compare the LMM and the JM for what concerns βt^3. We compare the WM and the JM concerning the α parameter. As in the previous section, we created three DGPs according to the different degrees of association between the longitudinal and the dropout mechanisms in this scenario.

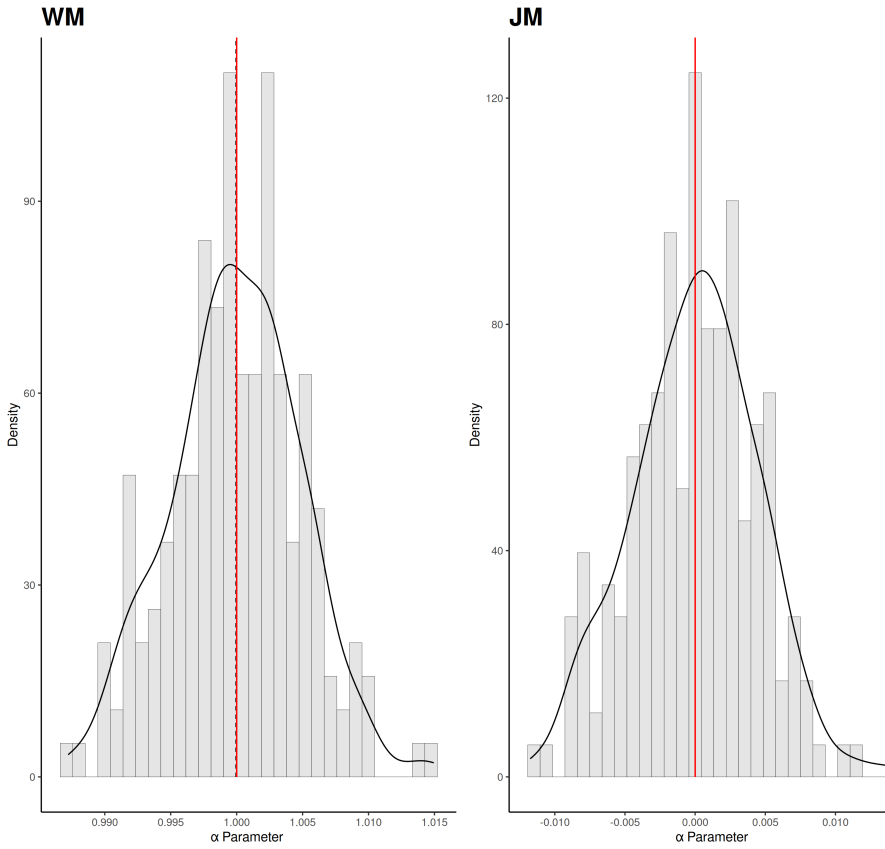
No Association

The DGP assuming MAR sets the parameter that shapes the influence of the longitudinal outcome on the drop out rate (the)α parameter) at zero. Figure III.9 shows the distribution of the cubic term of the time of measurement computed by the LMM and the JM.

Figure III.9: Distribution of simulated cubic term βt^3 estimated by the LMM an the JM and the α parameter by the WM and JM.



(a) Distribution Parameter βt^3. On the left: LMM. On the right: JM



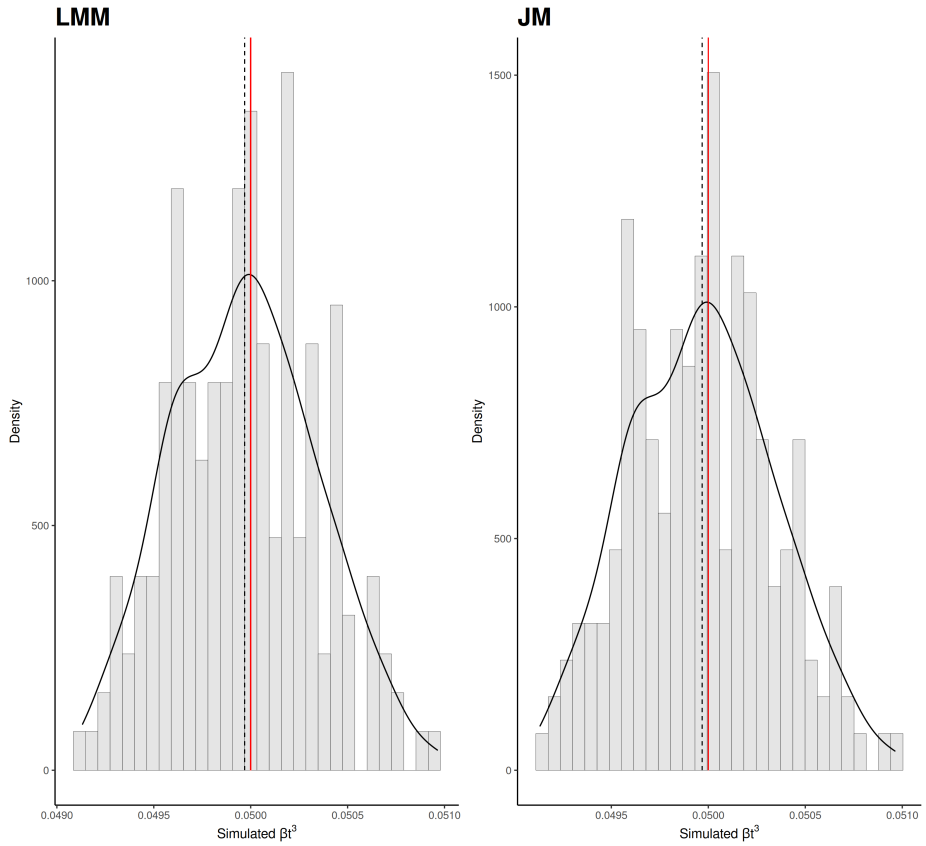
(b) Distribution Parameter α. On the left: WM. On the right:

The graph shows that the two models behave similarly, as the distributions are both unbiased. Regarding the $\alpha parameter$, the JM and the WM are very similar, presenting no bias.

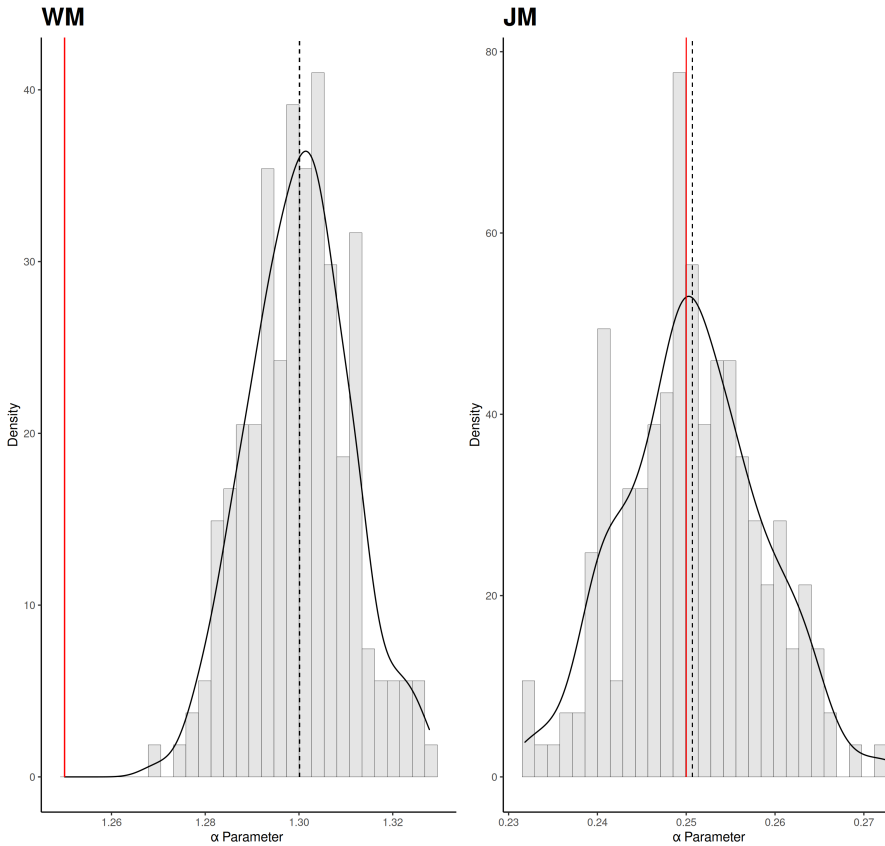
Moderate Association

When we model a moderate association between the longitudinal outcome and the dropout rate, figure III.10 shows that both the JM and the LMM provide reasonable estimations of the cubic term related to the longitudinal pattern.

Figure III.10: Distribution of simulated cubic term βt^3 estimated by the LMM an the JM and the α parameter by the WM and JM.



(a) Distribution Parameter βt^3



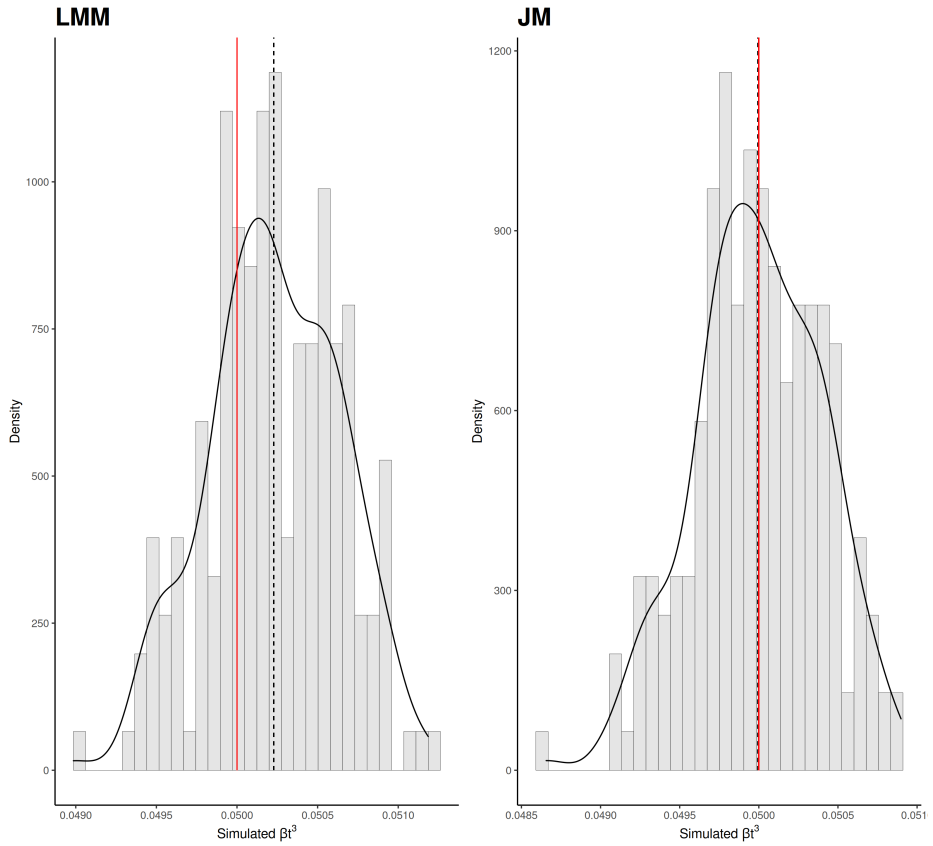
(b) Distribution Parameter α

The bottom part of the figure, representing the α parameter, shows evidence of upward bias for what concerns the WM estimator but no bias for the JM.

