THE FADE-AWAY PHENOMENON OF INITIAL NON-RESPONSE BIAS IN PANEL SURVEYS

A Dissertation submitted in partial fulfillment
of the requirement for the degree

Dr. rer. pol.

to the
Chair of Applied Statistics
School of Business and Economics
Freie Universität Berlin

Submitted by
Mursala Khan
from Mardan, Pakistan

Berlin, 2020

Mursala Khan, *The Fade-away Phenomenon of Initial Non-response Bias in Panel Surveys*,
February, 2020

Supervisors:
1. Prof. Dr. Ulrich Rendtel (Freie Universität Berlin)
2. Prof. Dr. Timo Schmid (Freie Universität Berlin)

Location:
Berlin

Date of defense:
11 February, 2020

Dedication

I would like to dedicate my thesis to
my beloved family and my respectable supervisor

Acknowledgments

I am very thankful to the following people for their great assistance and support during my Ph.D studies. I have written this dissertation while working at the Chair of Applied Statistics, Freie Universität Berlin.

First and foremost, I would like to extend my sincere and deepest gratitude to my supervisor, Prof. Dr. Ulrich Rendtel (Freie Universität Berlin, Germany), for his suggestion to do the Ph.D in a very interesting area and for his support and invaluable guidance throughout my doctoral research. It was the time that I had no idea about "Panel Surveys". He introduced me to the world of "Panel Surveys". Thank you Professor Ulrich Rendtel for giving me an exemplary motivation and all-out support in tough times and constructive advice during this time. Without your support and generosity the completion of this PhD thesis wouldn't have seen the light of the day. I have learned a lot from you and enjoyed my research project with you very much.

The same applies to Prof. Dr. Timo Schmid (Freie Universität Berlin), my co-supervisor, whom I also very much thankful for his guidance and great comments at every facet of my Ph.D project.

Special thanks go to the Higher Education Commission (HEC) of Pakistan for supporting this thesis project.

Further, I am also very thankful to my colleagues and friends, for making my life pleasant in Berlin, and providing their useful suggestions and guidance throughout my Ph.D studies at Freie Universität Berlin, especially my colleagues: Ann-Kristin Kreutzmann, Sören Pannier, Natalia Rojas-Perilla, Nora Würz, Felix Skarke, Paul Walter, Noah Mutai, Alejandra Arias, and the entire faculty and staff at the Chair of Statistics, particularly the members of the Statistical Consulting Unit fu:stat.

I am especially very grateful to my cousin Tufail Khan for his useful suggestions and continuous help throughout the writing of this thesis.

Last but certainly not least, I am very much thankful to my family members, whereby I want to mention my parents in particular. I thank these wonderful people for their continuous support and hearty encouragement throughout this journey. My wife Saira Khan has most closely accompanied me and extended a helping hand while I was working on my dissertation, which is why I am very much thankful to her. Her love, sincerity, care, and friendship have motivated me every day afresh and her support has contributed to the thesis in many respects. Finally, the two little kids who make my life full of happiness and helped me achieve my goal before than what I expected are: my son Hashir Khan and daughter Anaya Khan.

Contents

List of Figures

List of Tables

List of symbols and abbreviations

List of symbols

Y : Dependent variable

X : Matrix of covariates

a_t : Vector of model intercepts

b_t : Vector of regression coefficients

M : Permanent component of the covariates

Z : Transient component of the covariates

κ : Stability of the permanent components of the covariates

ρ : Stability of the transient components of the covariates

σ_m^2 : Variance of the permanent component M

σ_z^2 : Variance of the transient component Z

e : Residual term of the model

σ^2 : Variance of the residual term

V : Permanent component of the residual term

U : Transient component of the residual term

γ : Stability of the permanent components of the residual term

ϕ : Stability of the transient components of the residual term

σ_v^2 : Variance of the permanent component V

σ_u^2 : Variance of the transient component U

ϵ : Error term of the first order auto-regressive model of the transient components of the covariates

σ_ϵ^2 : Variance of the error term ϵ

ξ : Error term of the first order auto-regressive model of the transient components of the residual term

σ_ξ^2 : Variance of the error term ξ

R : Response indicator

α : Initial non-response parameter

β : Initial non-response (selective non-response) parameter

$\alpha^\star$: Non-selective attrition parameter

$\beta^\star$: Selective attrition parameter

$B_{t,com}$: Percent relative bias of the regression estimates of b_t

$B_{t,sim}$: Percent relative bias of the regression estimates of b_t

$\lambda_{t,com}$: Relative factor of the fade-away of bias under approximation formula

$\lambda_{t,sim}$: Relative factor of the fade-away of bias under simulation study

R^2 : Coefficient of determination

σ_v^2 : Variance of the individual random effect v_i

σ_η^2 : Variance of the within individual error term $\eta_{i,t}$

$W_{i,t}$: Hourly wage of the individual i at time t

$W_{i,(t-1)}$: Lagged hourly wage of the individual i at time $t-1$

List of abbreviations

MCAR : Missing completely at random

MAR : Missing at random

NMAR : Not missing at random

PSID : Panel Study of Income Dynamics

ECHP : European Community Household Panel

EU-SILC : European Union Statistics of Income and Living Conditions

SOEP : German Socio Economic Panel

FRG : Federal Republic of Germany

DGP : Data generating process

CDF : Cumulative distribution function

OLS : Ordinary least squares

IPW : Inverse probability weighting

WLS : Weighted least squares

GLS : Generalized least squares

FGLS : Feasible generalized least squares

MLE : Maximum likelihood estimation

RLS : Restricted least squares

RCM : Random coefficient model

RR : Ridge regression

PLR : Piecewise linear regression

SLR : Simple linear regression

MLR : Multiple linear regression

NLPM : Non-linear probability model

PO : Proportional odds model

UOL : Un-weighted ordered logit model

WOL : Weighted ordered logit model

AR(1) : Auto-regressive model of order one

Fixone : One way fixed effects model

Ranone : One way random effects model

RE : Random effects

FE : Fixed effects

RB : Relative bias

MSE : Mean squared error

Part I

Introduction and Theoretical Foundations

CHAPTER 1

Introduction and Motivation

In practice, almost every survey suffers from the problem of non-response. The problem of non-response arises mainly due to the refusal of the persons to respond, and sometimes when there is unavailability of some persons, households, firms because of invalid addresses or wrong telephone numbers, inability of the interviewer to reach the household in remote areas or failure to collect the required information from a sample member in the mail surveys (Armstrong and Overton, 1977; Hox and deLeeuw, 1994). In the context of panel surveys, non-response occurs when the sample members don't participate in a particular wave of the study. This kind of non-response is called wave non-response. On the other hand, when a sample member participates in the initial wave of the survey but refuses to participate in the later waves of the survey, this kind of non-response is called panel attrition. Panel attrition is a common problem in panel surveys, which reduces sample size and can lead to biased inferences when the propensity to drop out is systematically related to the substantive outcome of interest (Behr et al., 2005; Olsen, 2005; Hogan and Daniels, 2008).

In this thesis, we are mainly concerned with the problem of non-response in panel surveys. Like any non-mandatory survey, a panel survey suffers from substantial non-response at its start. 30 to 70% of the initial sample persons refuse to cooperate.

The motivation and causation for this behaviour don't distinguish from standard cross-sectional surveys. However, in panel surveys, the respondents are repeatedly interviewed in later waves. With this repeated measurement it is possible to analyze gross-change, i.e., individual change, for example, changes between poverty and non-poverty. These individual changes have a substantial impact on the distribution of the variable interest in later waves of the panel. As a consequence, an initial bias resulting from selective non-response at the start of the panel may "fade-away" in later panel waves. The fade-away phenomenon can be empirically observed for those rare cases where a panel is selected from the register and where it is possible to make statistical inferences also for the non-responders based on the register information. Motivated by examples from the Finnish sub-samples of the European Community Household Panel (ECHP), the European Statistics on Income and Living Conditions (EU-SILC) and the German Panel Labor Market and Social Security (PASS) Alho et al. (2017) have developed a statistical framework in the context of Markov chains which explains the fade-away effect of initial non-response bias. The term "fade-away effect" is used by Rendtel (2003), details are given in Subsection 1 of this chapter.

There are three general approaches for analyzing incomplete data depending on the assumed missing data mechanisms which are summarized by Little and Rubin (Rubin, 1987; Little, 1988; Little and Rubin, 1987, 1989) and by some other researchers, e.g., (Ibrahim et al., 2005; Reiter, 2007; Durrant, 2009; Allison, 2000). These approaches are: complete case analysis (direct analysis of the incomplete data), weighting and imputation. The first case which is also called a complete case method is the simplest method for the analysis of incomplete data. According to this method, researchers completely ignore the missing part of the data and use only the complete cases of data for the analysis, e.g., listwise deletion (Bartels, 1993; Wawro, 2002; Briggs et al., 2003). Such a method works well when data are missing completely at random (MCAR), otherwise, it may introduce biases in survey estimates Nakai and Weiming (2011).

Weighting by the inverse of the estimated response probability, on the other hand, is considered as a traditional method for dealing with missing data and unit non-response. Weighting in the context of regression analysis is the commonly used procedures to compensate for the non-response bias of the estimated slope coefficients. In the case of panel surveys, there is considerable information available from previous panel waves, such as information on the lagged earnings. These

variables are good predictors of non-response and including them in the model as a weighting variable, the weighted least squares (WLS) estimator may contribute in reducing the amount of unexplained variation in the data due to non-response. Weighting may also help in reducing the non-response bias in the survey estimates through the use of calibration of sampling weights on external auxiliary information. Calibration estimation is commonly used in survey sampling whereby probability sampling weights are adjusted to increase the precision of estimates.

In the literature there exist several methods of calibration of weights, see, e.g., the unified approach of calibration proposed by Deville and Särndal (1992) by making use of prior information on auxiliary population totals and shares. They coined the term "calibration estimation" as a procedure of minimizing the distance measure between the initial survey weights and the final weights subject to the calibration equations. Firth and Bennett (1998) extend the idea of calibration estimation in the context of non-linear models. A more detailed overview of the calibration estimators can be found in Isaki and Fuller (1982), Särndal et al. (1989), Rao (1994), Chambers (1996), Estevao and Särndal (2000, 2006), Wu and Sitter (2001), Montanari and Ranalli (2005), Park and Fuller (2005), Kott (2006), Särndal (2007) and Kim and Park (2010).

Fitzgerald et al. (1998) developed an econometric framework for the analysis of attrition bias in a panel survey. In their influential study on attrition bias in the Panel Study of Income Dynamics (PSID) sample, they developed some tests for attrition bias that draws a sharp distinction on missingness on un-observables and observables. They concluded that WLS estimators can produce consistent parameter estimates when selection is based on the observables.

To understand this approach, let y and x be the activity and covariates of interest, and let us define R as a response indicator which is equal to 1 if there is a response and 0 if there is a non-response. Non-response is said to be missing at random (MAR) or ignorable, if non-response is independent of the variable of interest conditional on the observed variables. Mathematically, this can be written as $P(R=1|y,x) = P(R=1|x)$. Fitzgerald et al. (1998) and later on Moffitt et al. (1999) expand the definition of this probability function by proposing "selection on observables" such that the non-response depends not only on the x that are included in the model but also depends on the additional auxiliary variables z. These additional auxiliary variables are assumed to be observable for all units in the sample,

but are not included in the regression model. Further, the z variables are to be distinct from x but are to be endogenous to the y variable. Mathematically, selection on observables approach can be written as $P(R = 1|y, x, z) = P(R = 1|x, z)$. On the other hand, selection on un-observables occur if the assumption of selection on observables doesn't holds.

The method of imputation is to replace missing values in the data set by some suitable prediction and then applying standard methods to the complete data.

An essential strategy to minimize non-response consists of planning preventive actions to deal with the problem of non-response at the survey design stage. The goal of a well-designed survey is to reduce non-response rates by selecting the most appropriate fieldwork period, method of data collection, questionnaire design and layout, interview mode, interviewer training, protection of confidentiality of information provided, follow-up procedures and the effectiveness of respondent incentives, etc. Empirical studies of Groves and Couper (1998), Campanelli and O'Muircheartaigh (1999), Groves et al. (2002) and Riphahn and Serfling (2005), show that all these factors of survey design are typically crucial to explain response rates attained in sample surveys. Instead of these preventive measures adopted for reducing non-response the response rates rarely near to a hundred percent.

This describes why most of the survey literature on non-response concentrates on the development of statistical methods for ex-post adjustments of non-response, for detail, see Chapter 8 of Lessler and Kalsbeek (1992) and Little and Rubin (2002). Further, discussion on non-response and panel attrition in surveys can be seen elsewhere (Watson, 2003; Rendtel et al., 2004; Luca and Peracchi, 2007; Junes, 2012).

The "fade-away effect" in panel surveys

Non-response in surveys may create a bias in the estimates. However, one advantage of panel surveys over cross-sectional surveys is that under some regularity conditions an initial non-response bias may fade-away over later panel waves. Sisto (2003) and Rendtel (2013) studied the effect of initial non-response on the income quintiles estimates from the European Community Household Panel (ECHP) and poverty rates from the European Union Statistics of Income and Living Conditions (EU-SILC). They reported that the effect of initial non-response bias declines very fast for income

quintiles and poverty states in the subsequent panel waves. Such a hypothesis of the fade-away effect doesn't only base on the information provided by the respondent sample but also depends on the information obtained from the non-respondent sample, where information about the non-respondents is available via registers.

Rendtel (2013) used the concept of the Markov chain to explain the fade-away phenomenon. The purpose of using this approach is the possibility to use the steady-state distribution of the Markov chain. If the transition law of the Markov chain is stable over time, then under some regularity conditions the distribution on the state space of the Markov chain converges to a stable distribution, called the steady-state distribution. The convergence takes place irrespective of the starting distribution of the Markov chain. However, the state transition law for the respondents and non-respondents between panel waves must be the same.

Alho et al. (2017) present an extension of the fade-away phenomenon which is not restricted to a time-homogeneous transition law of the Markov chain. They provide some theoretical results on the speed of the convergence to the steady-state distribution. Alho (2015) extends the approach to regression analysis. He uses a two wave panel to explain the fade-away phenomenon of initial non-response bias in the framework of regression analysis with a single covariate. In the proposed regression model the covariate and the error term are decomposed into permanent and non-permanent variance components. Alho concludes that the initial non-response bias fades-away in the case of low non-permanent components of the covariate and/or the error term.

Outline of the thesis

The thesis is divided into three parts: Part I contains the theoretical foundations for the fade-away effect of initial non-response bias in panel surveys. In part II of this thesis, a simulation study is conducted to investigate the fade-away effect of the initial non-response bias in a multi-wave panel survey. The purpose of the simulation study is to investigate the accuracy of the bias approximation in a simulation setting and check the size of the fade-away effect in later panel waves with no analytical bias approximation. Alho (2015) has investigated the bias of cross-sectional OLS estimates under not missing at random (NMAR) non-response at the start of the

panel. He derived analytical bias approximation for the OLS estimate of the slope coefficient of the variable of interest. His underlying model used a variance component model with two components: a fixed individual component and an auto-regressive shock component (Alho's model will be discussed in Subsection 2.3.2 of Chapter 2). However, in multi-wave panel surveys, the analytical expression for Alho's bias approximation formula becomes very intractable for later waves. Therefore, we extend the results to a longer panel wave via a simulation study.

The remainder of Chapter 3 in Part II is structured as follows: Section 3.2, presented an extension of the fade-away phenomenon to a multi-wave panel survey. To judge the performance of the estimators, a Monte Carlo simulation study is conducted in Section 3.3. In Section 3.4, we discussed the fade-away effect of the cross-sectional OLS estimators with and without panel attrition. We then compared the estimation results of weighted and un-weighted cross-sectional OLS estimators in Section 3.5. In the next section (Section 3.6) we discussed the behaviour of different panel estimators. Finally, Section 3.7 is devoted to the estimation of the non-linear ordered logit model. Here we compared the fade-away effect of the un-weighted estimates of ordered logit model with several weighting approaches. The estimation is done with **SAS** and with the procedure: **PROC REG**, **PROC LOGISTIC** and **PROC PANEL**, respectively.

In Chapter 3 of this thesis, we have conducted a simulation study to verify the approximate results of Alho (2015), and investigated the accuracy of the bias approximation in a simulation setting. We checked the size of the fade-away in later panel waves with no analytical bias approximation. The speed of the fade-away effect of the initial non-response bias is then investigated for different stability scenarios of covariates and error terms, with and without any attrition patterns in later panel waves. As the speed of the fade-away depends on the stability of the covariates and error terms it is important to investigate this effect not only for simulated data but also for real longitudinal data. Therefore, in the application part (Part III) of this thesis, we switch to real data from the German Socio Economic Panel (SOEP): specifically to income data and life satisfaction scores data of the SOEP. Here we used the following two settings:

- Income data and their explanation via regression.
- Life satisfaction scores and their explanation by an ordered logit model.

While Alho's approximation was developed for cross-sectional OLS estimates, in a panel more complex estimators are used, for example, estimators that use information from several panel waves and incorporate dependent error terms. For such models and estimators, analytical bias approximation is out of scope. Therefore, simulation will be also used here to study the fade-away effect and its size.

In the first part of Chapter 4 in Part III, we will investigate the fade-away effect of initial non-response on the estimation of a wage equation using data from the first 10 panel waves of the SOEP. To examine the fade-away effect, we use three different model settings: a random effects (RE) model for wages with and without lagged dependent variable, a RE model with auto-correlated errors for wages and a fixed effects (FE) wage model. Here we will not use any simulation of the dependent variable and the covariates. The only part which is simulated is the endogenous drop-out of observations at the start of the panel. Furthermore, we will switch to a model with multiple covariates. In order to demonstrate the fade-away effect, we will gradually extend the database from 1 to 10 panel waves. This covers the lengths 1984 to 1993 of the Sub-sample A-B of the SOEP. Contrary to the previous work in Chapter 3, where the dependent variable and the covariate are simulated here, we use real data from the SOEP. The estimations are done with **SAS** and with the procedure: **PROC REG**, **PROC PANEL**, and **PROC HPMIXED**, respectively.

The second part of Chapter 4 is to explore the effect of initial-response on the estimation of a model which explains life satisfaction scores by using SOEP data from the year 2000 to 2010 of the Sub-sample F. To examine the fade-away effect, we use two models: the ordered logit model for cross-sectional data and the random effects (RE) model for longitudinal data. In order to demonstrate the fade-away effect we gradually extend the database from 1 to 11 panel waves. This covers 11 panel waves of the SOEP starting from the year 2000 to 2010. The estimations are done with SAS and with the procedure: **PROC HPLOGISTIC** and **PROC GLIMMIX**, respectively.

Finally, The thesis ends with a conclusion and some open research questions in Chapter 5.

CHAPTER 2

Theoretical Foundations

2.1. Missing data mechanism

In this section, we discuss the modeling of the missing values which is useful to explain which value(s) in a data set are observed and which value(s) are missing. As we have discussed in the introduction chapter, the problem of non-response doesn't only occur in cross-sectional data but also in panel data. In this thesis, we are mainly concerned with the problem of unit non-response and attrition in panel surveys. Therefore, it is important to highlight the essential differences between the unit non-response in the initial wave of a panel and the sample attrition in the subsequent waves of a panel. First, the study of unit non-response in the initial wave of a panel is usually aggravated by the lack of adequate information on the units who refuse to participate in the survey, whereas information collected during panel waves preceding attrition can be used to analyze attrition.

One important issue in studying non-response is to establish whether the data generating mechanism is missing at random or not. One possibility is to use the terminology introduced by Rubin (1976) or the second possibility is to use the more econometric approach developed by Fitzgerald et al. (1998) which will be discussed in Subsection 2.1.2 of this chapter.

Using the terminology proposed by Rubin (1976) and Little and Rubin (2002), one can explain three possible missing data mechanisms. Consider a data set, say Y, which is decomposed into two parts, observed Y_{obs} and unobserved Y_{mis}, by a dummy variable R, such that if $R=0$ the data is missing due to initial non-response or attrition and $R=1$ if data is observed. The conditional distribution of the missing data mechanism is denoted by $f(R|\pi)$, where π represents the unknown parameters.

A missing data mechanism is said to be missing completely at random (MCAR) if missingness doesn't depend on the observed Y_{obs} and the missing Y_{mis} of the variable Y. The MCAR assumption can then be described by the expression:

$$f(R=1|Y,\pi)=f(R=1|\pi), \quad \text{for all} \quad Y,\pi \tag{2.1}$$

That is, the conditional distribution of R given Y doesn't depend on the observed nor the unobserved values of Y, but only depends on the unknown parameter π.