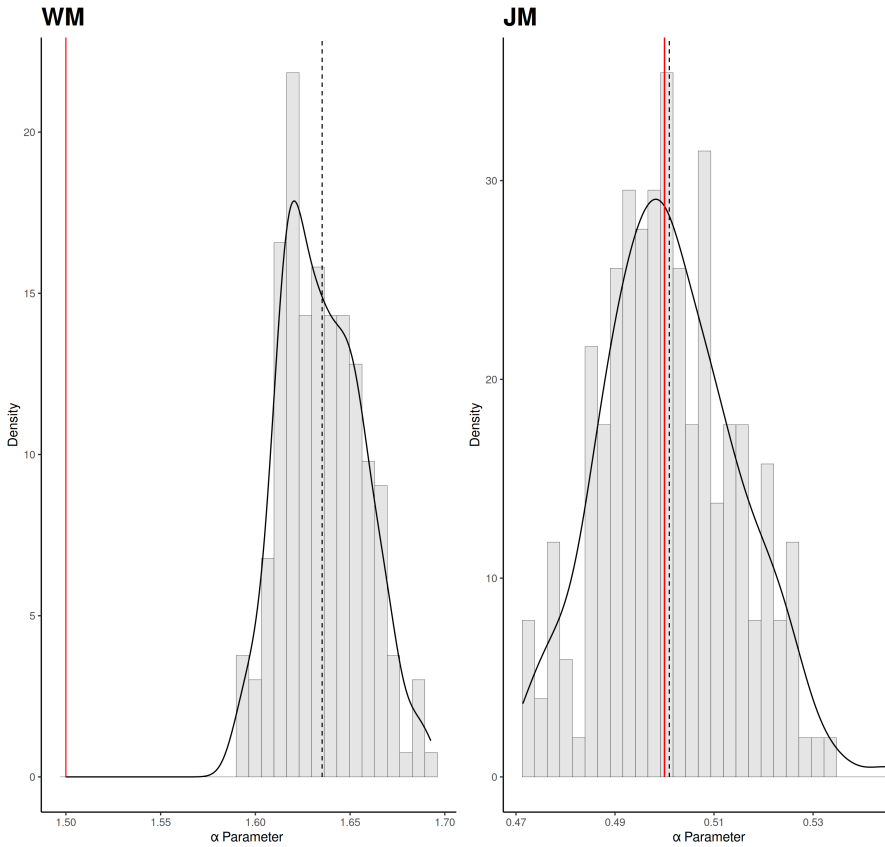
Strong Association

The last DGP sets a strong association between the longitudinal and the dropout mechanisms. Figure III.11 shows the results under this DGP specification. The evidence suggests a different depiction than the previous findings. Indeed, when the informative dropout is relatively strong, the LMM shows a slight upward bias in detecting the cubic shape of the longitudinal pattern, while the JM presents no bias at all.

Figure III.11: Distribution of simulated cubic term βt^3 estimated by the LMM an the JM and the α parameter by the WM and JM.



(a) Distribution Parameter βt^3



(b) Distribution Parameter α

When we consider the α parameter, the findings suggest an exacerbation of the upward bias of the WM estimator, while the JM is capable of correctly detecting the parameter set in the Monte Carlo simulation.

III.4 Discussion & Conclusions

In the statistical literature, growing attention focused on the threats introduced by missing data to the causal estimation of phenomena measured over time. Specifically, the efforts of statisticians, medical and social scientists regard proposing statistical models devoted to dealing with missing data that are informative of an underlying longitudinal process, which might affect correct survey data analysis. This striving effort among scholars of different disciplines is due to informative dropout's endogeneity and sample homogenization bias. This paper assessed the potential benefits and limitations of a relatively new statistical model, the Joint Modeling (JM) approach. The JM approach uses a two-stage estimation method to accommodate for informative dropout. The approach simultaneously estimates a Linear Mixed Model and survival analysis to account for the dropout mechanism. The idea of a two-stage approach to deal with informative missing is not new in statistics (see Little 1995) and in the social sciences (Heckman, 1979). This study has compared the JM approach with the traditional linear mixed model (LMM) and the Weibull regression model (WM). We have tested these statistical models in two main scenarios, constructed through Monte Carlo techniques: endogeneity (via omitted variable bias) and time specification of the longitudinal pattern. The data generating process of the Monte Carlo study provides three DGPs that are common to each scenario. The first DGP assumes no association between the longitudinal and the dropout mechanisms. The second DGP allows for a moderate association between the two simulated mechanisms. Finally, the third DGP provides a strong association between the longitudinal phenomenon and the dropout from the study. The unobserved heterogeneity scenario included three covariates: one exogenous and one endogenous time-varying covariate and one time-invariant dichotomous factor. We then omitted the time-varying covariate from the less-than-optimal models. In the time specification scenario, we created a longitudinal pattern following a cubic spline approximation to see the models' behavior when dealing with more complex phenomena potentially found in the real-world data analysis. The results suggest that the LMM provides robust estimations of true differences set with the dichotomous covariate from the first scenario. The omitted variable bias introduced does not significantly affect the estimation. The findings suggest that the JM approach and the LMM return

unbiased and robust estimates. These findings are quite surprising as these findings are in contrast with Touloumi et al. (1999). However, a more recent study from Stolz et al. (2018) confirms the similarity between the LMM and the JM. In this sense, this is coherent with the estimation strategy of the JM, which indeed uses a linear mixed model to detect the longitudinal pattern and correct it through the analysis of the dropout rate with a survival model. The findings clearly and constantly favor the JM regarding the α parameter. Indeed, in some cases, the Weibull regression returned biased estimations even n the case of no omitted variable bias. These findings are in line with the previous biomedical literature, and it generally discourages the use of survival analysis with time-varying covariates (see Prentice 1982 for a formal analysis). In the second scenario, the findings depict a similar overview of the models' behavior as we have found with the first scenario. Indeed, the LMM and the JM can correctly model the more complex shape of the longitudinal outcome variable. However, when the association between the longitudinal pattern and the informative dropout is strong (i.e., with $\alpha = 0.5$), the findings suggest a slight upward bias of the cubic term for what concerns the LMM, while the JM presents unbiased estimations. When comparing the WM and the JM, the evidence highlights the JM's capability to detect the proper association parameter better than the WM. Indeed, as the association becomes more robust, the bias of the WM estimations becomes wider. This study has several implications for applied social scientists. The first implication concerns the comparison between the LMM and the JM. We did not find good reasons to discourage using the LMM, which proved to be a reliable and consistent statistical model. However, with this study, we hope to have highlighted social scientists of the potential threats of informative dropout to provide a starting point for a methodological discussion. One point favoring the JM over the LMM is detecting more complex longitudinal patterns. Indeed, only a few phenomena follow linear trends but nonlinear and more complex patterns. The second implication concerns the comparison between the WM and the JM. It is of special interest to social scientists in the field of education (i.e., the dropout rate from the higher educational system, as an example), medical sociologists, and demographers (especially for those dealing with datasets presenting high mortality rates), and economists (e.g., scholars interested in unemployment spells). From the findings we showed earlier, it is clear (and consistent with the biomedical literature) that informative dropout heavily biases the naive time-varying survival regression model, returning inconsistent estimates. Our results suggest considering informative dropout when using survival regression models.

We would like to address the limitations of our study as input for further methodological research on this model. First, we did not test the model with

simulated discrete-time survival data by using age as the timescale. This approach could be of particular interest for sociologists and demographers evaluating life-expectancy inequalities. Second, our study simulated mortality rates using parametric Weibull distribution; an appealing alternative for further researchers could be to use dropout rates from real-world databases. Third, we considered only one specification of the association parameter: the current value association. The biostatistical literature developed alternative estimation methods to assess the association between the longitudinal outcome variable and dropout mechanisms. We address this limitation as a suggestion for comparing the efficiency, in particular sets, of these estimation methods. We conclude that the JM model can be a valuable resource for social scientific research. For example, the JM approach is beneficial to sociological fields in which the focus is on dynamics of social change and the life course (which typically deal with complex longitudinal patterns). The JM approach could be even more beneficial for social scientists whose field uses time-to-event regression for panel data, where attrition and MNAR are rampant.

Notes

[14]Rubin, 1976 and Little, 1995 classify three main dropout mechanisms which are well established in the statistical literature.

[15]Ibrahim, Chu, and Chen (2010) and Asar et al. (2015) provide an accessible introduction to the Joint Modeling approach.

[16]The JM approach, among other fields, has been used recently in aging research (Arbeev et al. 2014; van den Hout and Muniz-Terrera 2016)

[17]The survival submodel can be either a parametric (Weibull, Gompertz, or exponential) baseline hazard, semi-parametric or flexible parametric (see Royston and Parmar 2002)

[18]In this sense, equation III.1 shows the different assumption with the Two-Stage Model: $m_i = y_{it}$ while in the JM y_{it} is the proxy to reconstruct the longitudinal history of m_i.

[19]Nowadays statistical software such as R and Stata offer different packages to implement this model (Crowther, Abrams, and Lambert 2013) for Stata; see Rizopoulos 2011) for an application in R).

Bibliography

Abbring, J. H., & Van Den Berg, G. J. (2007). The unobserved heterogeneity distribution in duration analysis. *Biometrika*, *94*(1), 87–99.

Arbeev, K. G., Akushevich, I., Kulminski, A. M., Ukraintseva, S. V., & Yashin, A. I. (2014). Joint analyses of longitudinal and time-to-event data in research on aging: Implications for predicting health and survival.

Asar, Ö., Ritchie, J., Kalra, P. A., & Diggle, P. J. (2015). Joint modelling of repeated measurement and time-to-event data: An introductory tutorial. *International Journal of Epidemiology*, *44*(1), 334–344.

Balan, T. A., & Putter, H. (2019). Nonproportional hazards and unobserved heterogeneity in clustered survival data: When can we tell the difference? *Statistics in Medicine*, *38*(18), 3405–3420.

Billingham, L. J., & Abrams, K. R. (2002). Simultaneous analysis of quality of life and survival data. *Statistical Methods in Medical Research*, *11*(1), 25–48.

Blossfeld, H. P., & Hamerle, A. (1989). *Unobserved heterogeneity in hazard rate models: a test and an illustration from a study of career mobility* (tech. rep. No. 2).

Burton, A., Altman, D. G., Royston, P., & Holder, R. L. (2006). The design of simulation studies in medical statistics. *Statistics in Medicine*, *25*(24), 4279–4292.

Chi, Y. Y., & Ibrahim, J. G. (2006). Joint models for multivariate longitudinal and multivariate survival data. *Biometrics*, *62*(2), 432–445.

Cremers, J., Mortensen, L. H., & Ekstrøm, C. T. (2021). A Joint Model for Longitudinal and Time-to-event Data in Social and Life Course Research: Employment Status and Time to Retirement. *Sociological Methods and Research*.

Crowther, M. J., Abrams, K. R., & Lambert, P. C. (2013). Joint modeling of longitudinal and survival data. *Stata Journal*, *13*(1), 165–184.

Crowther, M. J., Andersson, T. M., Lambert, P. C., Abrams, K. R., & 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.

Crowther, M. J., & Lambert, P. C. (2013). Simulating biologically plausible complex survival data. *Statistics in Medicine*, *32*(23), 4118–4134.

Dempster, A., Laird, N., & Rubin, D. (1977). *Maximum Likelihood from Incomplete Data via the EM Algorithm* (Vol. 39).

Diggle, P., & Kenward, M. G. (1994). Informative Drop-Out in Longitudinal Data Analysis. *Applied Statistics*, *43*(1), 49.

Halaby, C. N. (2004). Panel models in sociological research: Theory into practice. *Annual Review of Sociology*, *30*(Villareal 2002), 507–544.

Haviland, A. M., Jones, B. L., & Nagin, D. S. (2011). Group-based trajectory modeling extended to account for nonrandom participant attrition. *Sociological Methods and Research*, *40*(2), 367–390.

Heckman, J. J. (1979). Sample Selection Bias as a Specification Error. *Econometrica*, *47*(1), 153–161.

Henderson, R., & Oman, P. (1999). Effect of frailty on marginal regression estimates in survival analysis. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, *61*(2), 367–379.

Hu, W., Li, G., & Li, N. (2009). A Bayesian approach to joint analysis of longitudinal measurements and competing risks failure time data. *Statistics in Medicine*, *28*(April), 1601–1619.

Huang, X., & Liu, L. (2007). A joint frailty model for survival and gap times between recurrent events. *Biometrics*, *63*(2), 389–397.

Ibrahim, J. G., Chu, H., & Chen, L. M. (2010). Basic concepts and methods for joint models of longitudinal and survival data. *Journal of Clinical Oncology*, *28*(16), 2796–2801.

Laird, N. M. (1988). Missing Data in Longitudinal Studies. *Statistics in Medicine*, *7*, 305–315.

Li, G., Lesperance, M., & Wu, Z. (2020). Joint Modeling of Multivariate Survival Data With an Application to Retirement. *Sociological Methods and Research*, 1–27.

Little, R. J. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, *90*(431), 1112–1121.

Liu, X. (2013). Survival Models on Unobserved Heterogeneity and their Applications in Analyzing Large-scale Survey Data. *Journal of Biometrics & Biostatistics*, *05*(02).

Liu, X., Engel, C. C., Kang, H., & Gore, K. L. (2010). Reducing selection bias in analyzing longitudinal health data with high mortality rates. *Journal of Modern Applied Statistical Methods*, *9*(2), 403–413.

Lugtig, P. (2014). Panel Attrition: Separating Stayers, Fast Attriters, Gradual Attriters, and Lurkers. *Sociological Methods & Research*, *43*(4), 699–723.

Marini, M. M., Olsen, A. R., & Rubin, D. B. (1980). Maximum-Likelihood Estimation in Panel Studies with Missing Data. *Sociological Methodology*, *11*(1980), 314.

Papageorgiou, G., Mauff, K., Tomer, A., & Rizopoulos, D. (2019). An overview of joint modeling of time-to-event and longitudinal outcomes. *Annual Review of Statistics and Its Application*, *6*(August 2018), 223–240.

Prentice, R. (1982). Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika*, 331–342.

Proust-Lima, C., Séne, M., Taylor, J. M., & Jacqmin-Gadda, H. (2014). Joint latent class models for longitudinal and time-to-event data: A review. *Statistical Methods in Medical Research*, *23*(1), 74–90.

Rizopoulos, D. (2011). Dynamic Predictions and Prospective Accuracy in Joint Models for Longitudinal and Time-to-Event Data. *Biometrics*, *67*(3), 819–829.

Royston, P., & Parmar, M. K. (2002). Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine*, *21*(15), 2175–2197.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592.

Schifeling, T. A., Cheng, C., Reiter, J. P., & Hillygus, D. S. (2015). Accounting for Nonignorable Unit Nonresponse and Attrition in Panel Studies with Refreshment Samples. *Journal of Survey Statistics and Methodology*, *3*, 265–295.

Schmoor, C., & Schumacher, M. (1997). Effects of covariate omission and categorization when analysing randomized trials with the Cox model. *Statistics in Medicine*, *16*(1-3), 225–237.

Schumacher, M., Olschewski, M., & Schmoor, C. (1987). The impact of heterogeneity on the comparison of survival times. *Statistics in Medicine*, *6*(7), 773–784.

Stolz, E., Mayerl, H., Rásky, V., & Freidl, W. (2018). Does Sample Attrition Affect the Assessment of Frailty Trajectories among Older Adults? A Joint Model Approach. *Gerontology*, *64*(5), 430–439.

Thiébaut, A. C., & 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.

Touloumi, G., Pocock, S. J., Babiker, A. G., & Darbyshire, J. H. (1999). Estimation and comparison of rates of change in longitudinal studies with informative drop-outs. *Statistics in Medicine*, *18*(10), 1215–1233.

Trappmann, M., Gramlich, T., & Mosthaf, A. (2015). The effect of events between waves on panel attrition. *Survey Research Methods*, *9*(1), 31–43.

Tsiatis, A. A. ., & Davidian, M. (2004). Joint modelling of longitudinal and time-to-event data : An Overview. *Statistica Sinica*, *14*(3), 809–834.

van den Hout, A., & 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.

Vandecasteele, L., & Debels, A. (2007). Attrition in panel data: The effectiveness of weighting. *European Sociological Review*, *23*(1), 81–97.

Viviani, S., Alfó, M., & Rizopoulos, D. (2014). Generalized linear mixed joint model for longitudinal and survival outcomes. *Statistics and Computing*, *24*(3), 417–427.

Wang, Y., & Taylor, J. M. (2001). Jointly modeling longitudinal and event time data with application to acquired immunodeficiency syndrome. *Journal of the American Statistical Association*, *96*(455), 895–905.

White, I. R. (2010). Simsum: Analyses of simulation studies including Monte Carlo error. *Stata Journal*, *10*(3), 369–385.

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

Xu, J., & Zeger, S. L. (2001). Joint analysis of longitudinal data comprising repeated measures and times to events. *Journal of the Royal Statistical Society. Series C: Applied Statistics*, *50*(3), 375–387.

Zheng, H. (2020). Unobserved Population Heterogeneity and Dynamics of Health Disparities. *Demographic Research*, *43*, 1009–1048.

Conclusions

4. Concluding Remarks

4.1 Key Findings and Implications

This Ph.D. thesis aimed to contribute to our understanding of the interplay between social stratification, mobility, and health inequalities. Chapter I addressed one of the underlying mechanisms that shape the social gradient of health, highlighting how individuals' social conditions get under the skin and affect mortality risks due to cardiovascular disease. Chapter I tried to examine the disparities in health between individuals with different socioeconomic statuses and differentials among individuals related to differentials at the same level of socioeconomic status. Chapters II and III provided methodological insights to advance our current knowledge on potential threats and limitations to statistical inference in empirical health and social research. Specifically, Chapter II tackled the identification problem, which hampers the capability of empirical models to correctly identify the unique effects of individuals' social origin, destination, and mobility. To address the identification problem, I tested the Diagonal Reference Model (DRM), a statistical model developed with the specific intent to overcome the identification problem. Chapter III engaged the methodological issue of informative dropout due to selective mortality, affecting statistical inference based on longitudinal data. Chapter III illustrates the potential advantages and the limitations of the Joint Modeling (JM) approach, comparing it with the Linear Mixed Effect model and the Weibull regression model.

This Ph.D. thesis examines individuals' social conditions and health outcomes in two mainframes. The first focuses on the association between social stratification and mortality risks. Here, social stratification refers to differentials between individuals in access to relevant resources that increase opportunities to pursue and achieve a healthy life due to their social position. Chapter I draws on this framework, and it assumes no change over time, as it measures individuals' socioeconomic status at a fixed point in time and space. The second framework allows for a temporal change in social positions, namely individuals' social mobility and health outcomes. In Chapters II and III, I

focused instead on the methodological issues that empirical researchers might face when the analytical strategy considers social mobility as an explanatory factor and when the empirical analysis implies the use of longitudinal data. These two frameworks have important implications, relevant for both academics and policymakers to implement more effective policies to reduce inequalities.

The first implication regards the differentials between individuals at the same level of socioeconomic status. Chapter I uncovers inequalities in mortality risks due to cardiovascular disease even if individuals belong to the same social position. This result indicates that other factors play a role in shaping health inequalities, revealing the complexity of the social gradient of health. These factors can mediate the relationship between socioeconomic conditions and mortality risks and belong to the spheres of social, mental, behavioral, or material resources that might influence the health outcomes. Therefore, Chapter I highlights that belonging to the same social position does not give individuals the same chances to access resources. Even more, early life social conditions experienced by individuals could significantly impact the availability and accumulation of these resources. The second implication of Chapter I concerns using biological data as an objective measure of health. Chapter I highlights the interplay between the social environment, the effects on physiological functioning, and health inequalities. Thus, results suggest a clear need to develop further our knowledge of how social conditions affect physiological reactions, as they have a remarkable influence on individuals' health and health inequalities.

The third implication regards the inclusion of social mobility as a critical factor to capture the dynamic influence of socioeconomic position on health. The identification problem, which hinders the correct estimation of the specific effects of individuals' social origin, destination, and mobility, is still unresolved. Thus, Chapter II provides the first step to revitalizing the methodological discussion on the identification problem. Given the similarity of the underlying problem, age-period-cohort literature has substantially progressed in that sense. Therefore, this study aims to provide potentially helpful input to facilitate further contributions that address the problem.

The last implication regards the threats to statistical inference due to informative dropout. Chapter III highlighted that the Joint Modeling approach is a valid method to account for nonignorable dropouts. Although the Linear Mixed model and the Joint Modeling perform quite similarly in the scenarios that constitute the statistical experiment, the Joint Modeling shows

a slight improvement over the Linear Mixed model in complex, nonlinear, longitudinal outcomes of interest. However, the Linear Mixed model is still robust against nonignorable dropout. Conversely, using the time-to-event regression framework comes with disadvantages. The first disadvantage of the time-to-event regression framework is that it is hard to assume that missing data are completely at random or ignorable. Unfortunately, no statistical procedure can assess whether the missing data are ignorable or not ignorable. Therefore, the Joint Modeling approach provides a valuable tool to handle nonignorable missing data, thereby avoiding the threats of biased inference. Secondly, even if the dropout rate has been simulated not to be endogenous to the longitudinal outcome of interest, the time-to-event regression model performed remarkably worse than the Joint Modeling approach. In the two different degrees of endogeneity (i.e., the association between the longitudinal outcome and the dropout rate), the Weibull regression model failed to infer the true estimates correctly. Therefore, my study suggests that the Joint Modeling approach is a reliable tool for analyzing panel data in the social sciences. The Joint Modelling approach is advantageous in the case the focus of the research relies on time-to-event panel data, but it comes with computational drawbacks. Indeed, the Joint Modeling approach is computationally demanding, particularly for large datasets.