The missing data mechanism is said to be missing at random (MAR), if the missingness depends only on the observed part of the data and is independent from the unobserved part of the data. The MAR assumption can be stated as:

$$f(R=1|Y,\pi)=f(R=1|Y_{obs},\pi), \quad \text{for all} \quad Y_{obs},\pi \tag{2.2}$$

The missing data under MCAR and MAR mechanisms are said to be ignorable, and estimates obtained in the presence of ignorable MCAR or MAR mechanisms have a negligible bias. However, this is not true in general. In contrast, if the assumption of MAR is violated, the missingness is said to be not missing at random (NMAR), where the missingness is additionally dependent on the unobserved part of the data that is:

$$f(R=1|Y,\pi)=f(R=1|Y_{obs},Y_{mis},\pi), \quad \text{for all} \quad Y,\pi \tag{2.3}$$

In the NMAR case, the missing data mechanism is not ignorable. Careful planning of the study can reduce any potential impact of an NMAR mechanism either by including direct measures of the potential causes of missingness or by including reasonable proxies or known correlates of the causes of such missingness. By inclusion of these proxies as a covariate in the resultant missing data analysis can reduce the size of the non-response bias Schafer (1997) and Enders (2010).

A good estimation method is required so that the missing data mechanism can

be modeled as part of the estimation process, otherwise, it may produce biased estimates. There are two well-known methods for analyzing data under the assumptions of NMAR. These methods are the: Heckman (1979) selection model and a pattern mixture model. According to Heckman's selection method, the purpose is to estimate a linear regression model with missing data under the NMAR mechanism on the outcome variable Y. A second method for analyzing data under the NMAR assumption is the pattern mixture model. This method is based on the distributional differences of the observed and the missing data by specificity of a separate regression model for each pattern. These models can then be used to constitute inferences. For a more comprehensive overview regarding pattern mixture models, see Molenberghs et al. (1988), Little (1995), Thijs et al. (2000), Daniels and Hogan (2000), Demirtas and Schafer (2003) and Carpenter and Kenward (2013, Subsection 1.4.3 of Chapter 1).

2.1.1. The likelihood approach of Rubin

One approach for handling missing data is the maximum likelihood (ML) approach. If the data is MAR, then the ML method yields estimates that are consistent, asymptotic efficient and asymptotic normal Allison (2001). According to this approach, all variables in a data set can be partitioned into two groups $Y = (Y_{obs}, Y_{mis})$. The first group consists of only those variables whose values are observed, say Y_{obs} (as defined previously). The second group then consists of only those variables whose values are missing, say Y_{mis}. We use the previously defined binary variable R, which is 1 if the variable is observed and 0 if it is missing. Then the probability density of complete data model $f(Y|\omega)$ can be written by $f(Y|\omega) = f(Y_{obs}, Y_{mis}|\omega)$ where ω is a parameter to be estimated. According to the likelihood-based approach, the joint distribution of Y and R can be written as

$$f(Y, R|\omega, \pi) = f(Y|\omega) f(R|Y, \pi), \quad (\omega, \pi) \in \Omega_{\omega,\pi}, \tag{2.4}$$

where the conditional density of R given Y is the model for missing data mechanism with an unknown parameter π which represents the distribution for the missing data mechanism, while $\Omega_{\omega,\pi}$ is the parameter space of ω and π. If there are missing values in the data, then the joint observed density function of Y_{obs} and R can be obtained

by integrating equation (2.4) with respect to Y_{mis}, i.e.,

$$f(Y_{obs}, R|\omega, \pi) = \int f(Y_{obs}, Y_{mis}|\omega) f(R|Y_{obs}, Y_{mis}, \pi) dY_{mis} \tag{2.5}$$

If the missing data mechanism is independent of the missing values and depends only observed values of Y_{obs} then the joint density of Y_{obs} and R in equation (2.5) becomes

$$\begin{aligned} f(Y_{obs}, R|\omega, \pi) &= f(R|Y_{obs}, \pi) \int f(Y_{obs}, Y_{mis}|\omega) dY_{mis} \\ &= f(R|Y_{obs}, \pi) f(Y_{obs}|\omega). \end{aligned} \tag{2.6}$$

As in the above case missingness only depends on the observed data in the sample, so the missing data process may be ignored. Rubin (1976) called this approach missing at random (MAR). The log-likelihood function of the observed density function can be attained by taking the log on both sides of the equation (2.6), i.e.,

$$log f(Y_{obs}, R|\omega, \pi) = log f(R|Y_{obs}, \pi) + log f(Y_{obs}|\omega) \tag{2.7}$$

In the above log-likelihood the parameter π (missingness-mechanism) and ω (data model) are now distinct. Therefore, the likelihood function can be maximized with respect to π if the likelihood function of the observed sample $f(Y_{obs}|\omega)$ is maximized. According to Rubin (1976) a missing-data mechanism is ignorable for likelihood-based inference if the following conditions are satisfied:

- The missing value mechanism is missing at random (MAR): $f(R|Y_{obs}, Y_{mis}, \pi) = f(R|Y_{obs}, \pi)$, for all Y_{obs},

- Distinctness: ω and π have distinct parameter spaces.

If the MAR condition holds but not the distinctness one, then the ML-based on ignorable likelihood is valid but not fully efficient. So MAR is the key condition Diggle et al. (2002). However, if the distribution of R depends on both Y_{obs} and Y_{mis} in such case the non-response is non-ignorable and the mechanism is known as non-ignorable missing data mechanism.

2.1.2. Missing on observables or unobservables

Another approach for the analysis of attrition bias is the typology introduced by Fitzgerald, Gottschalk, and Moffitt (1998). To understand this approach considers the framework: Let $Y_t = (Y_{1,t}, Y_{2,t}, \ldots, Y_{n,t})$ be the dependent variable at point t, and let $R_t = (R_1, R_2, \ldots, R_T)$ be the response indicator at time point t, which is 1 if Y_t is observed and 0 if Y_t is missing. Let X_t be a set of covariates of the model, then the linear regression model for the outcome variable is:

$$Y_t = \beta' X_t + \varepsilon_t, \tag{2.8}$$

where β is a set of regression coefficients, and ε_t is the error term of the model at time t. Also, let Z be a set of observed covariates that are used to explain attrition but not the behaviour of the outcome variable. To distinguish between attrition on observable and unobservable Fitzgerald et al. (1998) proposed the following attrition model:

$$R_t^\star = \gamma_1' X_t + \gamma_2' Z + \delta_t, \tag{2.9}$$

such that

$$R_t = \begin{cases} 1, & \text{if } R_t^\star > 0, \\ 0, & \text{if } R_t^\star < 0. \end{cases} \tag{2.10}$$

where $R_t^\star$ is the latent response indicator and attrition occurs if this indicator is less than zero, and δ_t is the random influence on the attrition probability.

In this context missing on observables occurs if Z is not independent of $\varepsilon_t | X_t$ and δ_t is independent of $\varepsilon_t | X_t$ i.e.,

$$\varepsilon_t \perp \delta_t | X_t \quad \text{and} \quad \varepsilon_t \not\perp Z | X_t$$

Stated alternatively, missing on observables simply holds if

$$f(R_t | Y_t, X_t, Z) = f(R_t | X_t, Z). \tag{2.11}$$

The dependence between ε_t and Z means that the inclusion of Z in the regression model would result in a non-zero coefficient for Z, while the independence of ε_t and δ_t means that the unobserved variables effects both the selection equation and the

main equation. If $Z = Y_{t-1}$ the lagged dependent variable, which is observed before attrition in wave t. Then equation (2.11) implies that attrition doesn't depend on the change $Y_t - Y_{t-1}$ of the dependent variables before the attrition occurs (Rendtel, 2002, p. 11). On the other hand, missing on unobservables occurs if Z is independent of $\varepsilon_t|X_t$ and δ_t is not independent of $\varepsilon_t|X_t$

$$\varepsilon_t \not\perp \delta_t|X_t \quad \text{and} \quad \varepsilon_t \perp Z|X_t$$

Stated alternatively, if the conditional independence of attrition function in equation (2.11) is not satisfied i.e.

$$f(R_t|Y_t, X_t, Z) \neq f(R_t|X_t, Z) \tag{2.12}$$

In the case of missing on unobservables, if $Z = Y_{t-1}$ this would mean that change $Y_t - Y_{t-1}$ has an impact on the attrition, while the whole impact of Y_{t-1} on Y_t is absorbed by the covariates (Rendtel, 2002, p. 10-11).

2.2. Regularity conditions for the fade-away effect

2.2.1. Markov chain model for state transition

A Markov chain is a mathematical model for random phenomena evolving over time (Norris, 1997, p. ix). In general, let $\{Y_t, t = 1, 2, 3, \ldots\}$ be a discrete-time stochastic process with finite state space $E = \{1, 2, 3, \ldots, I\}$. Also let R_t be the binary response indicator, where for $R_t = 1$ we have a response at wave t and for $R_t = 0$ we have non-response at wave t. If the conditional probabilities (also called transition probabilities) at time t satisfy:

$$\begin{aligned} P(Y_t = j|Y_1 = i_1, Y_2 = i_2, \ldots, Y_{t-1} = i) &= P(Y_t = j|Y_{t-1} = i) \\ &= P_{i,j}(t), \end{aligned} \tag{2.13}$$

then the process is called a Markov chain model.

The $I \times I$ dimensional matrix of transition probabilities from time $t-1$ to time t is $P(t) = (P_{i,j}(t))$. A Markov chain model is said to be time-homogeneous if the state transition probabilities are independent of the time t. Therefore, by removing the

time subscript t from the transition probabilities matrix $P(t)$ we simply get $P = (P_{i,j})$, for all $t = 1, 2, 3, \ldots, T$. Transition probabilities from time 1 to time t are given by $P^{(t)} = P(1)P(2), \ldots, P(t)$. In the case of a time-homogeneous Markov chain, it can be written as $P^{(t)} = P \times P \times P \ldots P \times P = P^t$. Therefore, for time-homogeneous chains, the t step transition probability is simply the $t-$fold product of the single step probabilities.

One important property of the time-homogeneous Markov chain is that for a long time run and for large $t = \infty$, the distribution of the chain converges to a limiting distribution of the chain, called the steady-state distribution. Mathematically, let Y_t be a finite state space time-homogeneous Markov chain with state space $E = \{1, 2, 3, \ldots, I\}$ and transition probability matrix P. Further, let $\pi^\star = \{\pi_1^\star, \pi_2^\star, \pi_3^\star, \ldots, \pi_I^\star\}$ be a vector of the probability distribution on the state space E then if it satisfies the following properties:

- $\pi_i^\star \geq 0$ for $i = 1, 2, 3, \ldots, I$ and $\sum_{i=1}^{I} \pi_i^\star = 1$,
- $P^t \times \pi^\star = \pi^\star$.

From the second property, we see that with performing one step of the Markov chain starting with the distribution $\pi^\star$ results in the same distribution $\pi^\star$ of the chain after time t. Then the distribution $\pi_i^\star$ is called the steady-state distribution of the Markov chain. Further insights about the steady-state distribution of the Markov chain are given in (Häggström, 2002, p. 29-30).

2.2.2. Contraction theorem

Let us consider two finite samples drawn from the same population of sufficiently large size. Further, let the members of the two samples follow a Markov chain model, and change their state according to the same transition probabilities $P(t)$. The initial sample which is denoted by **Full-Sample**, consists of all those individuals who were selected by the sampling design from the target population. From the Full-Sample we obtain the Response-Sample which consists of all persons who respond at the start of the panel. This sample is denoted by **Resp-Sample**.

The initial starting distributions of the two samples on the state space are the column vectors $\pi_{Full}(1)$ and $\pi_{Resp}(1)$, respectively. For the subsequent state

distributions on the state space of the two samples, it satisfies the recursions formula $\pi_{Full}(t) = P'(t)\pi_{Full}(t-1)$ and $\pi_{Resp}(t) = P'(t)\pi_{Resp}(t-1)$ for $t = 1, 2, \ldots, T$. When all the entries of the vector $\pi_{Resp}(t)$ are strictly positive, we have the lower and upper bounds of the inequalities:

$$m_t \equiv \min_i \frac{\pi_{Full,i}(t)}{\pi_{Resp,i}(t)} \leq \frac{\pi_{Full,j}(t)}{\pi_{Resp,j}(t)} \leq \max_i \frac{\pi_{Full,i}(t)}{\pi_{Resp,i}(t)} \equiv M_t, \tag{2.14}$$

where $j = 1, \ldots, I$. Then, we have the following theorem which states that under some regularity conditions, the distributions of the Full-Sample and the Resp-Sample converge to equal state distributions of the Markov chain.

Theorem 1 *Suppose that there is lower bound $0 < p_L \leq p_{i,j}(t)$ for all t. Then the two distributions $\pi_{Full}(t)$ and $\pi_{Resp}(t)$ converge uniformly in the sense such that:*

$$\lim_{t \to \infty} (M_t - m_t) = 0. \tag{2.15}$$

The above results are taken from Alho et al. (2017).

2.2.3. Initial non-response and its fade-away effect

If non-response at the start of the panel survey is not selective or ignorable for the estimation of population parameters, the distribution of the Resp-Sample will be equal to the distribution of the Full-Sample. In such a scenario there would be no bias in the Resp-Sample at any panel wave and hence, therefore, there would be no fade-away phenomenon present. Therefore, we assume that the initial non-response is not missing at random (NMAR) or highly selective for the estimation of population parameters. And therefore the initial distribution of the Resp-Sample at the start of the panel is somewhat away from the Full-Sample distribution. Under this respect, if there is no further selective attrition after wave 1, then according to the results of the contraction theorem (***Theorem 1***) the distorting effects of initial non-response in the Resp-Sample is expected to become smaller and smaller over time. And hence the distributional differences of the Full and the Resp samples are expected to fade-away over the passage of time.

2.2.4. Limitation: Mover-stayer models

The mover-stayer model which is an extension of the Markov chain model is frequently used in panel analysis. Actually, this is a mixture model where the population is divided into two groups/parts. The first group of the population is the group of stayers which consists of persons who don't change their state and thus their transition probability is zero. The second group of the population is the group of movers who change their state according to the Markov chain model with a positive transition probability. Blumen et al. (1955) used the discrete-time mover-stayer model to deal with the unobserved heterogeneity in the population. Their study revealed that the Markov chain models tend to predict too many changes after several transition periods. According to these authors, the reason of this problem was a mixture model with two chains, where the first chain "the movers" follows a simple Markov chain, predicting too much change after some transition periods. While, the second chain "the stayers" stay in their initial state and predicts no change at all, even after many periods. Frydman (1984) suggests the method of ML for estimating the discrete-time mover-stayer model. Later on, his work was extended by Frydman et al. (1985) for testing the adequacy of the discrete-time mover-stayer model for describing credit behavior. Altman and Kao (1991) implement the method in Frydman et al. (1985) for examining the behaviour of rating migration. For further discussion regarding the mover-stayer models see Goodman (1961), Spilerman (1972), Singer and Spilerman (1976), Poulsen (1983), Van de Pol and Langeheine (1989), Frydman and Kadam (2002) and references therein.

As the results on the fade-away effect depend on the transition law of the Markov chain model which was described earlier in this section. So if the sample is divided into two parts: the movers, who change their position through the Markov chain model and the stayers who don't change their position at all. Then, according to the results of contraction (***Theorem 1***), there would be a fade-away effect for the movers, but no fade-way effect is expected for the stayers of the mover-stayer model.

2.3. Regression setting

Regression analysis is a statistical technique that investigates the relationship between a dependent variable and one or more independent variable(s). The dependent

variable is also called the response variable, outcome variable, criterion variable, explained variable, regressand variable, or endogenous variable. While other names of the independent variables can also be called predictor variables, exogenous variables, control variables, covariates or regressors. This technique is widely used for time series modeling, forecasting and effect or trend forecasting and finding the causal effect relationship between the variables. Today, the technique is used in almost every field of life and has many applications in the field of engineering, business, as well as in the social, physical, and biological sciences, and many more. For the application of regression analysis in different fields of sciences see Kutner et al. (2005).

Based on the type of relationships between the response and the predictor variables regression models may be broadly divided into two categories, the linear regression models and the non-linear regression models. The response variable is usually linked to the predictor variables through some parameters. The regression models are said to be linear when it is linear in the parameters, while, in non-linear regression models non-linearity appears in parameters. It is to be noted that the models are still linear in parameters even when the predictor variables are in square, square root, exponent form, or any other non-linear form.

In the following subsection (Subsection 4.2.1), we provide a brief introduction on the linear models, the non-linear models are then discussed in Subsection 2.3.4 of this chapter.

2.3.1. The linear model

As discussed above, regression analysis is a statistical technique which investigates the relationship between a dependent variable and one or more independent variable(s). We start with a basic simple linear regression model where there is only one response variable and one predictor variable and the regression line is linear. The response variable is continuous, whereas the predictor variable may be either continuous or discrete.

Let Y indicate the response variable which is linearly related to the predictor variable say X_1 through the parameters β_0 and β_1, then the model can be stated as:

$$\begin{aligned} Y_i &= f(X_{i1}) + \varepsilon_i \\ &= \beta_0 + \beta_1 X_{i1} + \varepsilon_i, \quad \text{for} \quad i = 1, 2, 3, \ldots, N. \end{aligned} \tag{2.16}$$

This is called a simple linear regression (SLR) model. The function is $f(X_{i1})$ is the population regression equation of Y_i on X_{i1}. β_0 and β_1 are the regression parameters: β_0 is the intercept and β_1 is the regression slope coefficient associated with X. The term ε is the random error or residual term of the model that represents the fact that the data couldn't fit the model perfectly. The ε_i are assumed to be independent and identically normally distributed variables having mean zero and constant variance σ_ε^2, written as $\varepsilon \overset{iid}{\sim} N(0, \sigma_\varepsilon^2)$.

However, when the response variable Y is linearly related to the K explanatory variables $X_{i0}, X_{i1}, X_{i2}, \ldots, X_{iK}$ through the parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_K$ then the model is known as multiple linear regression (MLR) model. The general form of the model can be stated as:

$$\begin{aligned} Y_i &= f(X_{i0}, X_{i1}, X_{i2}, \ldots, X_{iK}) + \varepsilon_i, \\ &= \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_K X_{iK} + \varepsilon_i, \\ &= \beta_0 + \sum_{k=1}^{K} \beta_k X_{ik} + \varepsilon_i, \quad \text{for} \quad i = 1, 2, 3, \ldots, N. \quad \text{and} \quad k = 1, 2, 3, \ldots, K. \end{aligned} \tag{2.17}$$

Where the first $X_{i0} = 1$ is a constant unless otherwise stated, and $\beta_0, \beta_1, \beta_2, \ldots, \beta_K$ are the $K+1$ model parameters including the intercept β_0.

In matrix form, the multiple linear regression model can be rewritten as follows:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_N \end{bmatrix}_{N\times 1} = \begin{bmatrix} 1 & X_{11} & X_{12} & \ldots & X_{1K} \\ 1 & X_{21} & X_{22} & \ldots & X_{2K} \\ 1 & X_{31} & X_{32} & \ldots & X_{3K} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{N1} & X_{N2} & \ldots & X_{NK} \end{bmatrix}_{N\times(K+1)} \times \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{bmatrix}_{(K+1)\times 1} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_N \end{bmatrix}_{N\times 1}$$

Or, in more general form it can be expressed as:

$$Y = X\beta + \varepsilon, \tag{2.18}$$

where $Y = [Y_1, Y_2, Y_3, \ldots, Y_N]^{'}$ is the $N \times 1$ column vector of observations on the response variable and $X = [1_N, X_1, X_2, \ldots, X_K]$ is a data matrix of order $N \times (K+1)$ of N observations of the K explanatory variables. Also, $\beta = [\beta_0, \beta_1, \beta_2, \ldots, \beta_K]^{'}$ of

dimensions $(K+1)\times 1$ is the vector of population regression slope parameters and $\varepsilon = [\varepsilon_1, \varepsilon_2, \varepsilon_3, \ldots, \varepsilon_N]'$ is the column vector of $N \times 1$ of regression residual terms.

2.3.1.1. Estimation with ordinary least squares method

The population parameters of the model are usually unknown and have to be estimated by some suitable methods of estimation. In the literature, there exist various estimation methods for estimating model parameters namely: ordinary least squares (OLS), weighted least squares (WLS), generalized least squares (GLS), maximum likelihood (ML), restricted least squares (RLS), random coefficient (RC), ridge regression (RR), piecewise linear regression (PLR), variable parameter (VP) and regression models for qualitative variables. A general procedure for estimating model parameters are to find vector $\hat{\beta}$ which minimizes the residual sum of squares (this refers only to the method of OLS).

To estimate the parameter vector β with the method of least squares, the first step is to find the residual sum of squares and then find a set of estimators that minimize the squared distances. Let $\hat{\beta}$ be the estimate of the parameter vector β, then the estimated fitted model is given by:

$$\hat{Y} = X\hat{\beta} \tag{2.19}$$

Then the vector of least squares residuals $\hat{\varepsilon}$ is:

$$\hat{\varepsilon} = Y - X\hat{\beta} \tag{2.20}$$

Thus, the minimizing problem of the residual sum of squares $\hat{\varepsilon}'\hat{\varepsilon}$ of the general linear model in matrix form is as follows:

$$\begin{bmatrix} \hat{\varepsilon}_1 & \hat{\varepsilon}_2 & \hat{\varepsilon}_3 & \ldots & \hat{\varepsilon}_N \end{bmatrix}_{1\times N} \times \begin{bmatrix} \hat{\varepsilon}_1 \\ \hat{\varepsilon}_2 \\ \hat{\varepsilon}_3 \\ \vdots \\ \hat{\varepsilon}_N \end{bmatrix}_{N\times 1}$$

This can also be written as:

$$\begin{aligned}\hat{\varepsilon}^{'}\hat{\varepsilon} &= (Y - X\hat{\beta})^{'}(Y - X\hat{\beta}) \\ &= Y^{'}Y - \hat{\beta}^{'}X^{'}Y + Y^{'}X\hat{\beta} + \hat{\beta}^{'}X^{'}X\hat{\beta} \\ &= Y^{'}Y - 2\hat{\beta}^{'}X^{'}Y + \hat{\beta}^{'}X^{'}X\hat{\beta} \end{aligned} \tag{2.21}$$

Notice that the term $Y^{'}X\hat{\beta}$ is scalar, and we know that the transpose of a scalar is the scalar/number itself, so we can write $Y^{'}X\hat{\beta} = (Y^{'}X\hat{\beta})^{'} = \hat{\beta}^{'}X^{'}Y$.

Now in order to obtain $\hat{\beta}$ that minimizes the function $\hat{\varepsilon}^{'}\hat{\varepsilon}$, differentiate equation (2.21) w.r.t. to $\hat{\beta}$ and equating it equal to zero. That is:

$$\begin{aligned}\frac{\partial(\hat{\varepsilon}^{'}\hat{\varepsilon})}{\partial\hat{\beta}} &= 0 \\ &= -2X^{'}Y + 2X^{'}X\hat{\beta} = 0 \end{aligned} \tag{2.22}$$

This gives:

$$(X^{'}X)\hat{\beta} = X^{'}Y, \tag{2.23}$$

If the inverse of the matrix $X^{'}X$ exists, then by pre-multiplying both sides of the equation (2.23) by the inverse matrix of $(X^{'}X)^{-1}$, we get:

$$\hat{\beta} = (X^{'}X)^{-1}X^{'}Y, \tag{2.24}$$

which is known as ordinary least squares (OLS) estimator of β of the general linear regression model in matrix form.

2.3.2. Alho bias approximation

Alho (2015) considers a simple linear regression model of two panel waves, say at time 1 and time 2. The proposed regression models are:

$$Y_{i,t} = a_t + b_t X_{i,t} + e_{i,t}, \quad \text{for} \quad t = 1, 2. \tag{2.25}$$

The models describe the relationship between the response variable $Y_{i,t}$ and the

predictor variable $X_{i,t}$ that varies over time, where a_t is the intercept, b_t represents the response coefficient which measures the effect of $X_{i,t}$ on $Y_{i,t}$. The disturbance terms $e_{i,t}$ have expectations $E(e_{i,t}) = 0$ and variances $Var(e_{i,t}) = \sigma^2$. The explanatory variables $X_{i,t}$ have a common factor M_i and a time-varying components $Z_{i,t}$. It is assumed that the explanatory variable $X_{i,t}$ is uncorrelated with the stochastic disturbance terms $e_{i,t}$ and have the following form:

$$X_{i,t} = M_i + Z_{i,t},$$

where the permanent part M_i of the explanatory variable is assumed to be uncorrelated with the time-varying component $Z_{i,t}$ with expectations $E(M_i) = E(Z_{i,t}) = 0$, and variance components $Var(M_i) = \kappa$ and $Var(Z_{i,t}) = 1 - \kappa$, where $0 \leq \kappa \leq 1$.

Also, the disturbance terms $e_{i,t}$ of the regression equation have a variance component structure:

$$e_{i,t} = V_i + U_{i,t},$$

Following a similar argument V_i is the permanent part of the disturbance factors which is uncorrelated with the time-varying components $U_{i,t}$ with expectations $E(V_i) = E(U_{i,t}) = 0$ and variances $Var(V_i) = \gamma\sigma^2$ and $Var(U_{i,t}) = (1-\gamma)\sigma^2$, where $0 \leq \gamma \leq 1$.

Further, it is assumed that the time-varying components of the covariate and the disturbance term vary over time according to an auto-regressive process of order one (AR(1)). The respective models are as follows:

$$Z_{i,2} = \rho Z_{i,1} + \varepsilon_{i,2}, \tag{2.26}$$

and

$$U_{i,2} = \phi U_{i,1} + \xi_{i,2}. \tag{2.27}$$

where ρ and ϕ are the stability of the time-varying parts of the covariates and the residual factors. Also $\varepsilon_{i,2}$ and $\xi_{i,2}$ are the "fresh error terms" which are not correlated and have expectations $E(\varepsilon_{i,2}) = E(\xi_{i,2}) = 0$ with variance components $Var(\varepsilon_{i,2}) = (1-\kappa)(1-\rho^2)$ and $Var(\xi_{i,2}) = (1-\gamma)(1-\phi^2)\sigma^2$.

Here it is assumed that at the start of the survey at time 1, an individual decides

whether he/she would participate in the survey or not. Let $R_{i,1} = 1$ be the response indicator if the corresponding person i is willing to participate in the survey initially, and where $R_{i,1} = 0$ otherwise. Following Alho (2015), the response probability of a person i is defined by a linear model

$$P(R_{i,1} = 1|Y_{i,1}) = \alpha + \beta Y_{i,1}, \tag{2.28}$$

where $0 < \alpha < 1$ and $0 < \beta < 1$ are the initial non-response parameters whose values are to be chosen in such a way such that the response probabilities are in the interval between [0, 1]. Further, it is assumed that the distribution of $Y_{i,1}$ at the initial wave 1 is highly selective, i.e., non-response at the start of the panel is supposed to be non-ignorable for the estimation of regression coefficients. Then the marginal probability of response of a person i is

$$\begin{aligned} P(R_{i,1} = 1) &= E[P(R_{i,1} = 1|Y_{i,1})] \\ &= \alpha + \beta a_1. \end{aligned} \tag{2.29}$$

where α and β are small numbers so that the probabilities are in the interval [0, 1] and $a_1 = E(Y_{i,1})$. The following results are due to Alho (2015).
The OLS estimate of the true regression coefficient b_1, obtained from those who initially participate in the survey at time 1, is given by

$$\begin{aligned} p\lim_{n\to\infty}(\hat{b}_{1,com}) &\approx Cov(X_1, Y_1|R_1 = 1)/Var(X_1|R_1 = 1) \\ &\approx b_1(1 - \frac{\beta^2\sigma^2}{(\alpha + \beta a_1)^2 - \beta^2 b_1^2}), \end{aligned}$$

Thus one may derive the initial non-response bias of OLS estimate of b_1 is, as follows

$$Bias(\hat{b}_{1,com}) = \frac{\beta^2\sigma^2 b_1}{(\alpha + \beta a_1)^2 - \beta^2 b_1^2}. \tag{2.30}$$

Assuming that individuals who participate in the survey initially, are also willing to participate in the second wave of the survey at time 2, with the estimated regression

coefficient of b_2, may be approximated by:

$$\begin{aligned} p\lim_{n\to\infty}(\hat{b}_{2,com}) &\approx Cov(X_2, Y_2|R_1 = 1)/Var(X_2|R_1 = 1) \\ &\approx (b_2 - \frac{b_1\beta^2(\kappa+\rho(1-\kappa))(\gamma+\phi(1-\gamma))\sigma^2}{(\alpha+\beta a_1)^2 - \beta^2 b_1^2(\kappa+\rho(1-\kappa))^2}), \end{aligned}$$

with subsequent non-response bias of OLS estimate of b_2, which is stated by

$$Bias(\hat{b}_{2,com}) = \frac{b_1\beta^2(\kappa+\rho(1-\kappa))(\gamma+\phi(1-\gamma))\sigma^2}{(\alpha+\beta a_1)^2 - \beta^2 b_1^2(\kappa+\rho(1-\kappa))^2}. \tag{2.31}$$

Note that the condition $Bias(\hat{b}_{2,com}) < Bias(\hat{b}_{1,com})$, always holds for all values of the stability parameters in the interval $0 < \kappa < \rho < 1$ and $0 < \gamma < \phi < 1$. Detail analysis tables showing the size of the bias and its fade-away effect (by using bias approximation formula and through simulation study) are presented in Table 10 to table 17 in Section A.1 of Appendix A.

2.3.2.1. The speed of the fade-away effect

In order to find the speed of the fade-away effect of the initial non-response bias in a two wave panel data, we divide the bias at time 2 (=wave 2) by the bias at time 1 (=wave 1). The relative factors are the important parameters for the convergence of a distribution to its steady-state distribution on the state space. It tells us the possible number of waves that are necessary to get a possible reduction of the non-response bias. We denote this speed factor by λ_{1com} which is computed by:

$$\begin{aligned} \lambda_{1,com} &= \frac{Bias(\hat{b}_{2,com})}{Bias(\hat{b}_{1,com})} \\ &= \frac{b_1\beta^2(\kappa+\rho(1-\kappa))(\gamma+\phi(1-\gamma))\sigma^2/(\alpha+\beta a_1)^2 - \beta^2 b_1^2(\kappa+\rho(1-\kappa))^2}{\beta^2\sigma^2 b_1/(\alpha+\beta a_1)^2 - \beta^2 b_1^2}, \end{aligned}$$

Further simplifying, we get the simplified form of the ratio λ_{1com} as follows:

$$\lambda_{1,com} = \frac{((\alpha+\beta a_1)^2 - \beta^2 b_1^2)(\kappa+\rho(1-\kappa))(\gamma+\phi(1-\gamma))}{(\alpha+\beta a_1)^2 - \beta^2 b_1^2(\kappa+\rho(1-\kappa))^2}. \tag{2.32}$$

Small values of λ_{1com} indicate a high fade-away effect of the initial non-response bias.

Interpretation of the fade-away phenomenon under various scenarios:

- No bias
  - If $\sigma^2 = 0$, the probability of participating depends only on X, then there is no bias in the OLS estimates. The larger the σ, the larger is the bias.
  - If β is zero there is no bias present in OLS estimates.
- In case of independent covariate values ($\kappa = \rho = 0$) the bias would vanish in one period over time.
- In case of independent values of the residual term (small values of γ and ϕ) the bias would vanish.
- If permanent components of covariates and error terms are present ($0 < \kappa < 1$ or $0 < \gamma < 1$) then there will be always some kind of permanent bias present during the follow-ups.
- The speed of convergence to a steady-state distribution is fast in the presence of low stability (covariates or residual terms), e.g., when $\kappa = \gamma = \rho = \phi = 0.10$ the bias after one panel wave is almost zero.
- For moderate stability, the initial non-response bias decreases in a geometrical pattern in later panel waves. For example, in scenario $\kappa = \gamma = \rho = \phi = 0.50$, or in $\kappa = \gamma = \rho = \phi = 0.70$ the initial bias decreases in a geometrical fashion in following panel waves.
- There is no fade-away phenomenon present in the presence of high stability, e.g., when $\kappa = \gamma = \rho = \phi = 0.90$ the distorting effects of initial non-response don't faded-away in subsequent panel waves.
- **The fade-away of bias under various scenarios of stability (mixed cases):**
  - If the stability of covariate components are in between $0.10 \leq \kappa \leq \rho \leq 0.90$ and the stability of residual components are $\gamma = \phi = 0$. Then the speed of the fade-away effect is fast.

- If the stability of covariate components are $\kappa = \rho = 0$ and the stability of residual components are in between $0.10 \leq \gamma \leq \phi \leq 0.90$. Then the speed of the fade-away effect is fast.

- If κ is in between $0.10 \leq \kappa \leq 0.90$ other stabilities being $\gamma = \rho = \phi = 0$, or if ρ is in between $0.10 \leq \rho \leq 0.90$ other stabilities being $\kappa = \gamma = \phi = 0$. Then the speed of the fade-away effect is fast.

- If γ is in between $0.10 \leq \gamma \leq 0.90$ other stabilities being $\kappa = \rho = \phi = 0$, or if ϕ is in between $0.10 \leq \phi \leq 0.90$ other stabilities being $\kappa = \gamma = \rho = 0$. Then the speed of the fade-away effect is fast.

- If $\kappa = \gamma = 0.10$ and $\rho = \phi = 0$, or if $\kappa = \gamma = 0$ and $\rho = \phi = 0.10$ estimates having substantial initial non-response converge to the true solution without initial non-response just after one panel wave. Thus the speed of the fade-away effect is fast.

- In scenario say when $0.50 \leq \kappa \leq \gamma \leq 0.70$ and $\rho = \phi = 0.10$, the initial bias decreases in a geometric sequence in later panel waves. However, if $0.50 \leq \rho \leq \phi \leq 0.70$ and $\kappa = \gamma = 0.10$ the speed of the fade-away effect is very fast then the speed of the fade-away effect in scenario $0.50 \leq \kappa \leq \gamma \leq 0.70$ and $\rho = \phi = 0.10$. (Note: The smaller the value of κ or γ the higher is the speed of the fade-away effect).

- If the size of permanent components is large say $0.80 \leq \kappa \leq \gamma \leq 0.90$ and whatever is the size of transient components in the interval $0 \leq \rho \leq \phi \leq 0.90$, then the distribution of permanent components stays stable and therefore the distorting effects of initial non-response don't fade-away in later panel waves. While, this doesn't hold for the transient components which swing into a steady-state distribution. For example, when the size of transient components are between $0 \leq \rho \leq \phi \leq 0.90$ and the size of permanent components are very low say in the interval $0 \leq \kappa \leq \gamma \leq 0.90$, then over time the distorting effects of initial non-response melts down to zero just after one panel wave.

2.3.2.2. Coefficient of determination for the regression model

The coefficient of determination which is denoted by R^2 is a statistical measure, which is used to explain how much of the variability of the response variable is caused by its linear relationship to the explanatory variable.

The most general definition of R-Squared for the regression model is:

$$\begin{aligned} R^2 &= \frac{\text{Explained variation}}{\text{Total Variation}}, \\ &= \frac{b_t^2 Var(X_t)}{b_t^2 Var(X_t) + Var(e_t)}. \end{aligned} \tag{2.33}$$

In the following table, we calculate R^2 statistic for various values of residual variance σ^2. Assuming $b_t = 1$ and $Var(X_t) = 1$. R^2 is always between 0 and 100%. In our case R^2 is between 50% and 100%, which means that the regression model accounts for 50% of the variance when residual variance σ^2 is 1, while it accounts for 100% of the variance when the residual variance σ^2 is 0. The more the variance that is accounted by the regression model is the closer the data points to the fitted regression line. If the regression model could explain 100% (as well as in Table 1, when σ^2 is 0, the model accounts for 100% of the variance) of the variance, the fitted values would always equal to the observed values and all the data points would fall on the fitted line.

Table 1: Goodness-of-fit statistic R^2 for various values of σ^2

Residual variance σ^2	Percent of variance explained R^2
0.00	1.00
0.10	0.91
0.20	0.83
0.30	0.77
0.40	0.71
0.50	0.67
0.60	0.63
0.70	0.59
0.80	0.56
0.90	0.53
1.00	0.50

2.3.3. Panel model and panel estimates

In Subsection 2.3.1 of this chapter, we had provided an overview of the linear regression model for cross-sectional data. Our research, however, also focuses on linear panel models for longitudinal data. This subsection, therefore, addresses on linear panel models.

One of the unavoidable problems of the cross-sectional OLS estimator is that it doesn't account for the unobserved individual heterogeneity and thus the effect is neglected in regular OLS. Although, ignoring to control for such heterogeneity effects in the model may sometimes produce bias in the estimates. The pooled OLS estimator is simply an OLS technique run on panel data. It also doesn't control for the unobserved individual heterogeneity and thus the effect is simply ignored in pooled OLS.

For that reason, many assumptions about the error term are violated such as the orthogonality of the error term. Therefore, ignoring the individual effects in the pooled OLS may create bias in the estimates. Random effects model (see below)

solves this problem by implementing individual-specific effects in the regression model which are assumed to be random. However, the exogeneity assumption in the random effects model is extreme. In fact, every statistical model has some endogeneity problems, in such a situation the fixed effects model (see below) is the best method that gives us consistent parameter estimates. In the following, we distinguish between these two panel models.

Suppose the relationship between the response variable $Y_{i,t}$ and the set of K explanatory variables $X_{it,k}$ together with an error components model $\varepsilon_{i,t} = v_i + \eta_{i,t}$, can be modeled by the following linear regression:

$$Y_{i,t} = \beta_0 + \beta_1 X_{it,1} + \beta_2 X_{it,2} + \ldots + \beta_K X_{it,K} + v_i + \eta_{i,t}, \tag{2.34}$$

The index i stands for unit, k denotes the number of covariates and t refers to time periods. In addition, the error term $\varepsilon_{i,t}$ is decomposed into two components $v_i + \eta_{i,t}$. The individual-specific error component (fixed effects) v_i and the idiosyncratic error component of the individual which changes over time $\eta_{i,t}$ which is assumed to be normally distributed with mean 0 and variance σ_η^2 i.e., $\eta_{i,t} \overset{iid}{\sim} N(0, \sigma_\eta^2)$. The individual-specific error component v_i represents all unobservable time-invariant individual heterogeneity, while idiosyncratic error component $\eta_{i,t}$ of the individual comprises all unobserved factors of the individual that disturbs $Y_{i,t}$. In order to model the individual-specific heterogeneity v_i in the panel, we define these two models as follows:

Random effects (RE) model: It assumes that v_i are random variables and are uncorrelated with the covariates $X_{it,k}$ and with the idiosyncratic error term η_{it}. The value v_i is specific for the person i. The v's of different persons are independent and have a mean of zero, and their distribution is assumed to be normally distributed $v_i \overset{iid}{\sim} N(0, \sigma_v^2)$. Mathematically, the model in equation (2.34) is said to be RE model if the following orthogonality condition is satisfied:

$$Cov(v_i, X_{it,k}) = 0, \quad \text{for all} \quad k, i \quad \text{and} \quad t. \tag{2.35}$$

As long as the regressors are uncorrelated with the individual effects v_i and the error term $\eta_{i,t}$, we can get unbiased, consistent, and efficient parameter estimates by using generalized least squares (GLS) for fixed values of σ_v^2 and σ_η^2. However, in practice

the values of σ_v^2 and σ_η^2 are unknown so that GLS is not feasible, then in such a case σ_v^2 and σ_η^2 can be estimated by using the method called feasible generalized least squares (FGLS). Detailed introduction on GLS and FGLS estimators can be found in Grubb and Magee (1988) and Li et al. (2011) etc.

Fixed effects (FE) model: The RE model is based on the orthogonality assumption, consistency involves that the unobserved effects v_i should be uncorrelated with the observed covariates $X_{it,k}$ included in the model. However, if v_i is correlated with $X_{it,k}$, such that:

$$Cov(v_i, X_{it,k}) \neq 0, \quad \text{for all} \quad k, i \quad \text{and} \quad t. \tag{2.36}$$

Then regression parameters can be more efficiently estimated by using the FE model. A more detailed introduction on the RE and FE in linear panel data models, see the textbook of Hsiao (1986), Arellano (2003) and Wooldridge (2009).