4.2 Contributions to the Literature

Ultimately, this Ph.D. thesis contributes to the literature by providing innovative methodological tools to understand better the complexities inherent to health inequalities in four different ways. Chapter I's main contributions to the literature regard the innovative use of Bayesian models to uncover understated aspects of health inequalities and the innovative use of a biological marker as an objective health measure. Conceivably, the Bayesian framework (and the distributional model in particular) enhanced the range of the analysis by acknowledging the contextual and compositional aspects that affect inequalities in mortality risks among individuals. In this regard, Chapter I provides an alternative analytical strategy using the Bayesian framework to examine social stratification, mobility, and health inequalities. Chapter I contributes to the social scientific community regarding the measurement method of health outcomes. Indeed, the use of biomarkers as an objective measure of health can provide thorough information on how social conditions affect health outcomes without the issue of the inherent subjectivity underlying self-rated health measures.

The primary intention of Chapter II is to recrudesce the methodological discussion around the identification problem and the Diagonal Reference Model. Chapter II's contributions to further understanding the characteristics of the Diagonal Reference Model relate to the methodological evaluation of the model. That means the social scientific community might benefit from acquiring an insight into the capabilities of the Diagonal Reference Model to infer the effect of mobility on a particular outcome correctly(i.e., the degree of unbiasedness). Chapter II provides further understanding by providing evidence on the model's capability to detect statistically significant effects by including an analysis of the Empirical Coverage Rate.

From an empirical and substantial perspective, this Ph.D. thesis progresses by proposing another statistical framework, the Joint Modeling approach presented in Chapter III. With the Joint Modeling approach proposal, this dissertation contributes to the literature focused on the issue of nonignorable dropouts and puts it to the attention of empirical researchers. More advanced methodological tools are vital to capitalize on panel data's advantages to causal analysis fully.
The four contributions I exposed provide the social scientific community with alternative analytical strategies to tackle the potential pitfalls in the empirical analysis of social inequalities in health.

4.3 Limitations of the Studies

In Chapter I, an aspect that bounded the analysis regards the type of data at the basis of the analysis. Indeed, the analytical strategy in the first Chapter relies on cross-sectional data, which limits the analysis to the depiction of a static frame and thereby avoids the evolutionary aspect of health inequalities. Another limitation concerns the measurement of individuals' social conditions. Indeed, the analytical strategy deployed through the Bayesian framework has not considered early-life conditions. From a methodological perspective, the first Chapter draws only one (informative) prior distribution for each parameter, used to build up the Bayesian regressions.

The limitations of Chapter II rely on the set of the statistical experiment. The parameters utilized for the generation of the data at the basis of the Monte Carlo simulation do not rely on previous analysis of real-world data. Instead, they are fictitious. Thus, arbitrary values create a discrepancy between the real world and the simulated dataset, limiting the external validity of the results. Another limitation of the statistical experiment presented in

Chapter II concerns the measurement of social mobility. Indeed, the analytical strategy simulated two dichotomous variables, indicating whether upward, downward, or immobility occurred between the simulated cases. The dichotomous codification might influence the results of the statistical experiment.

Likewise, the limitations of Chapter III regard the construction of the experimental set, specifically for what concerns the modeling of the dropout phenomenon. Chapter II assumed that the dropout rate would be a continuous, Weibull distribution of events occurring (e.g., dropouts from the simulated study). The dropout occurrence - assumed to be Weibull distributed - admitted only one permanent exit from the study. However, the simulation study presented in Chapter II has not considered standard methodological procedures in survey design, such as sample refreshment or the phenomenon of intermittent dropout. Chapter III does not cover the use of age as the timescale, which is a common strategy in public health, demography, and sociology. Another limitation that bounds the experimental set concerns the simulation of only one pattern of dropout. Chapter III's limitation to the study's external validity concerns the dropout process simulation with fictitious true parameters rather than from previous real-world analysis.

4.4 Suggestions for Further Research

Further research might apply the Bayesian regression model and the distributional model to longitudinal data to expand our current knowledge on the social gradient of mortality risks due to cardiovascular disease. This aspect is particularly important to uncover how the compositional effects related to individuals at the same level of socioeconomic status change over time. Ideally, further research may use the results presented in Chapter I to refine better the prior distribution of the Bayesian model, an essential component of the Bayesian inference.

Further research that intends to expand our methodological knowledge about the behavior of the Diagonal Reference Model could base the Data Generating Process on previous analyses of real-world data. More realistic starting values could potentially expand the external validity of the experiment. Additionally, further research should develop statistical experiments that explore the degree of identification problem with a full-range mobility measure rather than indicator variables for upward and downward mobility.

To improve our understanding and contribute to the methodological lit-

erature on missing data, suggestions for further research should focus on analyzing different patterns of dropout rates. For instance, further research might include dropout rate under refreshment sample, intermittent dropout (such as exit and reentry to the study), or different severity of dropout. A thorough analysis of the potential missing data patterns would further help empirical researchers fully consider and appreciate threats to statistical inference.

This Ph.D. thesis benefited from an interdisciplinary perspective. Previous biomedical and public health knowledge has been essential to defining the biological markers for socially-patterned diseases. The methodological advances in biostatistics with the Joint Modelling approach are promising and potentially helpful in understanding how health inequalities change over time. The methodological insight I provided addresses academic research and policymakers to counteract and determine effective policies for the worrying resurgence of health inequalities in developed countries over time. Thus, it will undoubtedly be promising to enhance the interconnections between the different research fields to tackle health inequalities.
Interdisciplinarity is an invaluable resource to the sociological field to understand the dynamics of social mobility and stratification in health. A thorough understanding of the underlying mechanisms that link social position and health disparities is vital because it heavily affects the quality of life and well-being at every life stage.

Appendix

A. Supplementary Materials - Chapter I

Model 1

Figure A.1: Panel (a): Trace plot of the occupational status parameter σ_{β_i}. Panel (b): Autocorrelation plot of σ_{β_i} by number of chain of the MCMC.

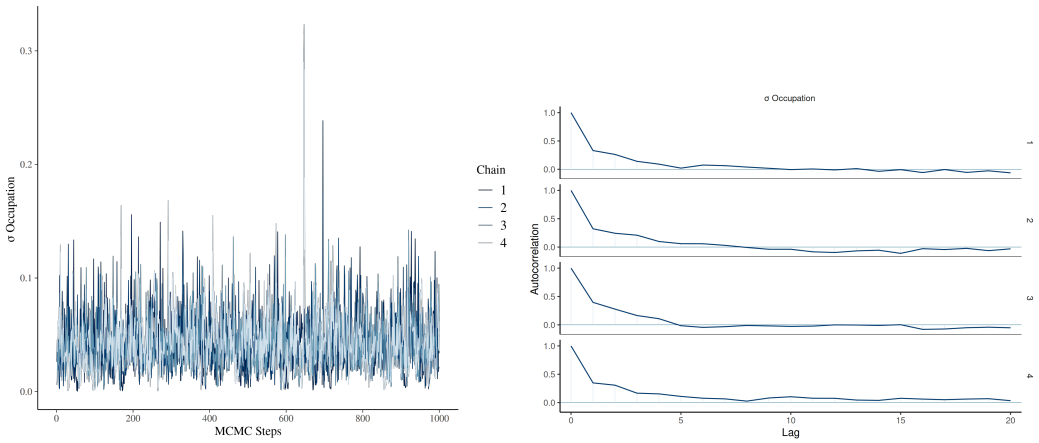


Figure A.2: Panel (a): Trace plot of the educational attainment parameter σ_{β_i}. Panel (b): Autocorrelation plot of σ_{β_i} by number of chain of the MCMC.

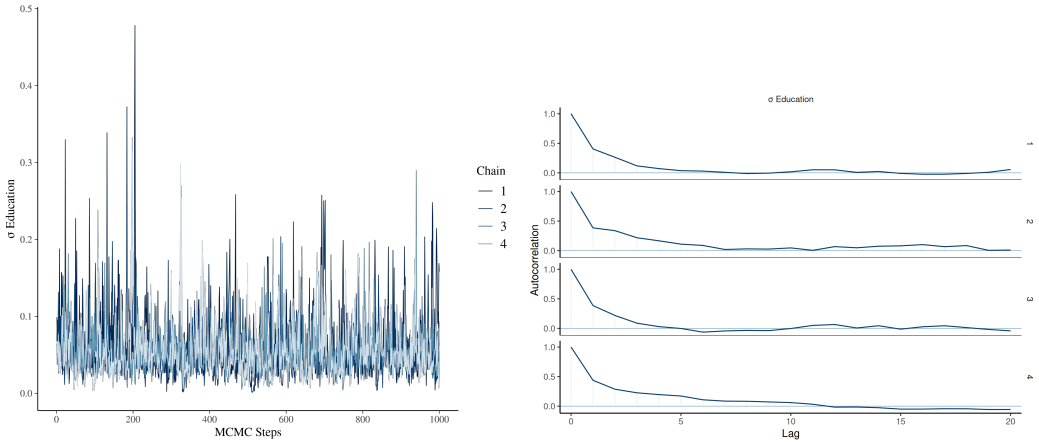
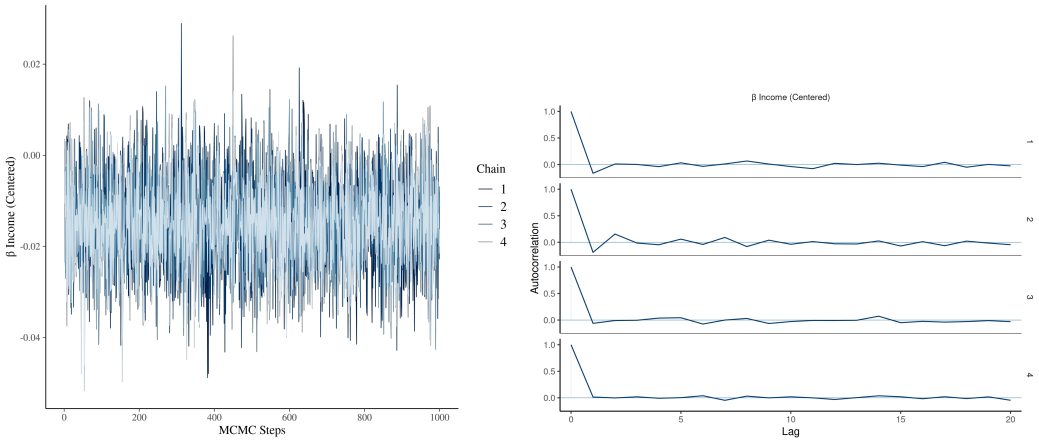


Figure A.3: Panel (a): Trace plot of the β income parameter. Panel (b): Autocorrelation plot of β by number of chain of the MCMC.