There are four different estimation methods of the FE model that give consistent estimators of the model parameters even when the FE v_i are correlated with the covariates $X_{it,k}$. These estimation methods are: The least squares dummy variable estimator; the FE or the Within estimator; the first-difference estimator; and the orthogonal deviations estimator. One of the motives of our research is to check the fade-away effect of the panel model estimators, especially the fade-away effect of the FE estimator. This subsection, therefore, addresses on the FE estimator which is also known as the Within estimator. The FE estimator is based on various steps. According to this method first we calculate the mean of each variable over time in the model, we then subtract the mean of each variable from the observed values in the model. This procedure is also known as Within transformation. Since the individual FE v_i is canceled out in differencing, so applying OLS on the transformed model will consistently estimate regression parameters.

In other words, to illustrate the Within estimator analytically consider again the multiple linear regression model in equation (2.34):

$$Y_{i,t} = \beta_0 + \beta_1 X_{it,1} + \beta_2 X_{it,2} + \ldots + \beta_K X_{it,K} + v_i + \eta_{i,t}, \tag{2.37}$$

with $E(v_i X_{it,k}) \neq 0$ and $E(\eta_{it} X_{it,k}) = 0$.
In the FE Within transformation the effect of v_i is removed from the model by taking

the mean of each cross-section i over time t, i.e., in terms of cross-section means we get the following model:

$$\bar{Y}_i = \beta_0 + \beta_1 \bar{X}_{i,1} + \beta_2 \bar{X}_{i,2} + \ldots + \beta_K \bar{X}_{i,K} + v_i + \bar{\eta}_i, \tag{2.38}$$

where, $\bar{Y}_i = \frac{1}{T}\sum_{t=1}^{T} Y_{it}, \bar{X}_{i,k} = \frac{1}{T}\sum_{t=1}^{T} X_{it,k}$ and $\bar{\eta}_i = \frac{1}{T}\sum_{t=1}^{T} \eta_{it}$, for $k = 1,2,3,\ldots,K$ and $t = 1,2,3,\ldots,T$.
Now, if we subtract equation (2.38) from equation (2.37), we obtain:

$$Y_{i,t} - \bar{Y}_i = \beta_1(X_{it,1} - \bar{X}_{i,1}) + \beta_2(X_{it,2} - \bar{X}_{i,2}) + \ldots + \beta_K(X_{it,K} - \bar{X}_{i,K}) + (\eta_{i,t} - \bar{\eta}_i), \tag{2.39}$$

or, by using the notation $\tilde{Y}_{it} = (Y_{it} - \bar{Y}_i)$, $\tilde{X}_{it,1} = (X_{it,1} - \bar{X}_{i,1})$, $\tilde{X}_{it,2} = (X_{it,2} - \bar{X}_{i,2})$, $\ldots$, $\tilde{X}_{it,K} = (X_{it,K} - \bar{X}_{i,K})$ and $\tilde{\eta}_{it} = (\eta_{it} - \bar{\eta}_i)$ we get the simplified form or demeaned model as follows:

$$\tilde{Y}_{i,t} = \beta_1 \tilde{X}_{it,1} + \beta_2 \tilde{X}_{it,2} + \ldots + \beta_K \tilde{X}_{it,K} + \tilde{\eta}_{i,t}. \tag{2.40}$$

Finally, in the transformed model the effect of the unobserved variable v_i is removed because it is time-invariant. Now applying OLS regression on the transformed model will consistently estimate model parameters.

Similarly, in matrix notation the demeaned model in equation (2.40) can be written by:

$$\tilde{Y} = \tilde{X}\beta + \tilde{\eta}, \tag{2.41}$$

where $\tilde{Y}$ is the $NT \times 1$ column vector of observations on the response variable, $\tilde{X}$ is a data matrix of order $NT \times K$ of N observations of the K explanatory variables, β is the $K \times 1$ column vector of regression slope coefficients, and $\tilde{\eta}$ is the $NT \times 1$ column vector of the error term. By applying the OLS regression on the transformed model, we get the FE Within estimator of β in vector form:

$$\hat{\beta}^{FE} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{Y}, \tag{2.42}$$

where, $\hat{\beta}^{FE} = [\hat{\beta}_1^{FE}, \hat{\beta}_2^{FE}, \hat{\beta}_3^{FE}, \ldots, \hat{\beta}_K^{FE}]'$. However, one drawback of the Within

estimator is that it can't estimate the individual FE. This is because all time-invariant variables are removed from the model by demeaning the variables within their group by using the Within transformation. As the FE model annihilate the individual FE v_i by demeaning the variables within their group by using the Within transformation. So one can't estimate FE directly from the model. Nevertheless, it can be recovered by using the following formula:

$$\hat{v}_i^{FE} = \bar{Y}_i - \hat{\beta}_1^{FE}\bar{X}_{i,1} - \hat{\beta}_2^{FE}\bar{X}_{i,2} - \ldots - \hat{\beta}_K^{FE}\bar{X}_{i,K}. \tag{2.43}$$

2.3.4. Non-linear models

As at the start of this section, we have shortly discussed that the basic idea of regression is to find the relationships if any, that exists between a response variable and one or more independent variable(s). Based on the type of relationships between the response and the predictor variables regression models may be broadly divided into two categories: the linear regression models and the non-linear regression models. In both types of models, the idea is the same that is to relate the response variable to the predictor variables through some parameters. When the model is linear in parameters, the linear regression analysis is the most powerful and widely used technique for causal inference and prediction. Whereas, non-linear models are designated by the fact that the models are non-linear in the parameters. In such situations, the methods of linear regression should be extended which introduces much complexity. In the following section we are discussing the most commonly used non-linear probability models (NLPMs) such as the logit, probit, ordered logit and ordered probit models.

2.3.4.1. The typical non-linear probability models

The OLS regression techniques are appropriate for continuous dependent variables such as age, height, weight, and income, etc. When the dependent variable is binary or ordinal other regression methods should be used for the estimation of regression parameters. In fact, the most important methods in this context are the logit/probit models. We use logit/probit models when the response variable is binary and ordered logit/probit models if it is ordinal or have at least more than two possible categories.

Regression models for ordered responses and methods have a long-lasting history

in the literature. They are very common in the field of biological sciences and social sciences. The early work on ordinal choice models in the biological sciences is pioneering the work of Aitchison and Silvey (1957) who proposed the ordered probit model to analyze experiments in which the responses of subjects to various doses of stimulus are divided into ordinal order. Snell (1964) suggested the use of logistic instead of a normal distribution of errors as an approximation for the mathematical simplification. While the early work on ordinal choice models in the in the field of social sciences is pioneering work of McKelvey and Zavoina (1975), who actually extended the probit ordered model proposed by "Aitchison and Silvey" to more than one explanatory variable. Their basic idea is to assume that there exists a continuous latent response variable which is related to a single index of explanatory variables and unobserved error term, so that one can obtain an observed ordinal response variable from the underlying continuous latent response variable by dividing it into finite number of intervals or cut points (thresholds).

McCullagh (1980) proposed the proportional odds (PO) model, which is also labeled as the cumulative odds model. The PO model belongs to the class of generalized linear models and is commonly used for the analysis of ordinal data. The proportional odds model is used to estimate the odds of being at or below a particular level of the response variable. McCullagh, directly modeled the cumulative probabilities of the ordered categories of outcome variable which are associated with the covariates via a linear predictor using a monotone link function. The most common link functions are the logit, probit and the complementary log-log.

Fullerton (2009) provides a detailed overview on several logistic regression models for ordinal data and their application in the field of sociology. The textbook of O'Connell (2006) provides applied researchers to the field of social, educational, and behavioral sciences with accessible and detailed coverage of analyses for ranked outcomes, while the textbook of Agresti (2010) on the analysis of categorical data provides a detailed treatment of the important methods for ordinal data. A further comprehensive discussion on the regression models for ordinal data can be found in Amemiya (1981), Maddala (1983), Winship and Mare (1984), Aldrich and Nelson (1984), Liao (1994), Long (1997), Bandeen-Roche et al. (1997), Williams (2006), Greene (2008) and Breen et al. (2018). We would first discuss logit and probit models for binary data and then will extend the model discussion for ordinal response data.

2.3.4.2. A latent variable formulation of binary logit and probit models

The logit and probit regression models are the appropriate technique when the response variable is binary such as: yes and no, agree and disagree, success and failure, etc for convenience these categories or outcomes are coded as 1 or 0. The logit model specifies the conditional mean response of a response variable as a logit function of model covariates. The probit model did the same with the slight difference that here the distribution of error term is standard normal instead of the error having logit distribution. Another standard approach of using logit/probit regressions is the latent variable model approach. This approach generates an ordinal dependent variable through the categorization of an underlying latent continuous variable Hosmer and Lemeshow (2000).

To understand the latent variable approach, let us suppose we have a continuous response variable, say $Y^\star$, which is linearly related to the K explanatory variables $X_{i0}, X_{i1}, X_{i2}, \ldots, X_{iK}$ through the parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_K$. Then we would write the model as follows:

$$Y_i^\star = \beta_0 + \sum_{k=1}^{K} \beta_k X_{ik} + \varepsilon_i, \tag{2.44}$$

where the first explanatory variable $X_{i0} = 1$ is a constant unless otherwise stated and $\beta_0, \beta_1, \beta_2, \ldots, \beta_K$ are the $K+1$ model parameters including the intercept β_0. The term ε is the residual term of the model. The ε_i are assumed to be independent and identically normally distributed variables having mean zero and constant variance σ_ε^2, written as $\varepsilon \overset{iid}{\sim} N(0, \sigma_\varepsilon^2)$. However, if $Y^\star$ is unobserved or latent in the interval from minus infinity to plus infinity, but it generates an observed variable Y which is equal to one if $Y^\star$ exceeds then a specific cut point or constant of $Y^\star$ (namely, the threshold parameter whose value is unknown) say, τ and zero if it is less than or equal to a threshold, i.e.,:

$$Y_i = \begin{cases} 1, & \text{if } Y_i^\star > \tau, \\ 0, & \text{if } Y_i^\star \leq \tau. \end{cases} \tag{2.45}$$

A universal method that is used for the estimation of such models is the method of ML estimation. This method requires assumptions about the distribution of residual terms. Mainly, the assumption that the residual term has a logit distribution follows a logit model, while the normality assumption of residual term results in a probit

model Long (1997). In both cases the concept is the same that is by fitting the model as follows:

$$h(Pr(Y_i = 1)) = \hat{\beta}_0 + \sum_{k=1}^{K} \hat{\beta}_k X_{ik}, \tag{2.46}$$

where the function $h(.)$ stands for the logit or probit transformation; i.e., the inverse of the distribution function of the standard logistic or the standard normal distribution. By using the latent variable approach of binary choice models, we redefine the residual term of the underlying linear regression model in equation (2.44) as $\varepsilon = s\zeta$. In the logit model, it is assumed that ζ is a standard logistic random variable with mean zero and variance $\pi^2/3 \approx 3.29$ and s is a scale factor, providing a variance $\sigma_\varepsilon^2 = s^2\pi^2/3$ for the residual term of the underlying linear model in equation (2.44). While, in the probit model it is assumed that ζ is a standard normal random variable with mean zero and variance one, providing a variance $\sigma_\varepsilon^2 = s^2$ for the residual term of the underlying linear model in equation (2.44).
Thus under the assumption that ζ to be a standard logit random variable, it induces the standard logit model of response as a function of the covariates:

$$\begin{aligned} P(Y_i = 1|X_i) = P(Y_i^\star > \tau|X_i) &= P\left(\frac{\varepsilon}{s} > -\left[\frac{\beta_0 - \tau}{s} + \frac{\sum_{k=1}^{K}\beta_k}{s} X_{ik}\right]\right) \\ &= \frac{exp\left(\frac{\beta_0-\tau}{s} + \frac{\sum_{k=1}^{K}\beta_k}{s} X_{ik}\right)}{1 + exp\left(\frac{\beta_0-\tau}{s} + \frac{\sum_{k=1}^{K}\beta_k}{s} X_{ik}\right)}, \end{aligned} \tag{2.47}$$

Now by taking the logarithm on both sides of equation (2.47), we get the logistic regression model:

$$logit(P(Y_i = 1|X_i)) = \frac{\beta_0 - \tau}{s} + \frac{\sum_{k=1}^{K}\beta_k}{s} X_{ik}. \tag{2.48}$$

Similarly, under the assumption that ζ to be a standard normal random variable, it

produces the standard probit model of response as a function of the covariates:

$$\begin{aligned} Probit(P(Y_i = 1|X_i)) &= P(Y_i^\star > \tau|X_i) \\ &= \Phi\left[\frac{\beta_0 - \tau}{s} + \frac{\sum_{k=1}^{K} \beta_k}{s} X_{ik}\right], \end{aligned} \tag{2.49}$$

where the function $\Phi(.)$ stands for the cumulative distribution function (CDF) of the standard normal distribution.

2.3.4.3. A latent variable formulation of ordered logit and probit models

So far, we have discussed logit and probit models for binary response data. We now focus our attention to ordered logit and probit regression models for ordinal response data. Ordinal logistic and probit models are a generalization of binary logistic and probit models, when the response variable has more than two possible discrete values or categories that contain ordinal information. Ordinal logistic regression models the relationship between a set of predictors and an ordinal response variable. These models are appropriate when the response variable takes more than two possible categories or outcomes with a natural sequential order of response values, such as single, married, divorced, separated and widowed.

Another example is the ranking and rating levels of health satisfaction in the German Socio Economic Panel (SOEP). In the SOEP survey, an individual's health satisfaction level is based on by answering the question "how much you are satisfied with your health keeping all other things being equal". An individual would express some degree of agreement/disagreement by choosing one of eleven options ranging from 0 to 10. Where 0 means completely dissatisfied and 10 means completely satisfied, while the intermediate numbers 1 to 5 and 6 to 9 means less satisfied and greatly satisfied, respectively.

By considering the continuous latent variable model in equation (2.44) without the intercept β_0:

$$Y_i^\star = \sum_{k=1}^{K} \beta_k X_{ik} + \varepsilon_i, \tag{2.50}$$

where $Y_i^\star$ is the unobserved latent response variable, but it generates an observed response variable having J categories for the i^{th} subject through the following censoring mechanism:

$$Y_i = \begin{cases} 0, & \text{if } -\infty < Y_i^\star \leq \tau_0 \\ 1, & \text{if } \tau_0 < Y_i^\star \leq \tau_1 \\ 2, & \text{if } \tau_1 < Y_i^\star \leq \tau_2 \\ \dots & \\ J, & \text{if } \tau_{J-1} < Y_i^\star \leq \infty, \end{cases} \tag{2.51}$$

where τ's are the unknown cut-off points namely the so-called model threshold parameters which divide the range of $Y^\star$ into disjoint and exhaustive intervals such that $\tau_0 < \tau_1, \dots, \tau_{J-1}$. Note, that the model in equation (2.51) doesn't include the constant β_0, because one can't estimate simultaneously the constant of the linear predictor as well as the thresholds parameters. This identification problem can be solved by removing either the constant of the linear predictor or setting the first threshold equal to zero.

As we have discussed earlier the difference between logit and probit models is the distribution of error terms. If ε is a logistic random variable then the cumulative probabilities of response will correspond to an ordered logit model. While, if ε follows a standard normal distribution then these probabilities will correspond to an ordered probit model.

For the ordinal response variable Y_i having J categories, the ordered logit model in logit form (Long, 1997) can be written as follows:

$$\begin{aligned} logit[P(Y_i \leq j|x_1, x_2, x_3, \dots, x_k)] &= log\left(\frac{P(Y_i \leq j|x_1, x_2, x_3, \dots, x_k)}{1 - P(Y_i \leq j|x_1, x_2, x_3, \dots, x_k)}\right) \\ &= \tau_j + \sum_{k=1}^{K} \beta_k X_{ik}, \end{aligned} \tag{2.52}$$

where $P(Y_i \leq j|x_1, x_2, x_3, \dots, x_k)$ is the cumulative probability of being at or below than category j, given a vector of covariates. τ_j are the thresholds parameters and $\beta_1, \beta_2, \beta_3, \dots, \beta_K$, are the logit coefficients.

Further simplifying equation (2.52), we get the cumulative probability of response

of being at or below category j:

$$P(Y_i \leq j) = \frac{\exp(\tau_j + \sum_{k=1}^{K} \beta_k X_{ik})}{1 + \left[\exp(\tau_j + \sum_{k=1}^{K} \beta_k X_{ik})\right]}, \quad \text{for} \quad j = 0, 1, 2, \ldots, J-1. \tag{2.53}$$

If the distribution of ε is a standard normal having mean zero and unit variance then the cumulative probabilities of response will be correspond to an ordered probit model. Therefore, the probability of response at a particular category j will be:

$$\begin{aligned} P(Y_i = j) &= \Phi(\tau_j + \sum_{k=1}^{K} \beta_k X_{ik}) - \Phi(\tau_{j-1} + \sum_{k=1}^{K} \beta_k X_{ik}), \\ &= 1 - \Phi(\tau_{j-1} + \sum_{k=1}^{K} \beta_k X_{ik}). \end{aligned} \tag{2.54}$$

For $j = 0, 1, 2, \ldots, J-1$.

2.4. Treatment of attrition

2.4.1. Weighting

In most surveys, a set of weights is a key component in order to produce unbiased population estimates, and correct for imperfections in the sample to estimate the descriptive statistics and the regression models. The regression coefficients obtained from the un-weighted least squares estimators may be biased if the inclusion probability of units in the sample is correlated with the outcome variable conditional on the explanatory variables. However, weighting by the reciprocals of the unit inclusion probabilities, the bias to be corrected and regression coefficients to be estimated consistently, see Section 6.3 of Fuller (2009). The weighted estimator in the regression models are used from different points of view: First, if the included variables in the model are measured with an error then the use of weighted regression is appropriate. Second, if the OLS assumption of homoscedastic error variance is violated, so the purpose of weighting, in this case, is to correct for the problem of heteroscedasticity and to achieve a more precise estimation of the coefficients of

regression models. Third, for all the available data if fewer than all the available data is used in the estimation, then a zero weight is assigned to each observation of the non-used data, while a weight of one is assigned to each observation used in the estimation.

2.4.2. Inverse probability weighting

Inverse probability weighting (IPW) is a standard to reduce non-response bias either by non-response analysis or by calibration. Now, the fade-away hypothesis assumes that an initial bias reduces by itself. So the question arises which gain we achieve from using the IPW.

Non-response may introduce bias in the estimation of population parameters when a large part of the data is missing. One of the goals of this thesis is to reduce non-response bias in the estimation of population parameters. There are various other approaches for dealing with missing data. Of these approaches, the simplest method for dealing with the analysis of incomplete data is the complete case method. According to this approach researchers completely ignore observations with missing values and use only the available cases of the data for the analysis. This method performs well when the data are MCAR, otherwise, it may lead to inconsistent parameter estimates Nakai and Weiming (2011).

Regression methods are widely used procedures that are used to reduce this non-response bias, by using prior information on model covariates. As in surveys, information on the non-respondents is not available for the outcome variable, however, for the covariates these information are usually available e.g., in the case of sampling from register data. Therefore, in order to reduce non-response bias, it is appropriate to model the response behaviour and incorporate the covariate information into the estimation.

Regression weighting is also widely used procedures in surveys to compensate for non-response/attrition bias or reduce the distorting effect of initial non-response. For a detailed overview on the use of weighting in the presence of non-response/attrition, see Kalton and Flores-Cervantes (2003) and Brick (2013). There exist several weighting methods in the context of regression analysis which are used to account for non-response/attrition. One such naive method is the inverse probability weighting (IPW), which is widely used in applications where data are missing due to non-

response and in the estimation of causal effects. An important tool in the construction of weights is inverse probability weighting, which is defined here as weighting by the inverse of the estimated response probability. In order to define this probability, we use previously defined R and Z (where Z is the vector of covariates which are always observed). Mathematically, it can be written as:

$$P(R_i = 1|Y_i, Z_i) = P(R_i = 1|Z_i), \tag{2.55}$$

where R_i selects out the observed data points.

The IPW approach has some similarity with the early work of Horvitz and Thompson (1952), who initially proposed the idea for the estimation of the population mean in a design-based approach, see for example Särndal et al. (1992, p. 42). IPW has been frequently used for regression models with data, especially when non-response or attrition process is assumed to depend only on the observed covariates, i.e., when outcomes are MAR (this assumption is defined earlier) in the sense of Rubin (1976). Later on the use of IPW to control for non-response/attrition under the MAR assumption has been widely used by several authors some of them are: Robins and Rotnitzky (1995) use the IPW estimator in the context multivariate regression models with missing data. Robins et al. (1995) suggested a class of inverse probability of censoring weighted estimators for the model parameters, and show that how theses estimators can be used to estimate the conditional mean in the presence of attrition in panel data. For a further detailed overview on the use of IPW to control for non-response/attrition under the MAR assumption, see for instance Rotnitzky and Robins (1995), Horowitz and Manski (1998), Scharfstein et al. (1999), Abowd et al. (2001), Hirano et al. (2003) and Wooldridge (2002, 2007), among many others. While in the case, when non-response/attrition is assumed to depend not only the observed covariates but also on the unobserved covariates, the work has been discussed very rarely. Of these: Das (2004) has examined a non-parametric regression model with an additive individual-specific component for panel data where attrition depends not only on the time-varying observables but also depends on the time-invariant unobservable which is potentially correlated with the individual effect. DiNardo et al. (2006) use conventional sample selection correction estimator techniques based on regression that assumes partial randomization of non-response (see sample selection models of Heckman: Heckman (1976, 1979)). These authors

discuss the application and usefulness of the IPW under attrition on missing on unobservables in an experimental context.

Part II

Simulation Data

CHAPTER 3

Fade-Away Effect of the Regression Estimators using Simulation Data

3.1. Motivation

In this chapter, a simulation study is conducted to investigate the fade-away effect of the initial non-response bias in a multi-wave panel survey. The purpose of the simulation study is to investigate the accuracy of the bias approximation in a simulation setting and to check the size of the fade-away effect in later panel waves with no analytical bias approximation. Alho (2015) has investigated the bias of cross-sectional OLS estimates under not missing at random (NMAR) non-response at the start of the panel. He derived analytical bias approximation for the OLS estimate of the slope coefficient of the variable of interest. His underlying model used a variance component model with two components: a fixed individual component and an auto-regressive shock component (Alho's model was discussed in Subsection 2.3.2 of Chapter 2). However, in multi-wave panel surveys, the analytical expression for Alho's bias approximation formula becomes intractable for later waves. Therefore, we extend the results to a longer panel wave via a simulation study.

The remainder of this chapter is structured as follows: Section 3.2, presented an extension of the fade-away phenomenon to a multi-wave panel survey. To judge

the performance of the estimators, a Monte Carlo simulation study is conducted in Section 3.3. In Section 3.4, we discussed the fade-away effect of the cross-sectional OLS estimators with and without panel attrition. We then compared the estimation results of weighted and un-weighted cross-sectional OLS estimators in Section 3.5. In the next section (Section 3.6) we discussed the behaviour of different panel estimators. Finally, Section 3.7 is devoted to the estimation of the non-linear ordered logit model. Here we compared the fade-away effect of the un-weighted estimates of ordered logit model with several weighting approaches. The estimation is done with **SAS** and with the procedure: **PROC REG**, **PROC LOGISTIC** and **PROC PANEL**, respectively.

3.2. Extension of the fade-away phenomenon to a multi-wave panel survey

Motivated by the fade-away phenomenon, as described in Subsection 2.3.2 of Chapter 2, we extend Alho's (2015) model to multi-wave panel in the framework of regression analysis. In order to verify the approximate results of Alho and to check the size of the bias in later panel waves through a simulation study, we proceed as follows. First, let us consider two finite samples drawn from the same population of sufficiently large size. Further, let the members of the two samples follow a Markov chain model and change their state according to the same transition probabilities. The first sample consists of all individuals (respondents and non-respondents) who are initially selected to the panel, we denote as Full-Sample. This sample consists of information on both the respondents and the non-respondents who were in the target population, so there is no bias in the Full-Sample estimates. While the second sample consists of individuals (only respondents) who participated in the survey initially, we denote this by Resp-Sample (respondent sample). Further, we assume that the distribution of the variable of interest is highly selective for response in wave one of a panel in the Resp-Sample, or non-response is not missing at random (NMAR) of the start of the Resp-Sample. For the initial non-response of a person i we suppose that it depends on the $Y_{i,1}$.

Let $R_{i,1}$ be the binary response indicator of a person i which is equal to one

$R_{i,1} = 1$ if the person is willing to responds at the start of the panel, and $R_{i,1} = 0$ otherwise. Then following the linear model of Alho (2015), the conditional response probability of a person i is defined by:

$$P_{i,1} = P(R_{i,1} = 1|Y_{i,1}) = \begin{cases} 0, & \text{if} \quad \alpha + \beta Y_{i,1} < 0, \\ \alpha + \beta Y_{i,1}, & \text{if} \quad 0 \leq \alpha + \beta Y_{i,1} \leq 1, \\ 1, & \text{if} \quad \alpha + \beta Y_{i,1} > 1, \end{cases} \tag{3.1}$$

where α and β are parameters. These non-response parameters are to be selected in such a way that the average probability of response from equation (3.1) is approximately of about 0.60 to 0.70. The reason for choosing the initial response probability to be between 0.60 and 0.70 is that we want to generate a substantial initial non-response in the first wave of a Resp-Sample. Otherwise, if non-response at the start of the Resp-Sample is not substantial or ignorable for the estimation of regression coefficients b_t, the distribution of the Resp-Sample would be equal to the distribution of the Full-Sample. In such a scenario there would be no bias in the Resp-Sample at any panel wave and hence there would be no fade-away phenomenon present. Therefore, we assume that the initial non-response is not NMAR or highly selective for the estimation of population parameters. And therefore the initial distribution of the Resp-Sample at the start of the panel is somewhat away from the distribution of the Full-Sample. Under this respect, if there is no further drop out after wave 1, then according to the results of the contraction theorem (see, Theorem 1 in Section 2.2 of Chapter 2) the distorting effects of initial non-response on the distribution of $Y_{i,t}$ in the Resp-Sample has to fade-away over time $t = 2, 3, 4, \ldots, T$.

In a similar fashion, it is assumed that all those who participated in the survey initially, will also take part in the survey as long as the survey ends. Let $R_{i,t}$ be the binary response indicator of the person i at time t $(t = 2, 3, 4, \ldots, T)$ which is equal to one $R_{i,t} = 1$ if the person i is agrees to participate and $R_{i,t} = 0$ otherwise. Assume that the participation probability depends on $Y_{i,t}$, for some attrition parameters $\alpha^\star$

and $\beta^\star$ which is defined by

$$P_{i,t} = P(R_{i,t} = 1|Y_{i,t}) = \begin{cases} 0, & \text{if} \quad \alpha^\star + \beta^\star Y_{i,t} < 0, \\ \alpha^\star + \beta^\star Y_{i,t}, & \text{if} \quad 0 \leq \alpha^\star + \beta^\star Y_{i,t} \leq 1, \\ 1, & \text{if} \quad \alpha^\star + \beta^\star Y_{i,t} > 1, \end{cases} \tag{3.2}$$

For, $t = 2, 3, 4, \ldots, T$. The parameters $\alpha^\star$ and $\beta^\star$ are chosen in such a way that the average response probability from equation (3.2) is approximately 0.90. Moreover, $\beta^\star$ is the selective effect of the panel attrition and hence large a value of $\beta^\star$ results in a large attrition probability and slow-down the speed of the fade-away effect.

3.3. Simulation study

A Monte Carlo simulation study is conducted to evaluate the fade-away effect of the proposed regression estimators up to ten panel waves. To generate the data from the model, we set the starting values of the regression model parameters to $a_t = 2$ for the intercept and $b_t = 1$ for the slope. We then generate a synthetic data set of size $N = 1{,}000$ units from the model, which is replicated over 100 Monte Carlo repetitions. Further, we assume that the distribution of the covariate $X_t = M + Z_t$ and the error term $e_t = V + U_t$ are normally distributed, and thus the data are generated from different univariate normal distributions. The covariate X_t consists of two components: the permanent component M and the transient component Z_t, both the components are uncorrelated with each other having expectations $\mu_m = \mu_{z_t} = 0$, and variances $\sigma_m^2 = \kappa$ and $\sigma_{z_t}^2 = (1 - \kappa)$, respectively.

$$M \sim \ N(0, \sigma_m^2) \quad \text{and} \quad Z_t \sim \ N(0, \sigma_{z_t}^2),$$

Likewise, the error term e_t is decomposed into two uncorrelated components: the permanent component V and the transient component U_t having expectations $\mu_v = \mu_{u_t} = 0$, and variances $\sigma_v^2 = \gamma\sigma^2$ and $\sigma_{u_t}^2 = (1 - \gamma)\sigma^2$, respectively.

$$V \sim \ N(0, \sigma_v^2) \quad \text{and} \quad U_t \sim \ N(0, \sigma_{u_t}^2),$$

Further, it is assumed that the error terms ε_t and ξ_t of the auto-regressive models are independent and identically normally distributed with expectations zero, and variances $\sigma_\varepsilon^2 = (1 - \kappa)(1 - \rho^2)$ and $\sigma_\xi^2 = (1 - \gamma)(1 - \phi^2)\sigma^2$, respectively.

$$\varepsilon_t \sim N(0, \sigma_\varepsilon^2) \quad \text{and} \quad \xi_t \sim N(0, \sigma_\xi^2),$$

To check the size of the non-response bias of the regression estimators and its fade-away effect in later panel waves, we consider different model stabilities of the covariate and error term. A total of eight Scenarios $A-H$ are considered here. In Scenario $A-D$ we assume equal stabilities of the covariate and residual components, while there are some intermixed cases (unequal stabilities) which are placed in Scenario $E-H$, for a more detailed explanation see Subsection 2.3.2.1 of Chapter 2.

- Scenario A: Low stability $\kappa = \gamma = \rho = \phi = 0.10$,
- Scenario B: Medium stability $\kappa = \gamma = \rho = \phi = 0.50$,
- Scenario C: High stability $\kappa = \gamma = \rho = \phi = 0.70$,
- Scenario D: High stability $\kappa = \gamma = \rho = \phi = 0.90$,
- Scenario E: Low permanent component $\kappa = 0.10$ of covariate and low stability $\rho = 0.10$ of AR(1) covariate, low permanent component $\gamma = 0.01$ of error and high stability $\phi = 0.70$ of AR(1) error,
- Scenario F: Medium permanent component $\kappa = 0.50$ of covariate and medium stability $\rho = 0.50$ of AR(1) covariate, low permanent component $\gamma = 0.01$ of error and high stability $\phi = 0.70$ of AR(1) error,
- Scenario G: Medium permanent component $\kappa = 0.70$ of covariate and medium stability $\rho = 0.70$ of AR(1) covariate, low permanent component $\gamma = 0.01$ of error and high stability $\phi = 0.70$ of AR(1) error,
- Scenario H: Medium permanent component $\kappa = 0.90$ of covariate and medium stability $\rho = 0.90$ of AR(1) covariate, low permanent component $\gamma = 0.01$ of error and high stability $\phi = 0.70$ of AR(1) error.

In order to get a better understanding of the fade-away effect of the regression estimators, we need to choose values of the initial non-response parameters α and β such that they create substantial initial non-response in the estimated model

parameters. Consequently, the parameters $\alpha = 0.05$ and $\beta = 0.40$ create a substantial initial non-response rate of approximately 25% in the first wave of the Resp-Sample. More importantly, the effect of α is non-selective for the response, while the effect of β is selective for the response. In the situation where the magnitude of β is small the model parameters are estimated with low bias. As a result of low initial biases in the estimates, there is no fade-away phenomenon present in later panel waves (for details see Table 9 of Section A.1 in Appendix A). Therefore, we choose a large value of $\beta = 0.40$ and investigate the fade-away effect in various scenarios of model stability of the covariate and error term, with and without panel attrition. In each scenario, we allow σ^2 to vary uniformly in the interval $0 \leq \sigma^2 \leq 1$, having values 0.1, 0.2, ...,1.0. We then estimate the slope coefficient b_r in each Monte Carlo replication r, $(r = 1, 2, 3, \ldots, R)$, where R stands for the number of Monte Carlo replications. Let $\hat{b}_{tr}$ be the regression estimator of b_{tr} in r^{th} simulation run with time t, then the average based on all replications is $\hat{b}_t = mean(\hat{b}_{tr}) = \frac{1}{100} \sum_{r=1}^{100} \hat{b}_{tr}$. Finally, we compute the following quantities based on all replications:

- Percent relative bias: $B_t = 100 \times (mean(\hat{b}_{tr}) - b_t)/b_t$, the bias in estimate $\hat{b}_t$ at time t. Where b_t indicates the true value of the regression coefficient at time t.

- The relative factor of the fade-away effect: In order to find the relative factor for the decline of initial non-response bias we divide the bias at time t by the bias at time $t-1$. It is written as $\lambda_{tsim} = B_{tsim}/B_{(t-1)sim}$ for $t \geq 2$, where B_{tsim} and $B_{(t-1)sim}$ are the percent relative biases of the estimates at time t and $t-1$, respectively.
  Note: To distinguish between the bias of the estimates obtained through approximation formula and through simulated study, we denote the bias of the approximation formula by B_{tcom} and their relative factor by λ_{tcom}, while the bias of the simulation study by B_{tsim} and their relative factor by λ_{tsim}.

3.4. Disscusion of the estimation results

In this section, we investigated the fade-away effect of the initial non-response bias of the cross-sectional OLS estimator in a four wave panel data. The motivation of using four wave panel data is to generalize the regression results of Alho (2015) to

longer panels and to check the validity of Alho's bias approximation formulas in a simulation setting. Further, for the generation of non-response, we use the linear approach of Alho's (details are given in Section 3.2 of this chapter). The size of the initial bias and its fade-away effect is computed by using the approximation formula in equation (2.32) as well as by the simulation study. As the bias approximation formula becomes intractable after wave 2, we only have to show the computed bias for the first two panel waves, and its corresponding relative factor, denoted by λ_{1com}.

Table 10 presents the bias of the cross-sectional OLS estimates for low stability of covariate and error term (Scenario A, $\kappa = \gamma = \rho = \phi = 0.10$). It can be seen from the table that there is no bias in the OLS estimates through approximation formula as well as through simulation study if the residual variance σ^2 is zero. For all other values of $\sigma^2 > 0$, there is some kind of small/large biases in the regression estimates, depending on the magnitude of residual variance and the stability of the covariate and error term. The larger the value of σ^2, the larger is the bias.

Therefore, we take the larger value of σ^2 in the interval $0 \leq \sigma^2 \leq 1$, i.e., $\sigma^2 = 1$. For this value of σ^2 the slope coefficients b_t are estimated with the largest initial bias. The corresponding size of the bias given by the approximation formula and by the simulation study is 28% and 24%, respectively. As the speed of the fade-away effect depends on the stability parameters $(\kappa, \gamma, \rho, \phi)$. If the stability parameters are low then estimates having substantial initial biases converge to the true solution in later waves. For example, in the last row of Table 2 where $\sigma^2 = 1$, the initial bias (through simulation study) is 24.19% which reduces to 1.42%, 0.51%, 0.02% in wave 2 wave 3 and wave 4, respectively. Here, the estimated slope coefficients in the last panel wave converge to the regression parameter b. This is simply because of the fade-away effect, and therefore under this aspect, the estimation results in later panel waves are more reliable and accurate than the corresponding first panel wave results. Thus, the speed of the fade-away effect is high in the presence of low stability $\kappa = \gamma = \rho = \phi = 0.10$. Moreover, in such a scenario the initial biases (both through approximation formula and through simulation study) faded-away in a non-geometrical fashion in later waves. Having low biases in later waves, the corresponding relative factors of the fade-away effect are very fast which are $\lambda_{1sim} = 6\%$, $\lambda_{2sim} = 36\%$ and $\lambda_{3sim} = 4\%$ (see the last row of Table 2).

Table 2: Scenario A: Impact of residual variance σ^2 on the bias of OLS estimates, with $\kappa = \gamma = \rho = \phi = 0.10$, without any attrition pattern.

σ^2	Relative bias*100						Relative factor (λ)			
	B_{1com}	B_{1sim}	B_{2com}	B_{2sim}	B_{3sim}	B_{4sim}	λ_{1com}	λ_{1sim}	λ_{2sim}	λ_{3sim}
0.10	2.84	3.98	0.08	0.00	0.00	0.00	0.03	0.00	0.00	0.00
0.20	5.69	6.88	0.16	0.13	0.10	0.30	0.03	0.02	0.77	3.00
0.30	8.53	9.88	0.24	0.04	0.04	0.05	0.03	0.01	0.01	1.25
0.40	11.31	12.68	0.32	0.20	0.50	0.57	0.03	0.02	2.50	1.14
0.50	14.22	14.82	0.41	0.90	0.16	0.06	0.03	0.06	0.17	0.38
0.60	17.07	17.14	0.49	0.30	0.35	0.08	0.03	0.02	1.17	0.23
0.70	19.91	18.99	0.57	0.16	0.17	0.37	0.03	0.01	1.06	2.18
0.80	22.76	20.67	0.65	0.81	0.18	0.79	0.03	0.04	0.22	4.39
0.90	25.60	22.80	0.73	1.35	0.51	0.14	0.03	0.06	0.38	0.28
1.00	28.44	24.19	0.81	1.42	0.51	0.02	0.03	0.06	0.36	0.04

However, by increasing the size of the permanent and the transient components $(\kappa, \gamma, \rho, \phi)$ it considerably increases the size of biases in later panel waves. Therefore, due to large biases in later waves, the speed of the fade-away effect is small. We demonstrate this in Table 3.

Table 3: The speed of convergence to steady-state distribution in Scenario A-D, with residual variance $\sigma^2 = 1$, without any attrition pattern.

Scenario	Relative bias*100						Relative factor (λ)			
$(\kappa, \gamma, \rho, \phi)$	B_{1com}	B_{1sim}	B_{2com}	B_{2sim}	B_{3sim}	B_{4sim}	λ_{1com}	λ_{1sim}	λ_{2sim}	λ_{3sim}
A (0.10)	28.44	24.19	0.81	1.42	0.51	0.02	0.03	0.06	0.36	0.04
B (0.50)	28.44	24.36	14.23	12.63	8.31	6.78	0.52	0.50	0.66	0.82
C (0.70)	28.44	24.54	22.46	19.23	16.24	14.15	0.79	0.78	0.85	0.87
D (0.90)	28.44	23.65	27.72	23.14	22.60	22.20	0.98	0.98	0.98	0.98

Apparently, it is visible from Table 3 that by increasing the size of the permanent and the transient components from $\kappa = \gamma = \rho = \phi = 0.10$ to $\kappa = \gamma = \rho = \phi = 0.50$ (Scenario B), we get larger biases for the estimates. Due to larger biases in later waves the speed of the fade-away effect is smaller than what it was in Scenario A. For instance, the initial and subsequent biases through simulation study are: 24.36%, 12.63%, 8.31%, 6.78% having relative factors: $\lambda_{1sim} = 50\%$, $\lambda_{2sim} = 66\%$,

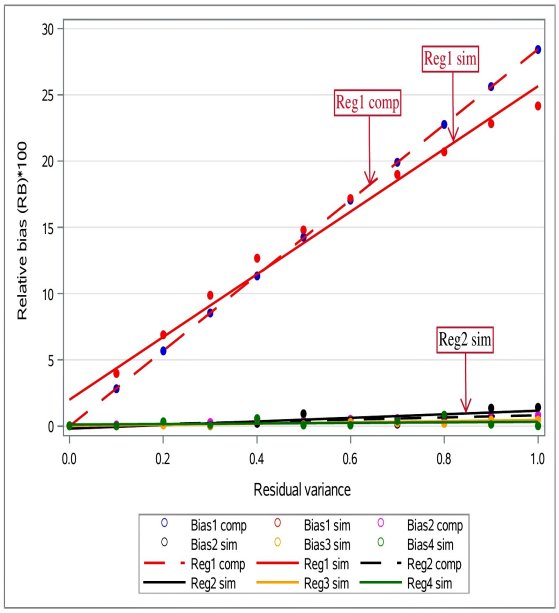
$\lambda_{3sim} = 82\%$. Interestingly, in this case the initial bias reduces in a geometrical fashion in successive panel waves. Scenario C ($\kappa = \gamma = \rho = \phi = 0.70$), repeats Scenario B with more stable results. Finally, Scenario D is the more extreme scenario where the size of stability parameters is very high $\kappa = \gamma = \rho = \phi = 0.90$. As already discussed in Subsection 2.3.2.1 of Chapter 2 the fade-away of the bias depends on the size of the permanent components of the covariate and the error term. If the size of the permanents is large then their distribution remains stationary and hence the effect of initial non-response doesn't reduce over time. While the situation is completely different for the transient components, which swing into the steady-state distribution of the Markov chain. This is documented in Scenario D. From this we can see that for the large permanent component the distorting effect of initial non-response remains permanent for the estimated slope coefficients for all panel waves. The biases under this scenario (Scenario D) are: 23.65%, 23.14%, 22.60% and 22.20%, respectively, while the speed of the fade-away effect are: $\lambda_{1sim} = \lambda_{2sim} = \lambda_{3sim} = 98\%$. Hence, there is no fade-away present. A complete list of analysis tables under all Scenario A-H can be found in Section A.1 of Appendix A. In order to get a better understanding of the fade-away effect in Scenario A-D, we graphically displayed the results in Figure 1.