Model 2

Figure A.4: Panel (a): Trace plot of the occupational status parameter σ_{β_i} on σ_y. Panel (b): Autocorrelation plot of σ_{β_i} by number of chain of the MCMC.

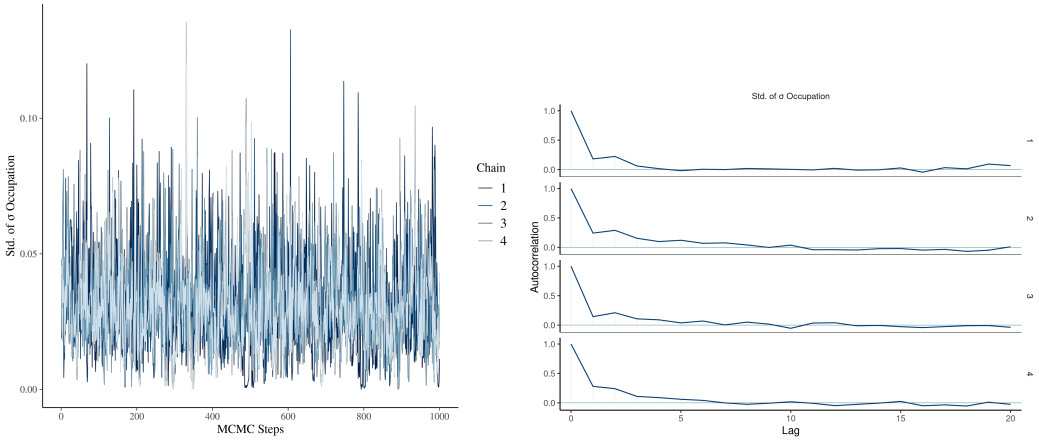


Figure A.5: Panel (a): Trace plot of the educational attainment parameter σ_{β_i} on σ_y. Panel (b): Autocorrelation plot of σ_{β_i} by number of chain of the MCMC.

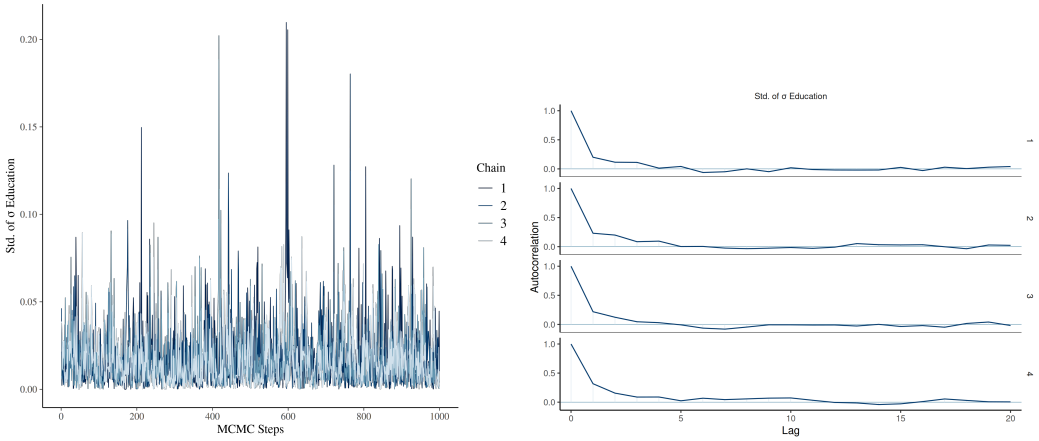
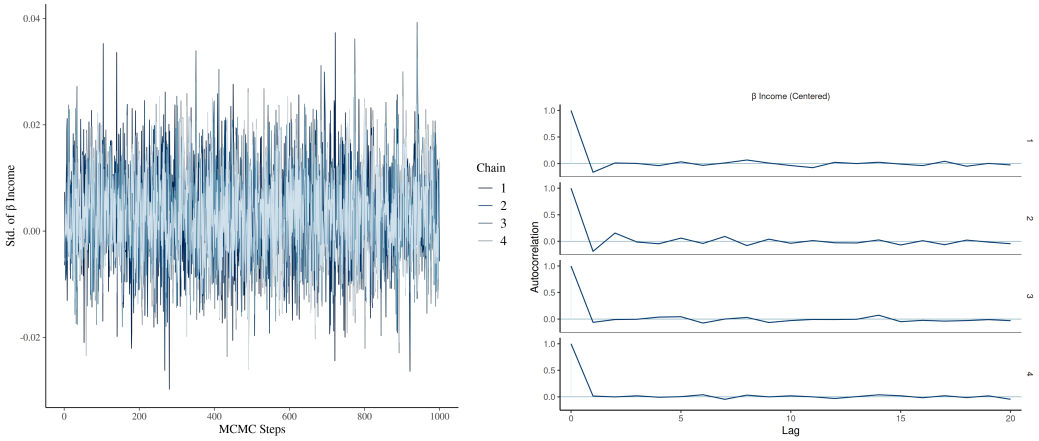


Figure A.6: Panel (a): Trace plot of the β income parameter on σ_y. Panel (b): Autocorrelation plot of β by number of chain of the MCMC.



B. Supplementary Materials - Chapter II

Appendix A - Summary Table of the Standardized Bias and the Empirical Coverage Rate of the Upward and Downward Mobility Coefficients and the Weighting Parameters ρ and $(1-\rho)$ for the Continuous Dependent Variable Scenario.

The columns ρ, r, γ_{Up}, γ_{Down} and N show the true population values. The columns δ_ρ, δ_r, δ_{up}, and δ_{down} show the standardized bias for the population parameters ρ, $(1-\rho)$, and upward and downward mobility coefficients. Columns $(1-\alpha)$Up and $(1-\alpha)Down$ show the ECR relative to the upward and downward mobility coefficients, respectively. To increase the readability of the table, we subdivided the rows so that the salience parameters ρ and r sum up to 1. The first four columns contain the true population parameters that we have set to guide the data-generating process.

Table B.1: Summary Results for the Continuous Dependent Variable Scenario

	ρ	r	γ_{Up}	γ_{Down}	N	δ_ρ	δ_r	δ_{up}	δ_{down}	$(1-\alpha)Up$	$(1-\alpha)Down$
1	0.7	0.3	−0.1	0.1	500	0.02774	−0.02774	−0.01998	0.01931	0.94400	0.95150
2	0.5	0.5	−0.1	0.1	500	0.01621	−0.01621	−0.02289	0.01476	0.95650	0.95050
3	0.3	0.7	−0.1	0.1	500	0.01989	−0.01989	−0.02763	0.00394	0.95650	0.95200
4	0.7	0.3	−0.5	0.1	500	0.00281	−0.00281	0.01108	−0.00336	0.94850	0.94900
5	0.5	0.5	−0.5	0.1	500	0.01264	−0.01264	−0.01098	0.01325	0.95250	0.95100
6	0.3	0.7	−0.5	0.1	500	−0.01230	0.01230	−0.00560	−0.00760	0.95650	0.93850
7	0.7	0.3	−0.1	0.5	500	−0.02278	0.02278	0.02449	−0.02263	0.93900	0.94400
8	0.5	0.5	−0.1	0.5	500	−0.00356	0.00356	0.00810	0.01720	0.95100	0.94450
9	0.3	0.7	−0.1	0.5	500	−0.00113	0.00113	−0.02311	−0.00680	0.94200	0.95150
10	0.7	0.3	−0.5	0.5	500	0.01454	−0.01454	−0.00845	−0.00905	0.94400	0.94450
11	0.5	0.5	−0.5	0.5	500	0.01622	−0.01622	−0.01287	−0.01041	0.95000	0.94300
12	0.3	0.7	−0.5	0.5	500	0.01085	−0.01085	−0.02006	−0.00438	0.94350	0.95000
13	0.7	0.3	−0.1	0.1	750	−0.00124	0.00124	−0.01223	−0.02100	0.94750	0.94650
14	0.5	0.5	−0.1	0.1	750	0.00367	−0.00367	−0.02641	0.00102	0.94900	0.94950
15	0.3	0.7	−0.1	0.1	750	−0.00652	0.00652	−0.00961	0.00843	0.93750	0.94650
16	0.7	0.3	−0.5	0.1	750	−0.00341	0.00341	−0.01929	−0.00582	0.94900	0.94700
17	0.5	0.5	−0.5	0.1	750	0.00082	−0.00082	−0.00848	0.01942	0.94700	0.94750
18	0.3	0.7	−0.5	0.1	750	0.02182	−0.02182	−0.02805	0.01622	0.94300	0.94700
19	0.7	0.3	−0.1	0.5	750	0.01170	−0.01170	−0.02164	0.03033	0.94700	0.94950
20	0.5	0.5	−0.1	0.5	750	−0.01387	0.01387	0.00303	−0.01679	0.95250	0.94950
21	0.3	0.7	−0.1	0.5	750	−0.02744	0.02744	0.02628	−0.04290	0.94700	0.95600
22	0.7	0.3	−0.5	0.5	750	−0.00813	0.00813	−0.00390	−0.02369	0.94650	0.95200
23	0.5	0.5	−0.5	0.5	750	−0.01246	0.01246	0.02768	−0.01811	0.94400	0.94700
24	0.3	0.7	−0.5	0.5	750	−0.00724	0.00724	0.00417	−0.00607	0.95000	0.95150
25	0.7	0.3	−0.1	0.1	1000	−0.02105	0.02105	0.02234	−0.02738	0.95100	0.95800
26	0.5	0.5	−0.1	0.1	1000	0.03539	−0.03539	0.00115	0.04738	0.95700	0.94750
27	0.3	0.7	−0.1	0.1	1000	−0.01023	0.01023	−0.02459	−0.03658	0.94900	0.95750
28	0.7	0.3	−0.5	0.1	1000	0.01425	−0.01425	−0.01218	0.01807	0.94400	0.96150
29	0.5	0.5	−0.5	0.1	1000	−0.00836	0.00836	0.01583	0.00754	0.94600	0.94000
30	0.3	0.7	−0.5	0.1	1000	−0.01300	0.01300	0.01598	−0.00074	0.94550	0.94550
31	0.7	0.3	−0.1	0.5	1000	0.01779	−0.01779	−0.00140	0.02561	0.95200	0.95000

32	0.5	0.5	−0.1	0.5	1000	0.01253	−0.01253	0.01616	0.02180	0.95400	0.95450
33	0.3	0.7	−0.1	0.5	1000	−0.01539	0.01539	0.00792	−0.02631	0.95600	0.94200
34	0.7	0.3	−0.5	0.5	1000	0.01300	−0.01300	−0.00747	0.01349	0.93700	0.94100
35	0.5	0.5	−0.5	0.5	1000	0.00202	−0.00202	−0.01430	−0.01099	0.94000	0.94800
36	0.3	0.7	−0.5	0.5	1000	−0.01426	0.01426	0.00667	−0.00497	0.95300	0.94750

Appendix B - Summary Table of the Standardized Bias and the Empirical Coverage Rate of the Upward and Downward Mobility Coefficients and the Weighting Parameters ρ and $(1-\rho)$ for the Logistic Dependent Variable Scenario.