In the figure, the bias of the OLS estimates under different scenarios of Stability A-D has been plotted against σ^2 values. The vertical axis of the figures shows the relative bias of the estimates, while the horizontal axis shows the magnitude of the residual variance σ^2. The colored's circles points in the graph display the relative bias in panel wave t, (where $t = 1,2,3,4$) implied by our regression model. The solid lines in different colors correspond to the regression lines fitted to the biases obtained from the simulated study, while the dashed lines in different colors are the regression lines fitted to the biases given by the approximate formula. The figures display an impression of the fade-away effect. Intuitively, it is visible from the graph under Scenario A that the initial non-response biases which are of substantial size faded-away just in one panel wave, while in Scenario B these biases decrease in a geometrical pattern. Scenario C repeats Scenario B, with more stable results. Scenario D is the most extreme case, where there is apparently no reduction in the initial bias in following panel waves.

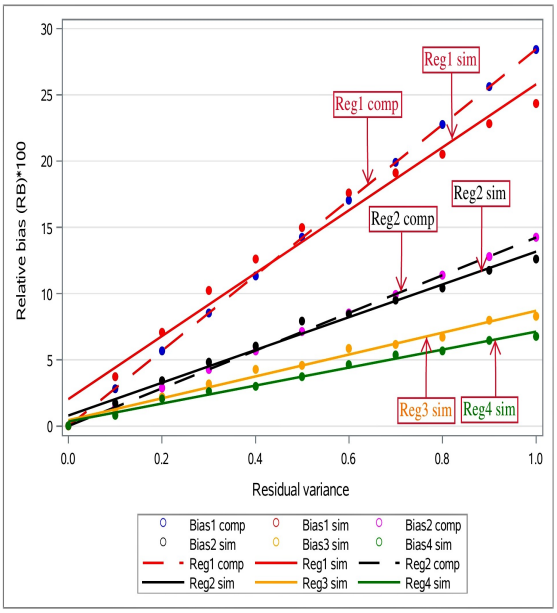
In Scenario A-D, we investigate the fade-away effect assuming equal stability of the permanent and transient components. In addition, it will also be worthwhile

to investigate the fade-away effect under different sizes of permanent and transient components. Interestingly, similar results on the fade-away (as in Scenario A-D) also hold for Scenario E-H. Except here the speed of the fade-away effect is faster than in the speed of the fade-away effect in Scenario A-D. This is due to the fact that in these scenarios (Scenario E-H) the size of the permanent component γ of the error term is very small which is $\gamma = 0.01$. We know that if non-response is related to the distribution of the error term then the slope coefficients are estimated with a bias. Further, if this bias depends on the permanent component of the error term, it will remain permanent too. However, if it depends on the stability of the temporal structure of the error term then it will quickly fade-away over time. We visualize the results of these intermixed cases (Scenario E-H) in Figure 2.

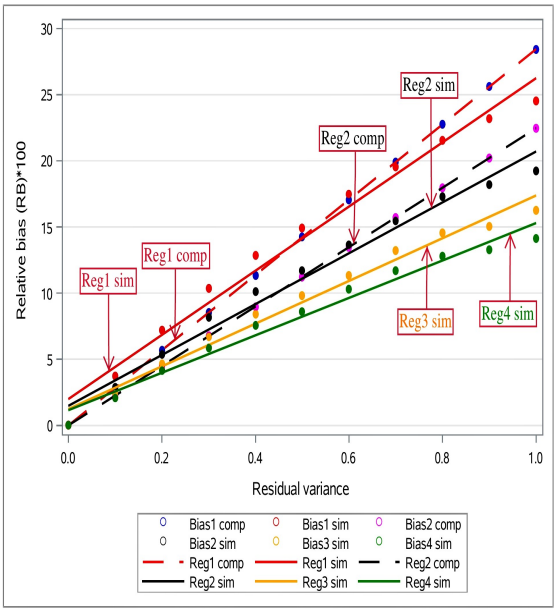
(a) Scenario A ($\kappa = \gamma = \rho = \phi = 0.10$).



(b) Scenario B ($\kappa = \gamma = \rho = \phi = 0.50$).



(c) Scenario C ($\kappa = \gamma = \rho = \phi = 0.70$).



(d) Scenario D ($\kappa = \gamma = \rho = \phi = 0.90$).

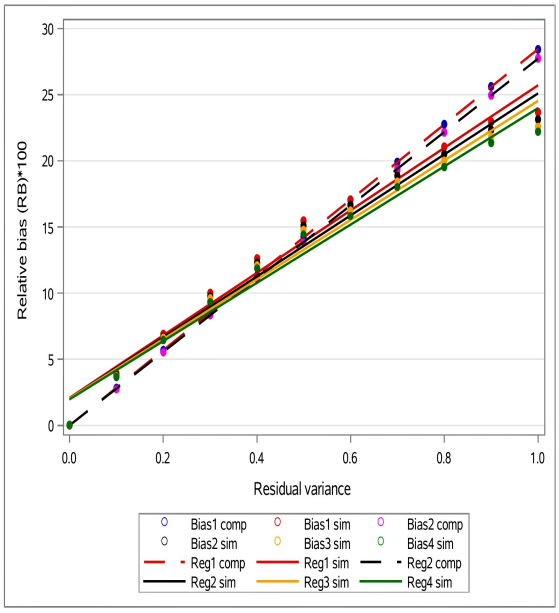
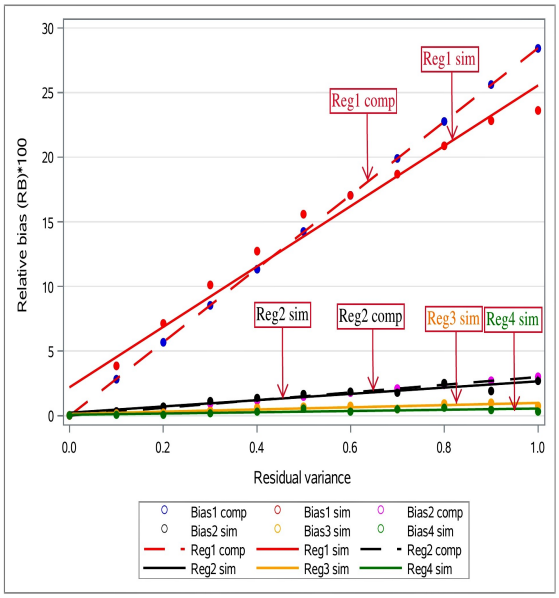


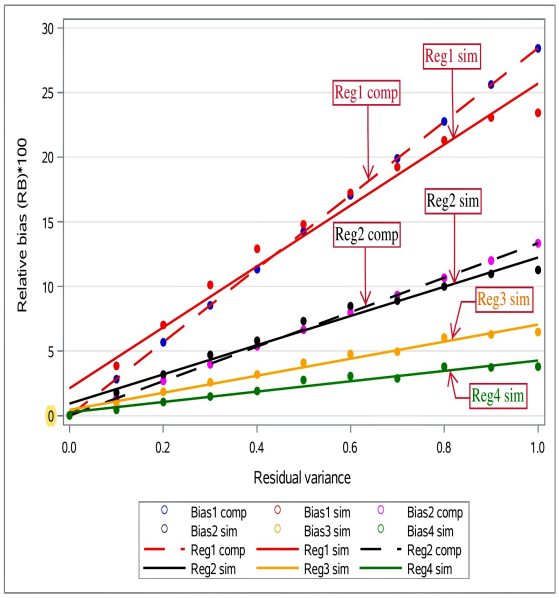
Figure 1: Graph of the impact of residual variance σ^2 on the bias of OLS estimates of $b_t = 1$ in Scenario A-D.

Number of observations in the sample are $n = 1{,}000$, which are replicated $R = 100$ times. Initial non-response rate is 25%. The colored's circles points on the graph display the bias of the estimate in a certain wave at time point t, $(t = 1, 2, 3, 4.)$ which is plotted against σ^2 values. Dotted line: Bias computed by approximation formula. Solid line: Bias computed through a simulation study.

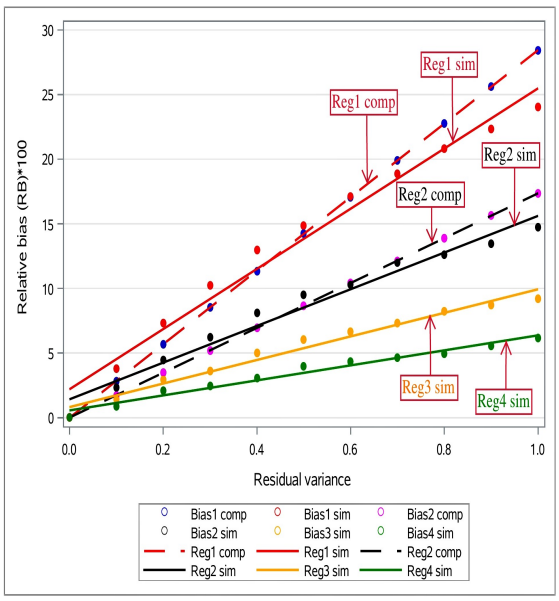
(a) Scenario E ($\kappa = \rho = 0.10, \gamma = 0.01$ and $\phi = 0.70$).



(b) Scenario F ($\kappa = \rho = 0.50, \gamma = 0.01$ and $\phi = 0.70$).



(c) Scenario G ($\kappa = \rho = 0.70, \gamma = 0.01$ and $\phi = 0.70$).



(d) Scenario H ($\kappa = \rho = 0.90, \gamma = 0.01$ and $\phi = 0.70$).

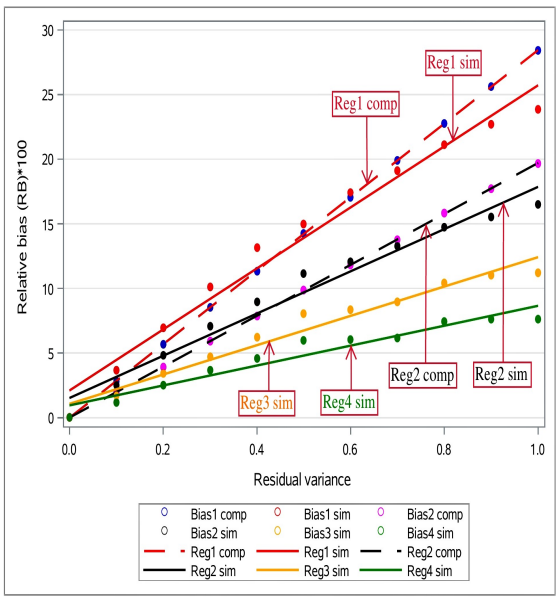


Figure 2: Graph of the impact of residual variance σ^2 on the bias of OLS estimates of $b_t = 1$ in Scenario E-H.

Number of observations in the sample are $n = 1{,}000$, which are replicated $R = 100$ times. Initial non-response rate is 25%. The colored's circles points on the graph display the bias of the estimate in a certain wave at time point t, $(t = 1, 2, 3, 4.)$ which is plotted against σ^2 values. Dotted line: Bias computed by approximation formula. Solid line: Bias computed through a simulation study.

Up to now, we could verify the fade-away effect of the initial non-response in four wave panel surveys. Under the assumption that there is no further drop-out/attrition of observations after the initial wave one. Under this aspect, we show that the size of non-response bias in the estimation of Resp-Sample gets smaller and smaller over time depending on Scenario A-D. As a consequence, the distribution of the Resp-Sample become more alike to the distribution of the Full-Sample. However, the assumption of no further attrition after wave one is extreme in panel surveys. Often panel surveys are also affected by selective attrition in later waves. If attrition is selective then, it may be the case that the distorting effect of non-response on the estimated slope coefficients at the start of the panel may be enhanced by attrition in later panel waves. Thus, it is more realistic to consider the fade-away effect under different selective attrition. Therefore, our next goal is to investigate the fade-away effect under the presence of selective attrition after the initial wave. Then under this aspect, our Resp-Sample will be further reduced by using the attrition model in equation (3.2). To investigate the fade-away effect in the existence of attrition, we artificially generate an attrition probability of about 10% in the Resp-Sample, for three different attrition scenarios:

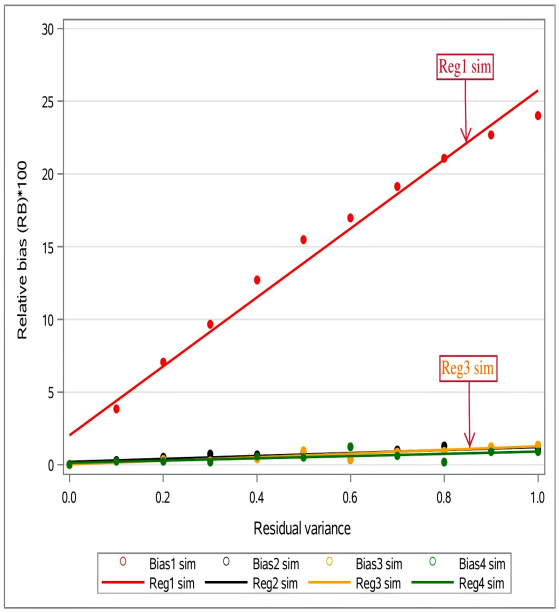
- Scenario 1: $\alpha^\star = 0.80$ and $\beta^\star = 0.05$,
- Scenario 2: $\alpha^\star = 0.70$ and $\beta^\star = 0.10$,
- Scenario 3: $\alpha^\star = 0.50$ and $\beta^\star = 0.20$,

where $\alpha^\star$ is non-selective and $\beta^\star$ is selective for the response after the initial wave. The fade-away hypothesis assumes that attrition (selective attrition) would not depend on the variable of interest. Usually, an attrition rate of about 5% doesn't harm the results or disturb the fade-away effect, but it may perhaps reduce case numbers. So, in our case the larger the value of $\beta^\star$ the higher will be the drop-out probability and consequently, the smaller will be the fade-away effect. The results of these attrition scenarios under model stability Scenario A-D are presented in Figure 3 to Figure 5.

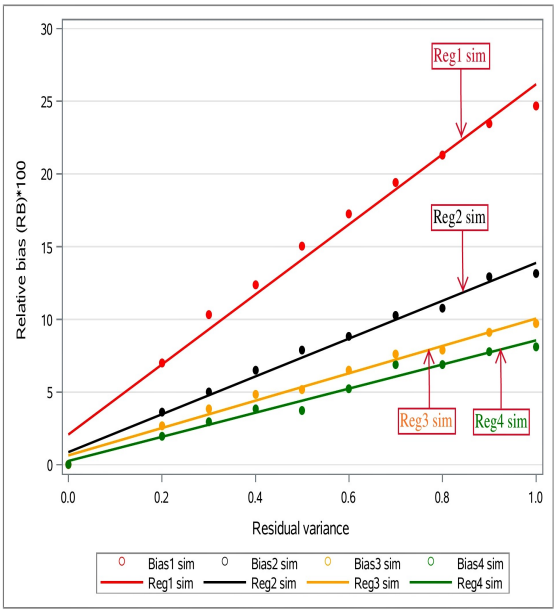
First, we will check whether the selective panel attrition up to $\beta^\star = 0.05$ in attrition Scenario 1 makes any real difference in the fade-away results. We plot the initial bias and the attrition biases of this scenario in Figure 3. By looking at the

results in Figure 3 we find that attrition Scenario 1 ($\alpha^\star = 0.80, \beta^\star = 0.05$) don't tend to create larger biases in the estimates due to attrition in later panel waves. This was done by comparing the results of Figure 3 with the results of Figure 1 (where we don't use any attrition pattern), which reveals that selective attrition up to 5% don't any make any real difference and therefore the speed of the fade-away effect is unaffected by this attrition scenario. Nevertheless, the fade-away effect was affected in attrition Scenario 2 ($\alpha^\star = 0.70, \beta^\star = 0.10$), however, the difference is not so large. This is documented in Figure 4. In the case of severe selective attrition parameter $\beta^\star = 0.20$ (Scenario 3), the distorting effect of initial non-response is propagated by the selective effect of attrition in later waves depending on the model stability of the covariate and the error term. However, it has been shown that for $\beta^\star = 0.20$ the estimates having substantial initial non-response don't converge to its parameter. This holds for all scenarios of Figure 5. As at the start of this section, we have discussed the fade-away effect in Scenario A-D without any attrition pattern. From there we confirmed that the size of the initial bias under Scenario C faded-away in a geometrical pattern, while the effect has fully disappeared in Scenario D (see Figure 1). However, the results of these Scenario C-D become worse when attrition Scenario 3 ($\alpha^\star = 0.50, \beta^\star = 0.20$) is used. For example, in Scenario D ($\kappa = \gamma = \rho = \phi = 0.90$) the size of initial bias of the OLS estimate at wave one (marked as color red) is smaller than the size of bias at wave 2 (marked as color black), wave 3 (marked as color orange) and wave 4 (marked as color green), respectively. So the strength of the fade-away effect is not affected by low selective attrition but it is really affected by massive selective attrition. Detailed results on the fade-away effect under different attrition rates can be seen from Table 18 to Table 29 in Section A.1 of Appendix A.

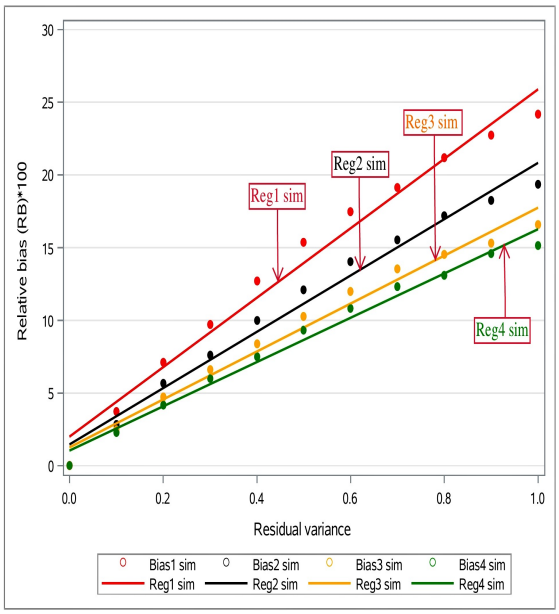
(a) Scenario A, with $\beta = 0.40$ and $\beta^{\star} = 0.05$



(b) Scenario B, with $\beta = 0.40$ and $\beta^{\star} = 0.05$



(c) Scenario C, with $\beta = 0.40$ and $\beta^{\star} = 0.05$



(d) Scenario D, with $\beta = 0.40$ and $\beta^{\star} = 0.05$

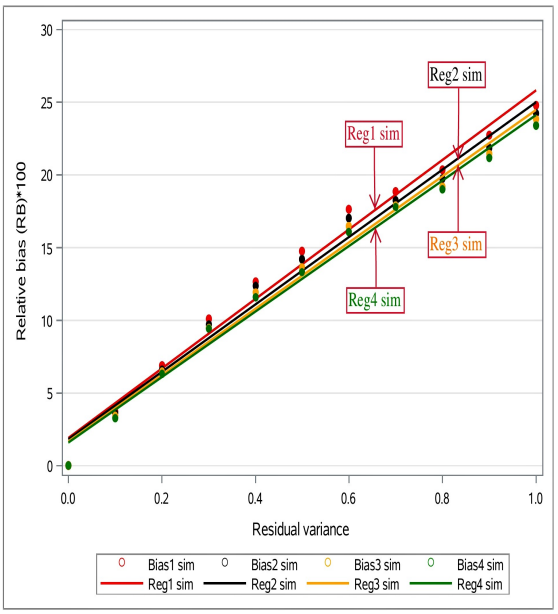
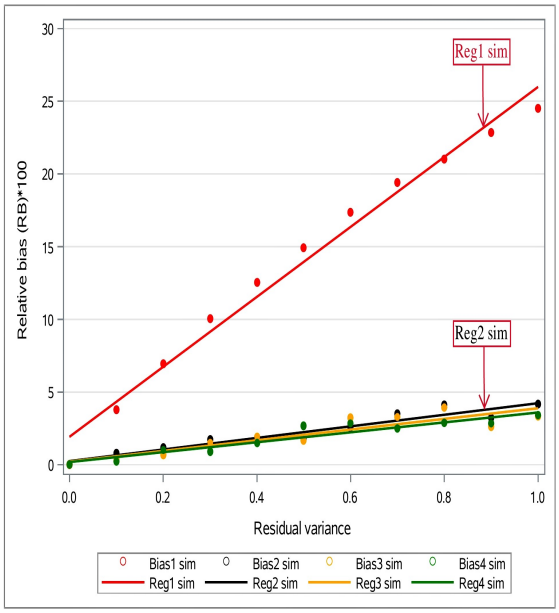


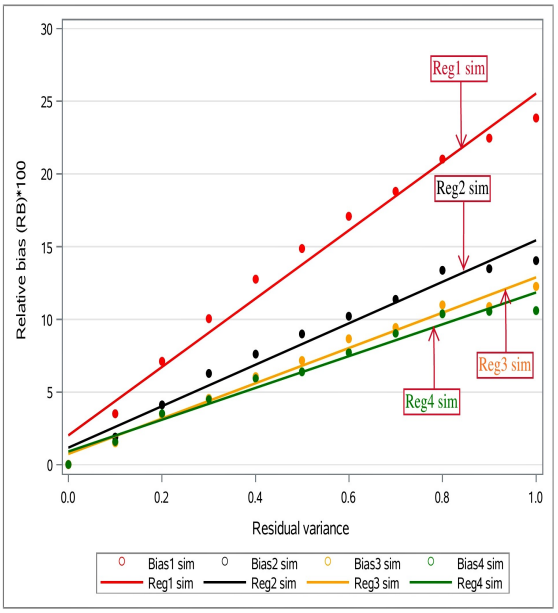
Figure 3: Graph of the impact of residual variance σ^2 on the bias of OLS estimates of $b_t = 1$ in Scenario A-D.