The columns ρ, r, γ_{Up}, γ_{Down} and N show the true population values. The columns δ_ρ, δ_r, δ_{up}, and δ_{down} show the standardized bias for the population parameters ρ, $(1-\rho)$, and upward and downward mobility coefficients. Columns $(1-\alpha)Up$ and $(1-\alpha)Down$ show the ECR relative to the upward and downward mobility coefficients, respectively. To increase the readability of the table, we subdivided the rows so that the salience parameters ρ and r sum up to 1. The first four columns contain the true population parameters that we have set to guide the data-generating process.

Table B.2: Summary Results for the Logistic Dependent Variable Scenario

	ρ	r	γ_{Up}	γ_{Down}	N	δ_ρ	δ_r	δ_{up}	δ_{down}	$(1-\alpha)Up$	$(1-\alpha)Down$
1	0.7	0.3	−0.1	0.1	500	−0.12426	0.12426	0.08899	−0.12033	0.94200	0.94900
2	0.5	0.5	−0.1	0.1	500	−0.01697	0.01697	−0.00733	−0.01813	0.94300	0.95500
3	0.3	0.7	−0.1	0.1	500	0.15197	−0.15197	−0.11176	0.12797	0.93800	0.94900
4	0.7	0.3	−0.5	0.1	500	−0.13650	0.13650	0.09226	−0.12281	0.93900	0.94800
5	0.5	0.5	−0.5	0.1	500	0.00196	−0.00196	−0.05770	0.02206	0.94550	0.94600
6	0.3	0.7	−0.5	0.1	500	0.17605	−0.17605	−0.18731	0.14244	0.94650	0.93500
7	0.7	0.3	−0.1	0.5	500	−0.14727	0.14727	0.09157	−0.10851	0.94200	0.94300
8	0.5	0.5	−0.1	0.5	500	−0.00606	0.00606	−0.00985	0.00437	0.94750	0.94500
9	0.3	0.7	−0.1	0.5	500	0.17250	−0.17250	−0.17398	0.13030	0.94000	0.94800
10	0.7	0.3	−0.5	0.5	500	−0.15231	0.15231	0.07243	−0.12726	0.94150	0.94200
11	0.5	0.5	−0.5	0.5	500	−0.02940	0.02940	−0.01316	0.00685	0.95600	0.95450
12	0.3	0.7	−0.5	0.5	500	0.13983	−0.13983	−0.15733	0.13777	0.95100	0.94500
13	0.7	0.3	−0.1	0.1	750	−0.07476	0.07476	0.06590	−0.05392	0.93850	0.93600
14	0.5	0.5	−0.1	0.1	750	−0.00936	0.00936	−0.02554	−0.03522	0.94200	0.94450
15	0.3	0.7	−0.1	0.1	750	0.06607	−0.06607	−0.07466	0.06743	0.94400	0.93600
16	0.7	0.3	−0.5	0.1	750	−0.08000	0.08000	−0.01237	−0.08248	0.93300	0.93150
17	0.5	0.5	−0.5	0.1	750	−0.01151	0.01151	−0.04016	−0.01199	0.94450	0.93700
18	0.3	0.7	−0.5	0.1	750	0.11285	−0.11285	−0.13256	0.07806	0.93950	0.92600
19	0.7	0.3	−0.1	0.5	750	−0.11322	0.11322	0.10219	−0.08297	0.93700	0.94750
20	0.5	0.5	−0.1	0.5	750	0.01004	−0.01004	−0.04258	0.02723	0.94300	0.94050
21	0.3	0.7	−0.1	0.5	750	0.06823	−0.06823	−0.03896	0.09521	0.94650	0.94300
22	0.7	0.3	−0.5	0.5	750	−0.03593	0.03593	0.00335	−0.00472	0.94150	0.94800
23	0.5	0.5	−0.5	0.5	750	−0.05449	0.05449	−0.00841	−0.01193	0.94350	0.94050
24	0.3	0.7	−0.5	0.5	750	0.09900	−0.09900	−0.10479	0.09666	0.95150	0.93950
25	0.7	0.3	−0.1	0.1	1000	−0.07458	0.07458	0.02166	−0.07888	0.93350	0.94200
26	0.5	0.5	−0.1	0.1	1000	0.01287	−0.01287	−0.02878	0.01285	0.94550	0.94900
27	0.3	0.7	−0.1	0.1	1000	0.03661	−0.03661	−0.04025	0.02851	0.94950	0.94300
28	0.7	0.3	−0.5	0.1	1000	−0.09584	0.09584	0.07256	−0.05399	0.93600	0.94500
29	0.5	0.5	−0.5	0.1	1000	−0.01313	0.01313	−0.02017	−0.03914	0.93650	0.94100
30	0.3	0.7	−0.5	0.1	1000	0.01219	−0.01219	−0.05022	0.01450	0.95100	0.93950
31	0.7	0.3	−0.1	0.5	1000	−0.00549	0.00549	−0.03228	0.01403	0.93900	0.94300

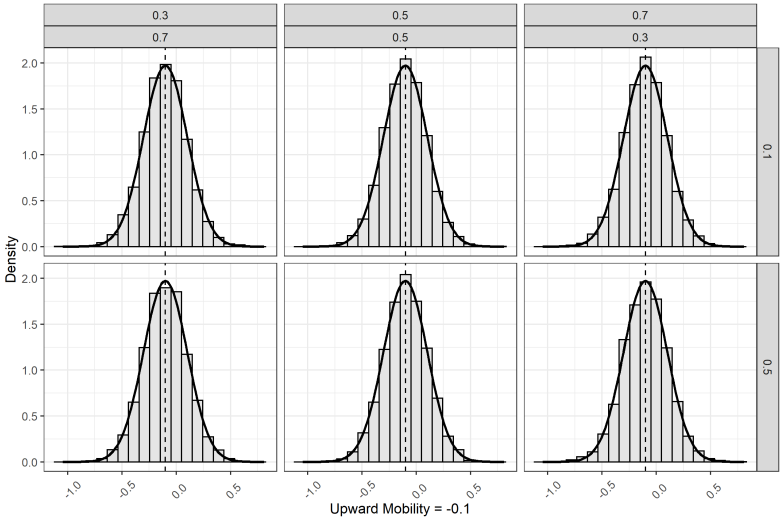
32	0.5	0.5	−0.1	0.5	1000	0.00242	−0.00242	−0.01792	0.00738	0.94300	0.93500
33	0.3	0.7	−0.1	0.5	1000	0.03876	−0.03876	−0.03614	0.05901	0.94250	0.94700
34	0.7	0.3	−0.5	0.5	1000	−0.05648	0.05648	0.02067	−0.03271	0.94800	0.95000
35	0.5	0.5	−0.5	0.5	1000	−0.00107	0.00107	−0.03248	0.02831	0.94100	0.94400
36	0.3	0.7	−0.5	0.5	1000	−0.00646	0.00646	−0.03199	0.02669	0.95250	0.94300

Appendix C: Histogram Distributions of Estimated Upward and Downward Mobility Parameters.

Linear Dependent Variable

Figure B.1: Histogram of the 2,000 simulated estimates of γ_{Up}. Panel a) shows the distribution when $\gamma_{Up} = -0.1$, Panel b) shows the distribution when $\gamma_{Up} = -0.5$

(a) $\gamma_{Up} = -0.1$



(b) $\gamma_{Up} = -0.5$

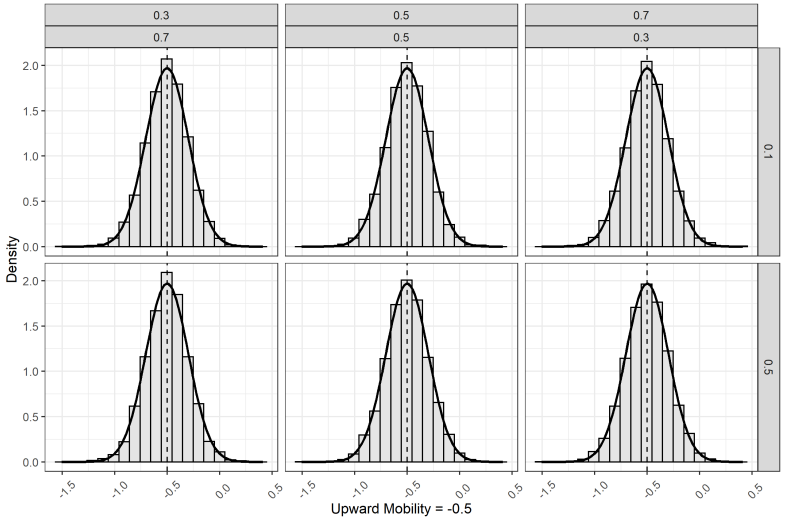
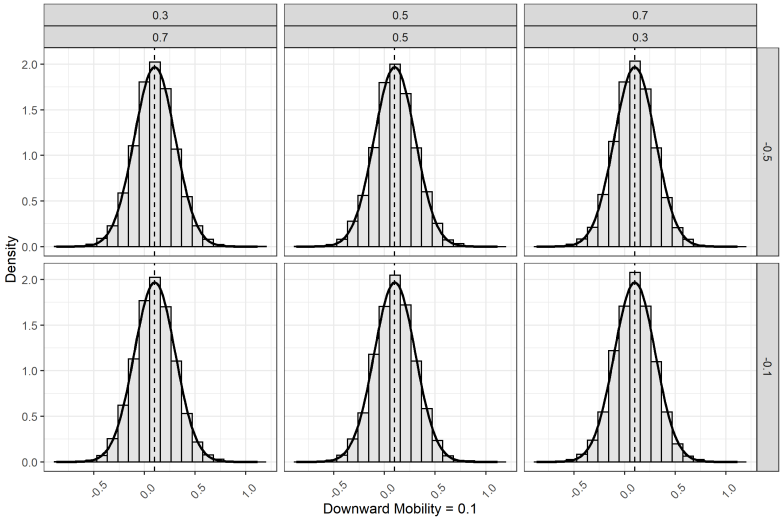
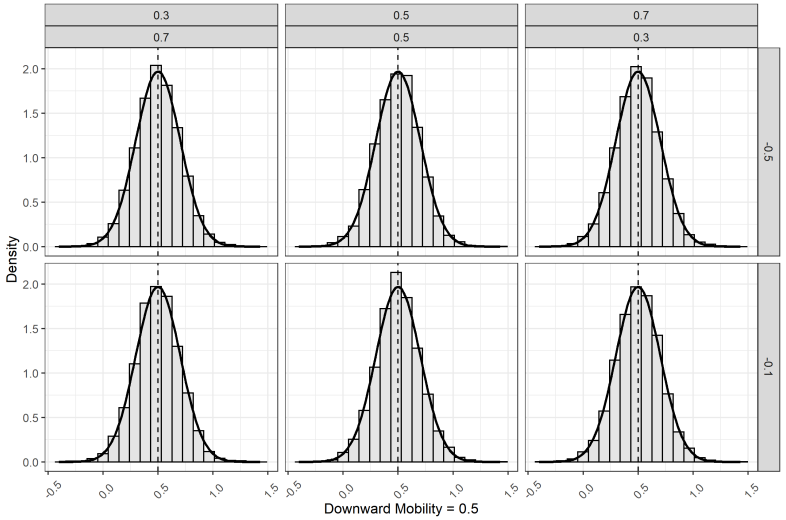