Number of observations in the sample are $n = 1{,}000$, which are replicated $R = 100$ times. Initial non-response rate is 25%. Attrition rate is 10%. The colored's circles points on the graph display the bias of the estimates in a certain wave at time point t, which are plotted against σ^2 values. Solid line: Bias computed through a simulation study.

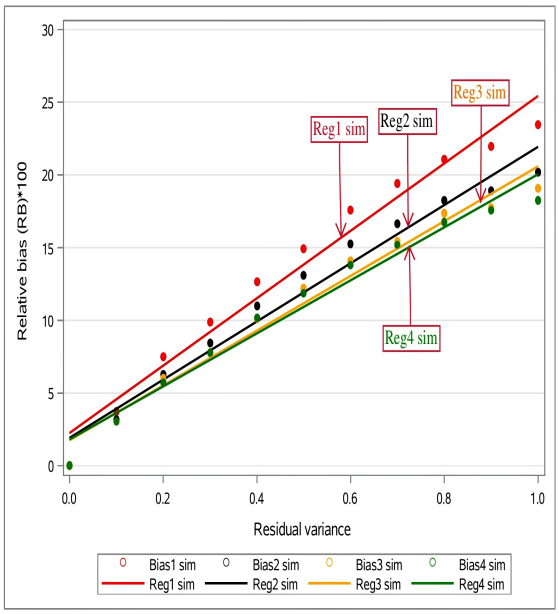
(a) Scenario A, with $\beta = 0.40$ and $\beta^{\star} = 0.10$



(b) Scenario B, with $\beta = 0.40$ and $\beta^{\star} = 0.10$



(c) Scenario C, with $\beta = 0.40$ and $\beta^{\star} = 0.10$



(d) Scenario D, with $\beta = 0.40$ and $\beta^{\star} = 0.10$

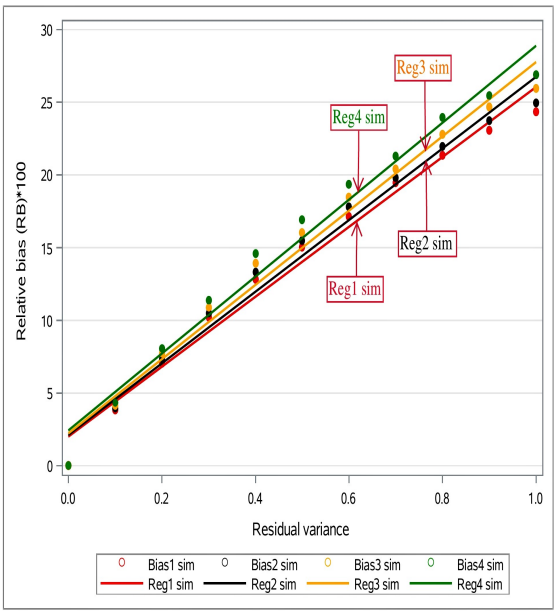
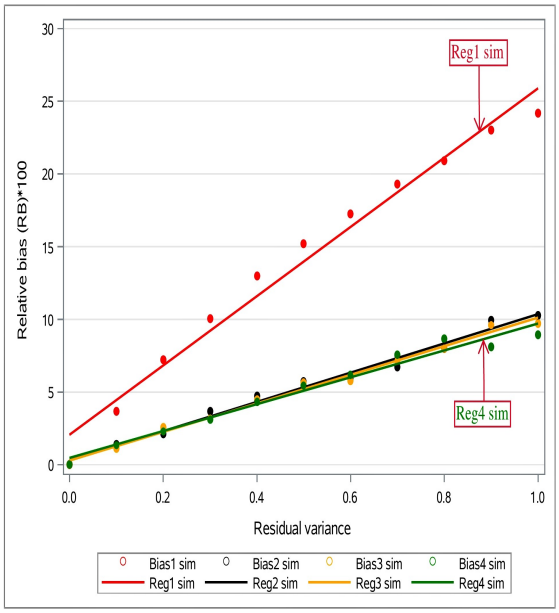


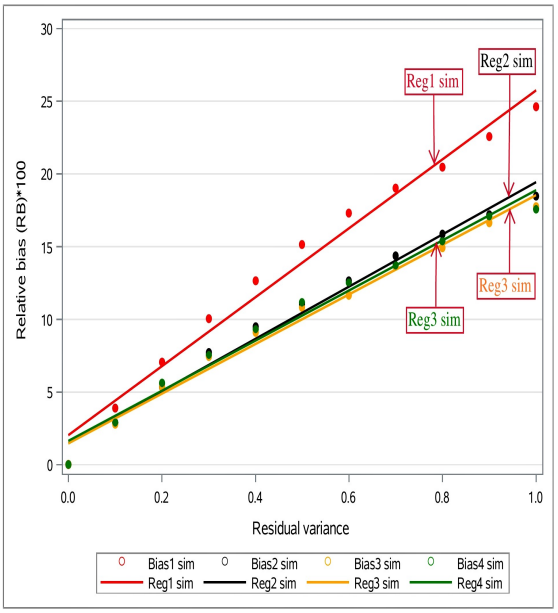
Figure 4: Graph of the impact of residual variance σ^2 on the bias of OLS estimates of $b_t = 1$ in Scenario A-D.

Number of observations in the sample are $n = 1{,}000$, which are replicated $R = 100$ times. Initial non-response rate is 25%. Attrition rate is 10%. The colored's circles points on the graph display the bias of the estimates in a certain wave at time point t, which are plotted against σ^2 values. Solid line: Bias computed through a simulation study.

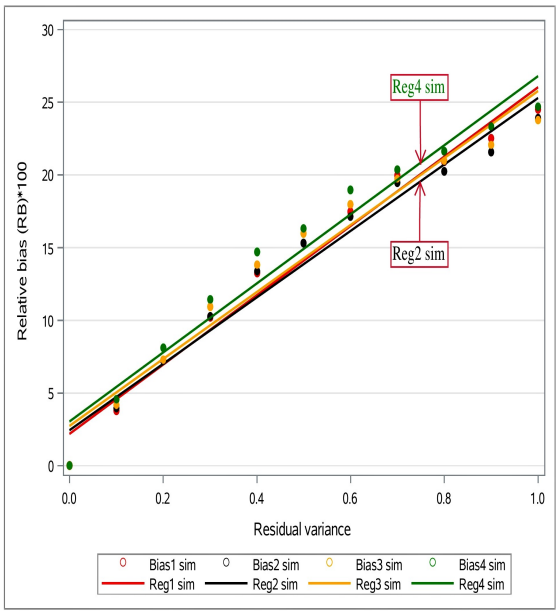
(a) Scenario A, with $\beta = 0.40$ and $\beta^{\star} = 0.20$



(b) Scenario B, with $\beta = 0.40$ and $\beta^{\star} = 0.20$



(c) Scenario C, with $\beta = 0.40$ and $\beta^{\star} = 0.20$



(d) Scenario D, with $\beta = 0.40$ and $\beta^{\star} = 0.20$

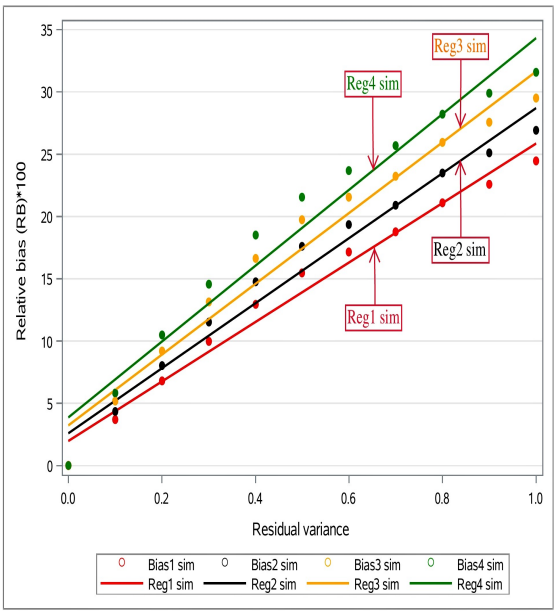


Figure 5: Graph of the impact of residual variance σ^2 on the bias of OLS estimates of $b_t = 1$ in Scenario A-D.

Number of observations in the sample are $n = 1{,}000$, which are replicated $R = 100$ times. Initial non-response rate is 25%. Attrition rate is 10%. The colored's circles points on the graph display the bias of the estimates in a certain wave at time point t, which are plotted against σ^2 values. Solid line: Bias computed through a simulation study.

3.5. Fade-away effect for the OLS and IPW estimates

In Section 3.3 of this chapter, we have conducted a Monte Carlo simulation study to verify the approximate results of Alho (2015), and investigated the fade-away effect of the initial non-response bias of the cross-sectional OLS estimator in a four wave panel survey. For the initial response, we have used the linear approach of Alho (2015). However, usually, the longer running panel provides more reliable results than panels with shorter periods. Therefore, the objective of this section is to extend the simulation database to up to ten panel waves and investigate the fade-away effect of the cross-sectional OLS estimator under panel attrition. Besides the estimation results of the un-weighted OLS estimator, this section also considers an IPW estimator and its comparison with the un-weighted OLS estimator. IPW is beneficial in correcting the bias of survey estimates. Therefore, we consider how well the IPW approach described in Section 2.4 of Chapter 2 does remove the non-response bias under not missing at random (NMAR) assumption at the start of the panel.
To evaluate the performance of these estimators we use the previously used simulation setup as was described in Section 3.3, i.e., we simulate a sample of size 1,000 units from the model over 100 Monte Carlo replications. Except for the estimation of the initial response probability we use the binary logit model, while previously we used the linear approach of Alho (2015). Therefore, for the initial non-response, we use a binary logit model, in which the non-response depends on the response variable $Y_{i,1}$. Mathematically, it can be written as:

$$P(R_{i,1}=1|Y_{i,1}) = \frac{\exp(\alpha+\beta Y_{i,1})}{[1+\exp(\alpha+\beta Y_{i,1})]}, \tag{3.3}$$

To introduce an initial non-response bias of substantial size at the start of the Resp-Sample we choose $\alpha=-4.50$ and $\beta=3.20$ which results in a non-response rate of 35%.
Further, for panel attrition after wave 1 ($t>1$) we use the logit regression to model

the attrition probability which is defined by:

$$P(R_{i,t}=1|Y_{i,t},R_{i,1}=1,R_{i,2}=1,R_{i,3}=1,\ldots,R_{i,t-1}=1)=\frac{\exp(\alpha^{\star}+\beta^{\star}Y_{i,t})}{[1+\exp(\alpha^{\star}+\beta^{\star}Y_{i,t})]}, \tag{3.4}$$

where $t=2,3,4,\ldots,10$.
For attrition parameters $\alpha^{\star}=0.90$ and $\beta^{\star}=0.90$ the model results in an attrition rate of about 5-10% depending on the stability of the covariates and the error term.

One of the objectives of this thesis is to adjust for non-ignorable non-response through the use of IPW, which weights observations of the sample by the reciprocals of the estimated response probabilities. The initial non-response weights for sample observations are obtained by the inverse of the estimated initial response probabilities, while for the later waves the weights are updated by the corresponding estimated attrition probabilities. To be more precise about the logit model for the non-response weights, here we used two weighting scenarios: (i) a realistic weighting scenario; (ii) and an unrealistic weighting scenario.

In the realistic scenario, we used two cases: in the first case, we correct for both the initial non-response and attrition through weighting, while in the second case we control only for the initial non-response (we don't care about attrition weighting). Under this aspect, the initial weights are constructed by using the information on the covariate vector $X_{i,t=1}$ which is supposed to be known for the respondents and the non-respondents. However, in most surveys information on the non-respondents is usually unavailable except from register data which contains information about the non-respondents. Therefore, the weights in the later panel waves are constructed by using the information of lagged dependent variable $Y_{i,t-1}$ as an explanatory variable.

While in the unrealistic case, for the initial non-response weights we use the true $Y_{i,1}$ for the prediction of non-response as a weighting variable. For the estimation of attrition weights, we use a logit model with lagged dependent variable $Y_{i,t-1}$ as an explanatory variable (similar to attrition weights in the realistic case). These two weighting scenarios are presented in the following table (Table 4).

Table 4: Non-response/attrition model in simulation.

	Initial non-response	Attrition
Non-response/attrition	First wave	Later waves
OLS (un-weighted case)	-	-
IPW 1 (realistic case)	X_1	Y_{t-1}
IPW 2 (realistic case)	X_1	-
IPW 3 (unrealistic case)	Y_1	Y_{t-1}

First, we will discuss the realistic cases, and then we will switch the discussion to the unrealistic case. In the realistic case, the initial non-response weights are equal to the inverse of the initial response probability $P(R_{i,1}=1|X_{i,1})$ which can be written as:

$$W_{i,1} = W(x_{i,1}) = \frac{1}{\hat{P}(R_{i,1}=1|X_{i,1})}, \tag{3.5}$$

where $\hat{P}(R_{i,1}=1|X_{i,1})$ is the logit estimate of $P(R_{i,1}=1|X_{i,1})$ and $X_{i,1}$ is the known predictor variable in the initial wave 1.

In the case of panel attrition, there is a considerable amount of information available on previous panel waves before attrition, such as information on lagged dependent variables. These variables are good predictors of attrition and using them for weighting variables may considerably reduce the attrition bias, because they are closely related to the response propensity score. Therefore, for the construction of attrition weights, we used lagged dependent $Y_{i,t-1}$ a weighting variable. Under this aspect, the estimation of weights in the following panel wave $2,3,4,\ldots,t$ is sequentially updated by the inverse of the estimated attrition probabilities attained from the model (3.4) as follows:

$$W_{i,t} = W(y_{i,t-1}) = \frac{1}{\hat{P}(R_{i,t}=1|Y_{i,t-1},R_{i,1}=1,R_{i,2}=1,R_{i,3}=1,\ldots,R_{i,t-1}=1)}, \tag{3.6}$$

where $\hat{P}(R_{i,t}=1|Y_{i,t-1},R_{i,1}=1,R_{i,2}=1,R_{i,3}=1,\ldots,R_{i,t-1}=1)$ is the logit estimate of $P(R_{i,t}=1|Y_{i,t-1},R_{i,1}=1,R_{i,2}=1,R_{i,3}=1,\ldots,R_{i,t-1}=1)$, and $Y_{i,t-1}$ is the value of lagged dependent variable at time point $t-1$.

Earlier in this section, we have discussed the properties of weighted and un-weighted cross-sectional OLS estimators in terms of bias and MSE. The un-weighted

OLS regression assigns equal weights to all observations, while the weighted regression assigns weights to the observations by the inverse of the estimated response/attrition probabilities. For the weights, we use the logit model using the covariate X_t vector, under the assumption that it is known both for the respondents and non-respondents. However, in most surveys information on the non-respondents is usually unavailable except from register data which also contains information about the non-respondents. Therefore, in the unrealistic case (denoted by IPW 3), we use true information on $Y_{i,1}$ for the construction of non-response weighting. Under this aspect, the respondents in the initial wave are weighted by

$$W_{i,1} = W(y_{i,1}) = \frac{1}{\hat{P}(R_{i,1} = 1|Y_{i,1})}, \tag{3.7}$$

where $Y_{i,1}$ is the response variable in the initial wave 1. This is an unrealistic scenario because $Y_{i,1}$ is unknown. Similarly, the respondents in the subsequent wave $2, 3, 4, \ldots, t$ is weighted by

$$W_{i,t} = W(y_{i,t-1}) = \frac{1}{\hat{P}(R_{i,t} = 1|Y_{i,t-1}, R_{i,1} = 1, R_{i,2} = 1, R_{i,3} = 1, \ldots, R_{i,t-1} = 1)}, \tag{3.8}$$

where $Y_{i,t-1}$ is the one year lagged dependent variable at time point t. The motivation of using the information on lagged dependent variable for the construction of attrition weights, is to assume that $Y_{i,t-1}$ is known in wave t.

By using the correct weights (unrealistic scenario) the corresponding weighted estimator performs better than the un-weighted OLS estimator. IPW is better because it achieves large gains in bias reduction with respect to un-weighted OLS. Further, despite using the true weighting variable the bias of the estimates has not completely vanished. However, weighting guarantees only consistent parameter estimates because a bias of 10% at wave 1 is always present in the estimates.
To demonstrate the benefits of using the IPW estimators, its properties in terms of bias and MSE are compared with the bias and MSE of the un-weighted OLS estimator. For this, we reported the bias and MSE of the estimators under Scenario A-D in Tables 30 to 33 in Section A.2 of Appendix A. To get a more clear picture of the fade-away effect of the non-response bias under the OLS and the IPW estimator, we present a graphical display of the fade-away effect in the following figures. For

each Scenario A-D, we plot the results in Figure 6 to Figure 9, respectively. The vertical axis of the figures shows the percent relative bias of the estimator, while the horizontal axis shows the wave of the panel. The points (as shown in colored dots) on the graph display the single bias in wave t, where $t = 1,2,3,\ldots,10$, implied by our regression model. The green solid line marked with the letter "OLS Full" corresponds to the bias of the OLS estimator under the Full-Sample. The red solid line marked with the letter "OLS Resp" corresponds to the bias of the OLS estimator under the Resp-Sample. For the realistic weighting scenarios, the blue solid line marked with the letter "IPW 1 Resp" corresponds to the bias of the IPW estimator (sequential weights with attrition) under the Resp-Sample, while, the black solid line marked with the letter "IPW 2 Resp" corresponds to the bias of the IPW estimator (only initial non-response weights with attrition) under the Resp-Sample. Similarly, for the unrealistic weighting scenario, the orange solid line marked with the letter "IPW 3 Resp" corresponds to the bias of the estimator under the Resp-Sample.