Figure B.2: Histogram of the 2,000 simulated estimates of γ_{Down}. Panel a) shows the distribution when $\gamma_{Down} = 0.1$, Panel b) shows the distribution when $\gamma_{Down} = 0.5$

(a) $\gamma_{Down} = 0.1$



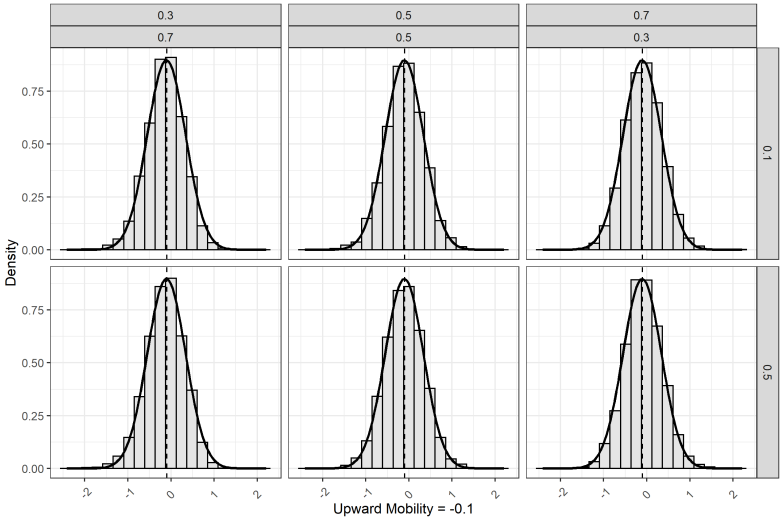
(b) $\gamma_{Down} = 0.5$



Binomial Dependent Variable

Figure B.3: Histogram of the 2,000 simulated estimates of γ_{Up}. Panel a) shows the distribution when $\gamma_{Up} = -0.1$, Panel b) shows the distribution when $\gamma_{Up} = -0.5$

(a) $\gamma_{Up} = -0.1$



(b) $\gamma_{Up} = -0.5$

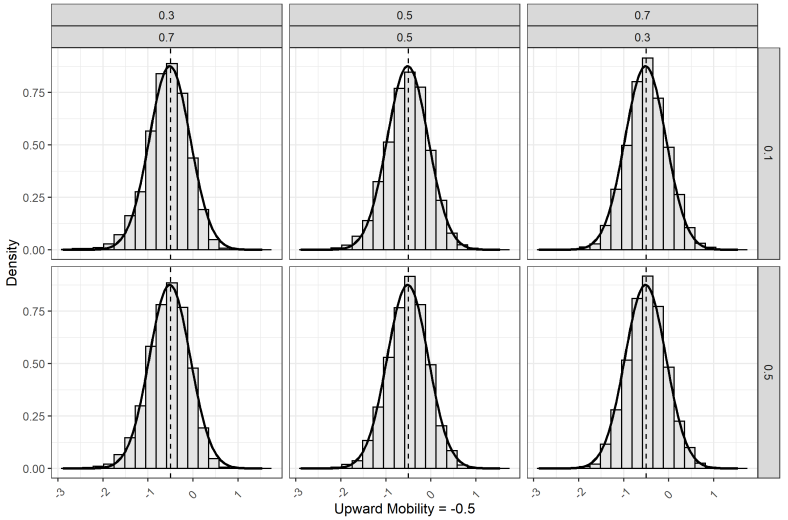
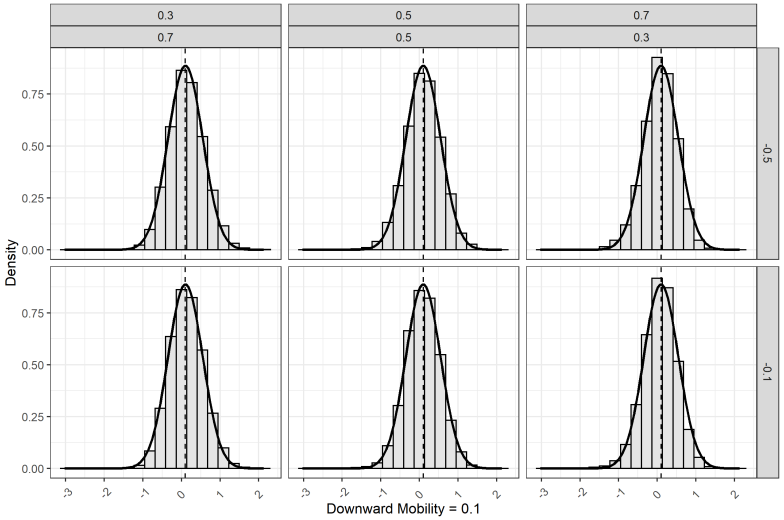
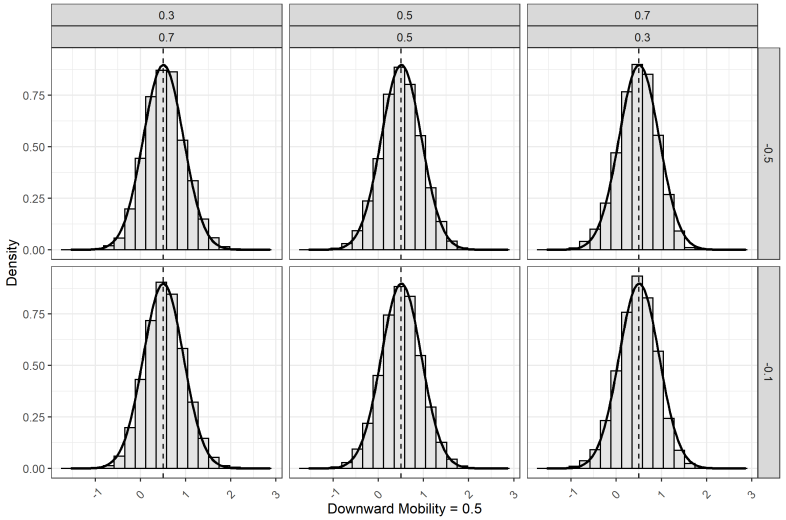


Figure B.4: Histogram of the 2,000 simulated estimates of γ_{Down}. Panel a) shows the distribution when $\gamma_{Down} = 0.1$, Panel b) shows the distribution when $\gamma_{Down} = 0.5$

(a) $\gamma_{Down} = 0.1$



(b) $\gamma_{Down} = 0.5$



C. Supplementary Materials - Chapter III

Appendix A: Review of the Missing Data Processes

In this Appendix we review the statistical concepts behind the missing data processes. The illustration relies on the work of Laird (1988), from which we borrow the notation and the key aspects. In the longitudinal survey design, the units included in the study are observed consequentially at different points in time (Diggle & Kenward, 1994; Laird, 1988; Schifeling et al., 2015). From a statistical perspective, the ideal outcome would be a $T \times 1$ vector for each unit of observation $\mathbf{Y}_i = [Y_{i1}, Y_{i2}, \ldots, Y_{it}]^T$ of interest. The subject-specific outcomes of interest can be combined in a unique vector $\mathbf{Y} = \mathcal{N}\{\mathbf{X}\boldsymbol{\theta}, \Sigma\}$, where $\mathbf{Y}$ can be conceived as multivariate normally distributed (assuming that it is continuous), the mean will be defined by the product of $\mathbf{X}$, a $N \times p$ matrix of explanatory variables by $\boldsymbol{\theta}$, a $1 \times p$ vector of parameters of interest. The variance-covariance matrix will be defined by Σ, which takes the form of $\Sigma = \begin{bmatrix} \sigma_1^2 & & & \\ 0 & \sigma_2^2 & & \\ \vdots & \vdots & \ddots & \\ 0 & 0 & \ldots & \sigma_n^2 \end{bmatrix}$ The function $\mathbf{Y} = \mathcal{N}\{\mathbf{X}\boldsymbol{\theta}, \Sigma\}$ defines then the data model. In order to define the missing data processes we need to define also a *non-response model* that is a statistical model that explain the missing data pattern. An indicator variable $\mathbf{R} \in \{0, 1\}$ can be defined, where 0 indicates if the datum is missing and 1 if it is observed. By using the indicator variable $\mathbf{R}$ we can then decompose the vector $\mathbf{Y} = (\mathbf{Y}_m, \mathbf{Y}_o)$. The non-response model is defined as $\mathbf{R} = f(\mathbf{Z}\boldsymbol{\phi})$, where $\mathbf{Z}$ is a matrix of covariates that influence the missing pattern and $\boldsymbol{\phi}$ is the associated vector of parameters. As Schafer and Graham (2002) noted, $\mathbf{R}$ is a random variable, therefore we need to define its probability density function (p.d.f.), which indicates the distribution of missingness or the probabilities of missingness (Schafer & Graham, 2002, p.

151). The p.d.f. of $\mathbf{R}$ (and the observed data) can thus be defined as:

$$f\big(\mathbf{Y}_o, \mathbf{R} \mid \mathbf{X}, \mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\phi}\big) = \int_R f\big(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\theta}\big)\, f\big(\mathbf{R} \mid \mathbf{Y}, \mathbf{X}, \mathbf{Z}, \boldsymbol{\phi}\big)\, d\mathbf{Y}_m$$

Where the integration has limits defined by $\mathbf{R}$ and then it is over $\mathbf{Y}_m$. The p.d.f. of $\mathbf{R}$ considers the relationship between the function concerning the data model and the same for the non-response model. From the works of Diggle and Kenward (1994), Little and Rubin (1987), and Rubin (1976) we can define three missing processes. Missing data are defined as Completely at Random if $f\big(\mathbf{R} \mid \mathbf{Y}, \mathbf{Z}, \mathbf{X}, \boldsymbol{\phi}\big) = f\big(\mathbf{R} \mid \mathbf{Z}, \mathbf{X}, \boldsymbol{\phi}\big)$. It is important to note that, in the case, Dropout is completely at random, the non-response model is not influenced by $\mathbf{Y}$, but it can be by $\mathbf{Z}$ and $\mathbf{X}$. A typical example of missing completely at random is the scenario in which the investigator decides *a priori* exclude respondents with specific characteristics by determining questions (Marini et al., 1980).
The dropout mechanism can be defined as Random (or ignorable) if $f\big(\mathbf{R} \mid \mathbf{Y}, , \boldsymbol{\phi}\big) = f\big(\mathbf{R} \mid \mathbf{Y}_o, \mathbf{Z}, \mathbf{X}, \boldsymbol{\phi}\big)$. The key difference is that the non-response model depends on $\mathbf{Y}_o$, but not on $\mathbf{Y}_m$. In this case, the non-response mechanism can be ignored. However, it is important to note that $\mathbf{Y}_o$ and $\mathbf{R}$ are not independent. This means that it is allowed to ignore the non-response model, not the missing on the data.
Lastly, informative (non-ignorable) dropout can be defined as $f\big(\mathbf{R} \mid \mathbf{Y}, \mathbf{Z}, \mathbf{X}, \boldsymbol{\phi}\big) = f\big(\mathbf{R} \mid \mathbf{Y}_o, \mathbf{Y}_m, \mathbf{Z}, \mathbf{X}, \boldsymbol{\phi}\big)$. In this scenario, the analytical strategy cannot ignore the non-response model, as endogeneity occurs.

Appendix B: Results in Tabular Format

Unobserved Heterogeneity: No Association

Table C.1: Performance Measure Table of the MC Simulation Assuming No Association between the Longitudinal Outcome and the Dropout Rate.

	UH Not Present					UH Present				
	β_3			α		β_3			α	
Perfmeasnum	LMM	JM	WM	WM	JM	LMM	JM	WM	WM	JM
Bias in point estimate	0.0095 (0.0113)	0.0095 (0.0113)	1.0011 (0.0016)	0.0028 (0.0037)	$7e^{-04}$ (0.0015)	0.0152 (0.0219)	0.0151 (0.0219)	0.0032 (0.0037)	0.9783 (0.0014)	-0.02 (0.0014)
Empirical standard error	0.1596 (0.008)	0.1596 (0.008)	0.0226 (0.0011)	0.0525 (0.0026)	0.021 (0.0011)	0.3083 (0.0155)	0.3083 (0.0155)	0.0527 (0.0027)	0.0199 (0.001)	0.0191 (0.001)
Mean squared error	0.0254 (0.0025)	0.0254 (0.0025)	1.0028 (0.0032)	0.0028 ($5e^{-04}$)	$4e^{-04}$ ($1e^{-04}$)	0.0948 (0.0088)	0.0948 (0.0088)	0.0028 ($5e^{-04}$)	0.9575 (0.0028)	$8e^{-04}$ ($1e^{-04}$)
RMS model-based standard error	0.1522 ($2e^{-04}$)	0.1522 ($2e^{-04}$)	0.0532 (0.0024)	0.0532 (0.0024)	0.1522 ($2e^{-04}$)	0.3227 ($6e^{-04}$)	0.3229 ($6e^{-04}$)	0.0536 (0.0024)	0.0536 (0.0024)	0.3229 ($6e^{-04}$)
Relative % error in standard error	-4.6781 (4.8046)	-4.6234 (4.8074)	135.1353 (15.9337)	1.4179 (6.8725)	623.96 (36.49)	4.6666 (5.2761)	4.7222 (5.2789)	1.7565 (6.8749)	170.0413 (18.2446)	1588.1 (85.094)
Coverage of nominal 95% confidence interval	94.4444 (1.6279)	94.4444 (1.6279)	0 (0)	69.697 (3.266)	100 (0)	95.9596 (1.3993)	95.9596 (1.3993)	69.697 (3.266)	0 (0)	100 (0)

Unobserved Heterogeneity: Moderate Association

Table C.2: Performance Measure Table of the MC Simulation Assuming Moderate Association between the Longitudinal Outcome and the Dropout Rate.

	UH Not Present					UH Present				
	β_3			α		β_3			α	
perfmeasnum	LMM	JM	WM	WM	JM	LMM	JM	WM	WM	JM
Bias in point estimate	−0.0068 (0.0114)	−0.002 (0.0114)	1.302 (0.0012)	0.0017 ($8e^{-04}$)	0.2511 ($9e^{-04}$)	0.0182 (0.0213)	0.0211 (0.0213)	0.0032 (0.0037)	1.1402 ($7e^{-04}$)	0.1301 ($6e^{-04}$)
Empirical standard error	0.1619 (0.0081)	0.1618 (0.0081)	0.0171 ($9e^{-04}$)	0.0107 ($5e^{-04}$)	0.0128 ($6e^{-04}$)	0.3006 (0.0151)	0.3006 (0.0151)	0.0527 (0.0027)	0.0095 ($5e^{-04}$)	0.0082 ($4e^{-04}$)
Mean squared error	0.0261 (0.0028)	0.0261 (0.0028)	1.6955 (0.0032)	$1e^{-04}$ (0)	0.0632 ($5e^{-04}$)	0.0902 (0.0086)	0.0904 (0.0086)	0.0028 ($5e^{-04}$)	1.3001 (0.0015)	0.017 ($2e^{-04}$)
RMS model-based standard error	0.152 ($2e^{-04}$)	0.152 ($2e^{-04}$)	0.0106 ($1e^{-04}$)	0.0106 ($1e^{-04}$)	0.152 ($2e^{-04}$)	0.3221 ($5e^{-04}$)	0.3224 ($5e^{-04}$)	0.0536 (0.0024)	0.0261 ($3e^{-04}$)	0.3224 ($5e^{-04}$)
Relative % error in standard error	−6.112 (4.7083)	−6.0757 (4.7101)	−37.8311 (3.2079)	−0.452 (5.1366)	1091.7 (59.764)	7.1709 (5.3748)	7.2286 (5.3777)	1.7565 (6.8749)	176.0681 (14.09)	3817.6 (196.47)
Coverage of nominal 95% confidence interval	92.5 (1.8625)	93 (1.8042)	0 (0)	95.5 (1.4659)	100 (0)	96 (1.3856)	96 (1.3856)	69.697 (3.266)	0 (0)	100 (0)

Unobserved Heterogeneity: Strong Association

Table C.3: Performance Measure Table of the MC Simulation Assuming Strong Association between the Longitudinal Outcome and the Dropout Rate.

	UH Not Present					UH Present				
	β_3			α		β_3			α	
perfmeasnum	LMM	JM	WM	WM	JM	LMM	JM	WM	WM	JM
Bias in point estimate	−0.0035 (0.0115)	0.0125 (0.0115)	0.0067 ($5e^{-04}$)	1.6633 (0.0018)	0.4975 (0.0012)	−0.0158 (0.023)	−0.0087 (0.023)	0.2678 (0.0027)	1.1704 ($7e^{-04}$)	0.1606 ($6e^{-04}$)
Empirical standard error	0.1627 (0.0082)	0.1621 (0.0081)	0.0077 ($4e^{-04}$)	0.0259 (0.0013)	0.0169 ($8e^{-04}$)	0.3256 (0.0163)	0.3255 (0.0163)	0.0378 (0.0019)	0.0093 ($5e^{-04}$)	0.0081 ($4e^{-04}$)
Mean squared error	0.0264 (0.0029)	0.0263 (0.0029)	$1e^{-04}$ (0)	2.7673 (0.0061)	0.2478 (0.0012)	0.1057 (0.0102)	0.1055 (0.0102)	0.0732 (0.0015)	1.37 (0.0015)	0.0258 ($2e^{-04}$)
RMS model-based standard error	0.1531 ($3e^{-04}$)	0.1527 ($3e^{-04}$)	0.0081 ($1e^{-04}$)	0.0081 ($1e^{-04}$)	0.1527 ($3e^{-04}$)	0.3228 ($6e^{-04}$)	0.323 ($6e^{-04}$)	0.0343 ($3e^{-04}$)	0.0343 ($3e^{-04}$)	0.323 ($6e^{-04}$)
Relative % error in standard error	−5.9334 (4.7177)	−5.7746 (4.7257)	6.145 (5.3867)	−68.6609 (1.5904)	804.6801 (45.372)	−0.857 (4.9727)	−0.776 (4.9768)	−9.3833 (4.5927)	267.35 (18.618)	3900.8 (200.67)
Coverage of nominal 95% confidence interval	92 (1.9183)	92.5 (1.8625)	93.5 (1.7432)	0 (0)	0 (0)	94 (1.6793)	94.5 (1.6121)	0 (0)	0 (0)	100 (0)

Time Specification: No Association

Table C.4: Performance Measure Table of the Time Specification Scenario Assuming No Association between the Longitudinal Outcome and the Dropout Rate.

	β_3		α	
perfmeasnum	LMM	JM	WM	JM
Bias in point estimate	0 (0)	0 (0)	1 ($3e^{04}$)	0 ($3e^{04}$)
Empirical standard error	$4e^{-04}$ (0)	$4e^{-04}$ (0)	0.0049 ($2e^{-04}$)	0.0045 ($2e^{-04}$)
Mean squared error	0 (0)	0 (0)	0.9999 ($7e^{-04}$)	0 (0)
RMS model-based standard error	$4e^{-04}$ (0)	$4e^{-04}$ (0)	0.0044 (0)	0.004 (0)
Relative % error in standard error	−2.1888 (4.904)	−2.2513 (4.9009)	−10.0917 (4.5107)	−10.572 (4.4866)
Coverage of nominal 95% confidence interval	96 (1.3856)	96 (1.3856)	0 (0)	91 (2.0236)

Time Specification: Moderate Association

Table C.5: Performance Measure Table of the Time Specification Scenario Assuming Moderate Association between the Longitudinal Outcome and the Dropout Rate.

	β_3		α	
Perfmeasnum	LMM	JM	WM	JM
Bias in point estimate	$2e^{-04}$ (0)	0 (0)	1.0501 ($8e^{-04}$)	$7e^{-04}$ ($6e^{-04}$)
Empirical standard error	$5e^{-04}$ (0)	$5e^{-04}$ (0)	0.0108 ($5e^{-04}$)	0.0078 ($4e^{-04}$)
Mean squared error	0 (0)	0 (0)	1.1028 (0.0016)	$1e^{-04}$ (0)
RMS model-based standard error	$4e^{-04}$ (0)	$4e^{-04}$ (0)	0.0108 (0)	0.0079 (0)
Relative % error in standard error	−7.6005 (4.6332)	−8.2924 (4.5985)	0.3784 (5.0364)	0.8274 (5.057)
Coverage of nominal 95% confidence interval	89 (2.2125)	90.5 (2.0733)	0 (0)	95.5 (1.4659)

Time Specification: Strong Association

Table C.6: Performance Measure Table of the Time Specification Scenario Assuming Strong Association between the Longitudinal Outcome and the Dropout Rate.

	β_3		α	
Perfmeasnum	LMM	JM	WM	JM
Bias in point estimate	$2e^{-04}$ (0)	0 (0)	1.1352 (0.0015)	0.001 (0.001)
Empirical standard error	$4e^{-04}$ (0)	$4e^{-04}$ (0)	0.0214 (0.0011)	0.0136 ($7e^{-04}$)
Mean squared error	0 (0)	0 (0)	1.2892 (0.0034)	$2e^{-04}$ (0)
RMS model-based standard error	$4e^{-04}$ (0)	$4e^{-04}$ (0)	0.0218 ($1e^{-04}$)	0.0136 (0)
Relative % error in standard error	9.8403 (5.5079)	10.396 (5.5358)	1.7669 (5.1075)	−0.2187 (5.0051)
Coverage of nominal 95% confidence interval	94.5 (1.6121)	98 (0.99)	0 (0)	95.5 (1.4659)