The pattern of bias lines that appears in the figures is truly astonishing. It can be seen from the figures that OLS and IPW bias lines are very different from each other, but the bias of the OLS estimates is smaller than the bias of the IPW estimates (unrealistic case). Hence, despite using extra knowledge from the covariate for the construction of weights, there is no improvement in bias reduction with respect to OLS. Even the bias of the IPW becomes worse, this holds for all the figures under Scenario A-D. Moreover, weighting under the realistic scenario doesn't guarantee consistent parameter estimates. The fade-away effect is present for both OLS and IPW estimates in panels for low and medium stability scenarios. However, OLS is better because the bias of the OLS estimates fades-away faster than the bias of the IPW estimates. For example, in Figure 6 we plot the bias of the estimators obtained under Scenario A (low stability: $\kappa = \gamma = \rho = \phi = 0.10$). It can be seen from Figure 6 that the initial non-response biases of the OLS and IPW estimators do fade-away rapidly in later panel waves. Moreover, the speed of convergence to a steady-state distribution is fast, this is because the variance of the permanent components M and V are very small, and the stability of AR(1) covariate Z_t and AR(1) error U_t is very low. I.e., the variables at time period t are less correlated with the variables at time period $t-1$ and thus the selection effects at time $t-1$ play a minor role in the regression at time t.

Turning to the IPW cases (realistic cases: IPW 1 and IPW 2), it doesn't help

in reducing the size of non-response bias, they are even more biased than the OLS estimator. First, despite using extra knowledge on the covariate there is no improvement in bias reduction. It may be that the covariate doesn't well explain the non-response. Second, the effect of weighting in reducing non-response bias largely depends on the correct specification of the response propensity model. If the model is misspecified, then the estimated slope coefficients of the IPW estimator is likely to be either overestimated or underestimated. Third, some respondents have very small estimated initial response probabilities and thus receive very large initial weights, which in turn may lead to higher variances of the IPW estimator. For the OLS case the initial bias (colored red marked with the letter "OLS Resp") of size -29.93% reduces to -0.80% in wave 10, while for the IPW case the initial bias (colored blue marked with the letter "IPW 1 Resp") is estimated -35.57% which reduces to -0.80% in wave 10. The IPW estimator performs very poorly with large attrition biases when we don't control for attrition, e.g., the size of initial bias -35.57% (colored black marked with the letter "IPW 2 Resp") declines to -5.75% in wave 10. This holds in all stability scenarios (Scenario A-D).

Further, it has also been shown that increasing the size of the permanent components can cause a considerable reduction in the speed of the fade-away effect. We report the fade-away effect under Scenario B (medium stability: $\kappa = \gamma = \rho = \phi = 0.50$) in Figure 7. Here the initial non-response bias decreases geometrically in later panel waves. Under this scenario, the bias of OLS estimator in wave 1 is -30.05% which faded-away to -11.08% in wave 10. Similarly, for the IPW estimator, the bias in wave 1 is -35.37% (marked with the letter "IPW 1 Resp") which faded-away to -8.75% in wave 10. Scenario C (medium stability: $\kappa = \gamma = \rho = \phi = 0.70$) is a more dramatic case in the same way, but the results are more stable (see Figure 8). On the contrary, Scenario D is the more extreme simulation example where there is no fade-away expected because of high stability $\kappa = \gamma = \rho = \phi = 0.90$. We demonstrate this in Figure 9. It is visible from the figure that large permanent components of the covariate and error term make the distorting effect of initial non-response permanent in subsequent panel waves. Due to large biases in later waves, there is no fade-away effect present. The reason for these high substantial biases is that if the size of the permanent component is high, the variables at time t is highly correlated to the variable at time $t-1$. Thus the probability of responding in wave t depends on the regression model at wave $t-1$, which results in a high potential bias in the later

panel waves. Furthermore, comparing the MSE's of the OLS and IPW estimators, we see that the MSE's of the IPW estimator are higher than the MSE's of OLS estimator. This is because there is some efficiency cost of conducting IPW regression. This is documented in Figure 10 to Figure 13.

In the regression model where weighting is related to Y_t, the non-response/attrition biases of the estimator are smaller than the biases of the corresponding OLS estimators that don't use any weights or using weights that don't depend on Y_t. This is not surprising because here we used the correct weighting model. Overall the weighting approach using true weights performs well in all stability scenarios (Scenario A-D). For example, looking to the estimation results under Scenario A (low stability: $\kappa = \gamma = \rho = \phi = 0.10$) which are plotted in Figure 6, we see that resulting bias of the IPW estimator with the correct weighting variable $Y_{i,1}$ in wave 1 is -9.82% (colored orange marked with the letter "IPW 3 Resp"), while the size of the bias for the un-weighted estimator in wave 1 is -30.66% (colored red marked with the letter "OLS Resp") which is almost 3 times larger than the bias of the IPW estimator with $X_{i,t}$ a weighting variable. Interestingly, under this scenario the attrition biases of two estimators in later waves remain the same which is -5%.

It is worth noting that by increasing the size of stability parameters $(\kappa, \gamma, \rho, \phi)$ the fade-away effect of IPW estimates have completely disappeared, while this does not hold for the OLS estimates (except in a most extreme case, Scenario D: $\kappa = \gamma = \rho = \phi = 0.90$). For example, by considering the estimation results under Scenario B (medium stability: $\kappa = \gamma = \rho = \phi = 0.50$) in Figure 7, we see that the bias of the OLS estimator in wave 1 is -30.56% (colored red marked with the letter "OLS Resp") which faded-away to -13.60% in wave 10, while for the IPW estimates the size of initial bias is -10.06% (colored orange marked with the letter "IPW 3 Resp") which increases to -13.33% in wave 10. Scenario C (high stability: $\kappa = \gamma = \rho = \phi = 0.70$), repeats Scenario B with more stable results. The results under Scenario C are documented in Figure 8. Scenario D (high stability: $\kappa = \gamma = \rho = \phi = 0.90$), is the more extreme scenario where there is no fade-away effect present both for the OLS and IPW estimates. Under this extreme scenario, the size of the initial non-response bias of the OLS estimate remains the same in all the panel waves. However, with respect to bias reduction, the IPW estimator (unrealistic case) performs best in reducing the impact of non-response on the estimated slope coefficient in wave 1. Although, in later waves, this advantage disappears quite soon depending on the

size of the permanent components of the covariate and the error term. Here, it is important to mention that for the unrealistic case the size of non-response bias in wave 1 is -10% while in wave 10 it is -20% (almost doubled in the last panel wave see, Figure 9 for detail). It is worth noting that the un-weighted OLS estimator is still more biased than the weighted estimator (unrealistic case: IPW 3). All the other weighting estimators (realistic cases: IPW 1 and IPW 2) perform worse than the un-weighted OLS estimator.

Overall, from our analysis we conclude that weighting is only beneficial in reducing the bias of the parameter estimates if we use the correct weighting model, otherwise, it may enhance bias of the estimates. The IPW estimators are the commonly used estimators in a situation where a large part of the data is missing due to non-response. These estimators are very helpful in reducing the bias and provide consistent parameter estimates in the simulation and empirical data if the model for response is correctly specified. Because if the model is misspecified, then the IPW estimator is likely to be more biased than the un-weighted OLS estimator. Also, the covariates of the model that are used for the construction of weights should well explain the non-response, otherwise weighting will provide larger bias than the un-weighted OLS estimator.

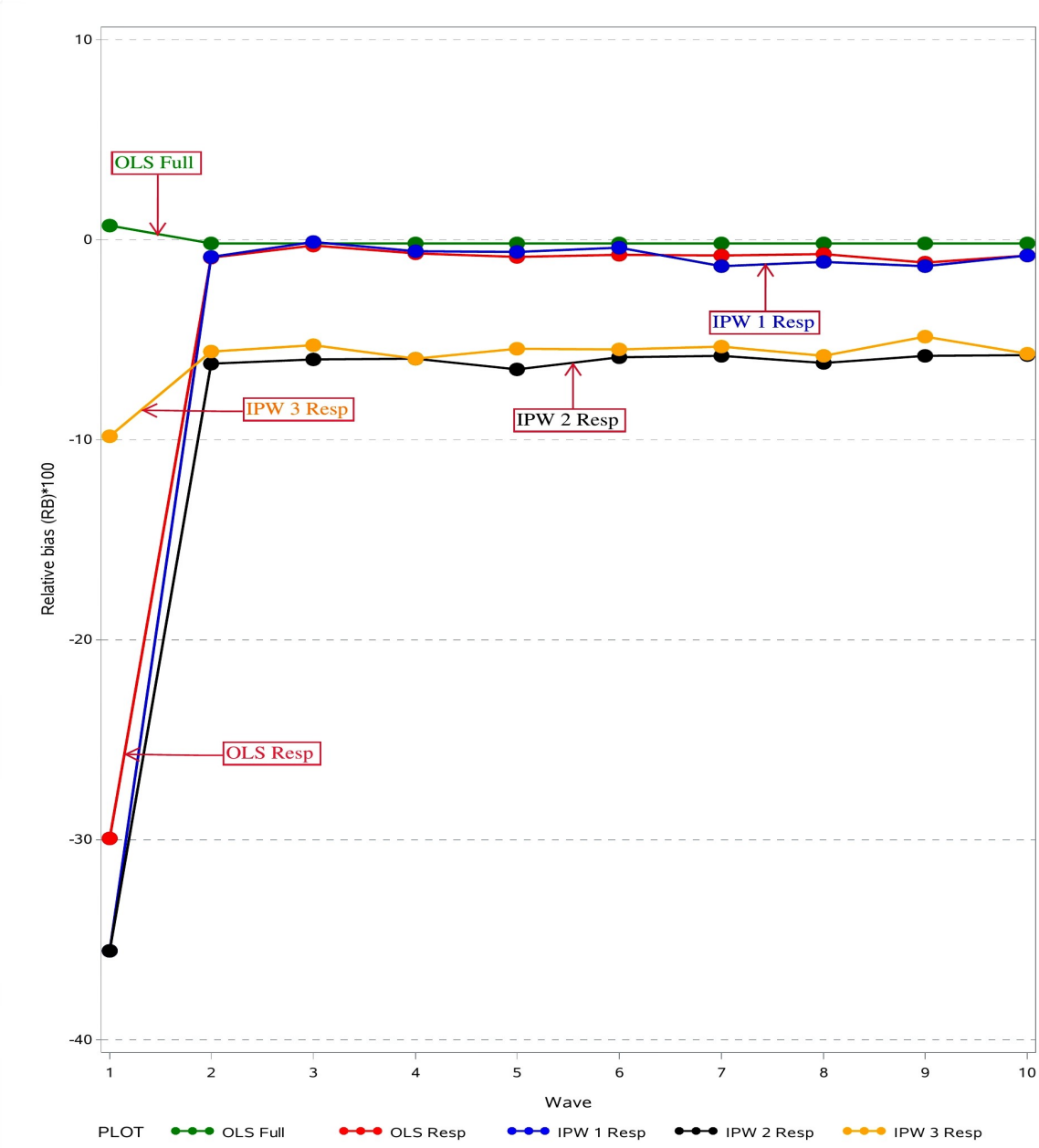


Figure 6: Fade-away effect for the OLS and IPW estimators in Scenario A ($\kappa = \gamma = \rho = \phi = 0.10$) for fix non-response parameters $\alpha = -4.50, \beta = 3.20$ and attrition parameters $\alpha^{\star} = 0.90, \beta^{\star} = 0.90$. For more detail about IPW realistic and unrealistic cases see Table 4.

Note: The vertical axis represents the percent relative bias of the estimators, while the horizontal axis represents the wave of the panel. Number of observations in the sample are $n = 1,000$, which are replicated $R = 100$ times. The bias of the OLS estimator under the Full-Sample and under the Resp-Sample are marked in green and in red colors respectively, while the bias of the IPW estimator under the Resp-Sample is marked in blue color.

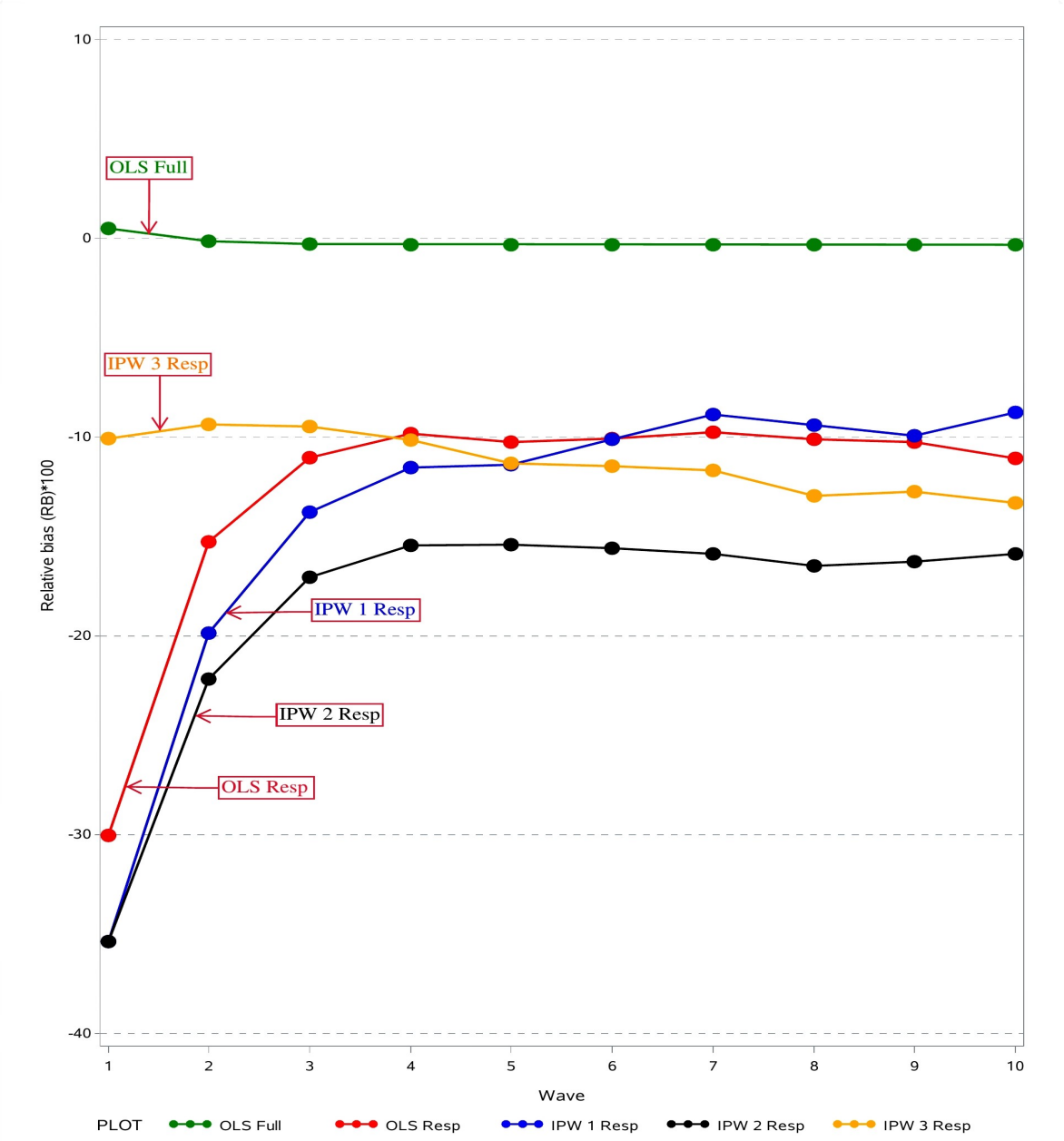


Figure 7: Fade-away effect for the OLS and IPW estimators in Scenario B ($\kappa = \gamma = \rho = \phi = 0.50$) for fix non-response parameters $\alpha = -4.50, \beta = 3.20$ and attrition parameters $\alpha^{\star} = 0.90, \beta^{\star} = 0.90$. For more detail about IPW realistic and unrealistic cases see Table 4.

Note: The vertical axis represents the percent relative bias of the estimators, while the horizontal axis represents the wave of the panel. Number of observations in the sample are $n = 1,000$, which are replicated $R = 100$ times. The bias of the OLS estimator under the Full-Sample and under the Resp-Sample are marked in green and in red colors respectively, while the bias of the IPW estimator under the Resp-Sample is marked in blue color.

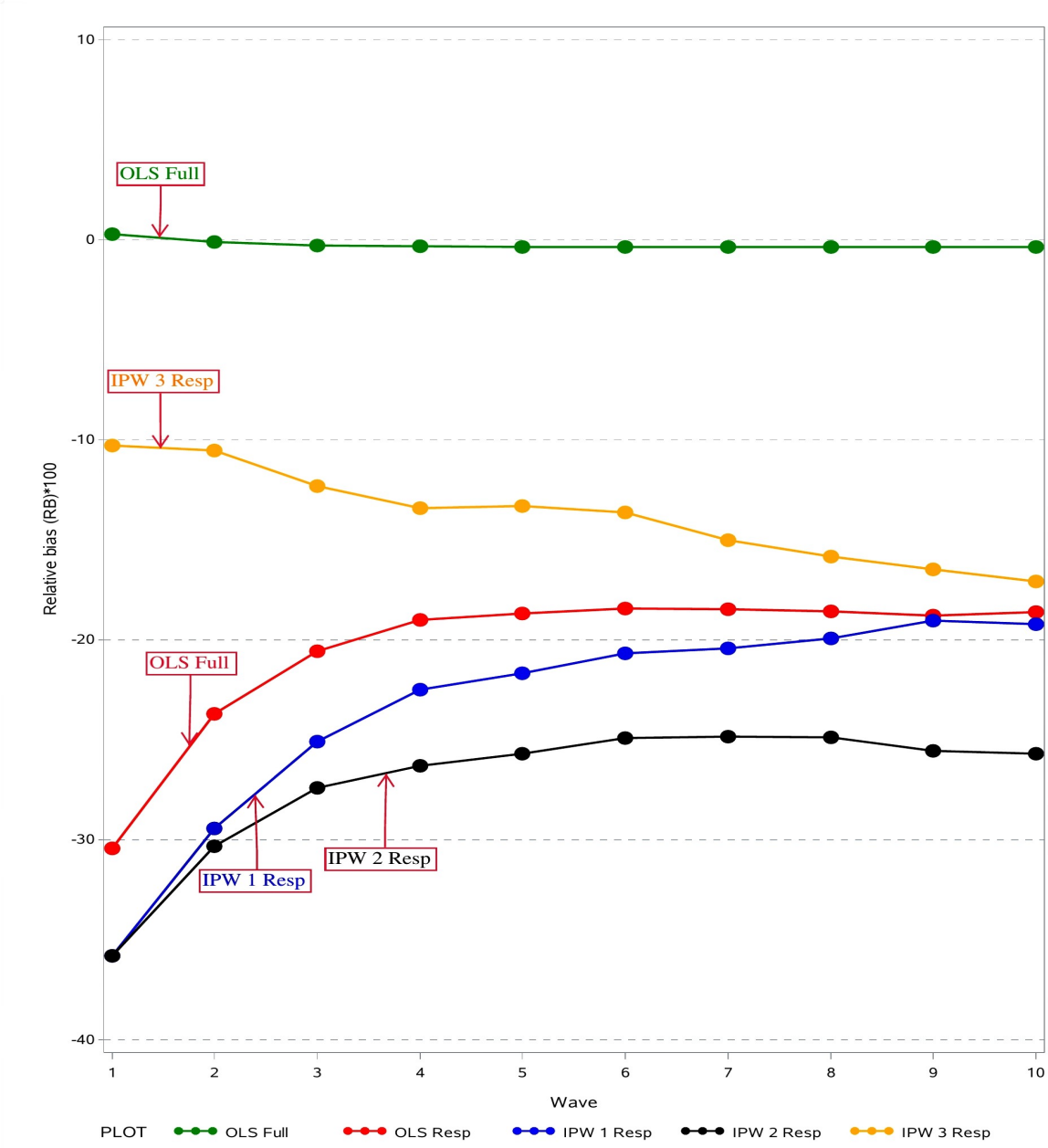


Figure 8: Fade-away effect for the OLS and IPW estimators in Scenario C ($\kappa = \gamma = \rho = \phi = 0.70$) for fix non-response parameters $\alpha = -4.50, \beta = 3.20$ and attrition parameters $\alpha^{\star} = 0.90, \beta^{\star} = 0.90$. For more detail about IPW realistic and unrealistic cases see Table 4.

Note: The vertical axis represents the percent relative bias of the estimators, while the horizontal axis represents the wave of the panel. Number of observations in the sample are $n = 1,000$, which are replicated $R = 100$ times. The bias of the OLS estimator under the Full-Sample and under the Resp-Sample are marked in green and in red colors respectively, while the bias of the IPW estimator under the Resp-Sample is marked in blue color.

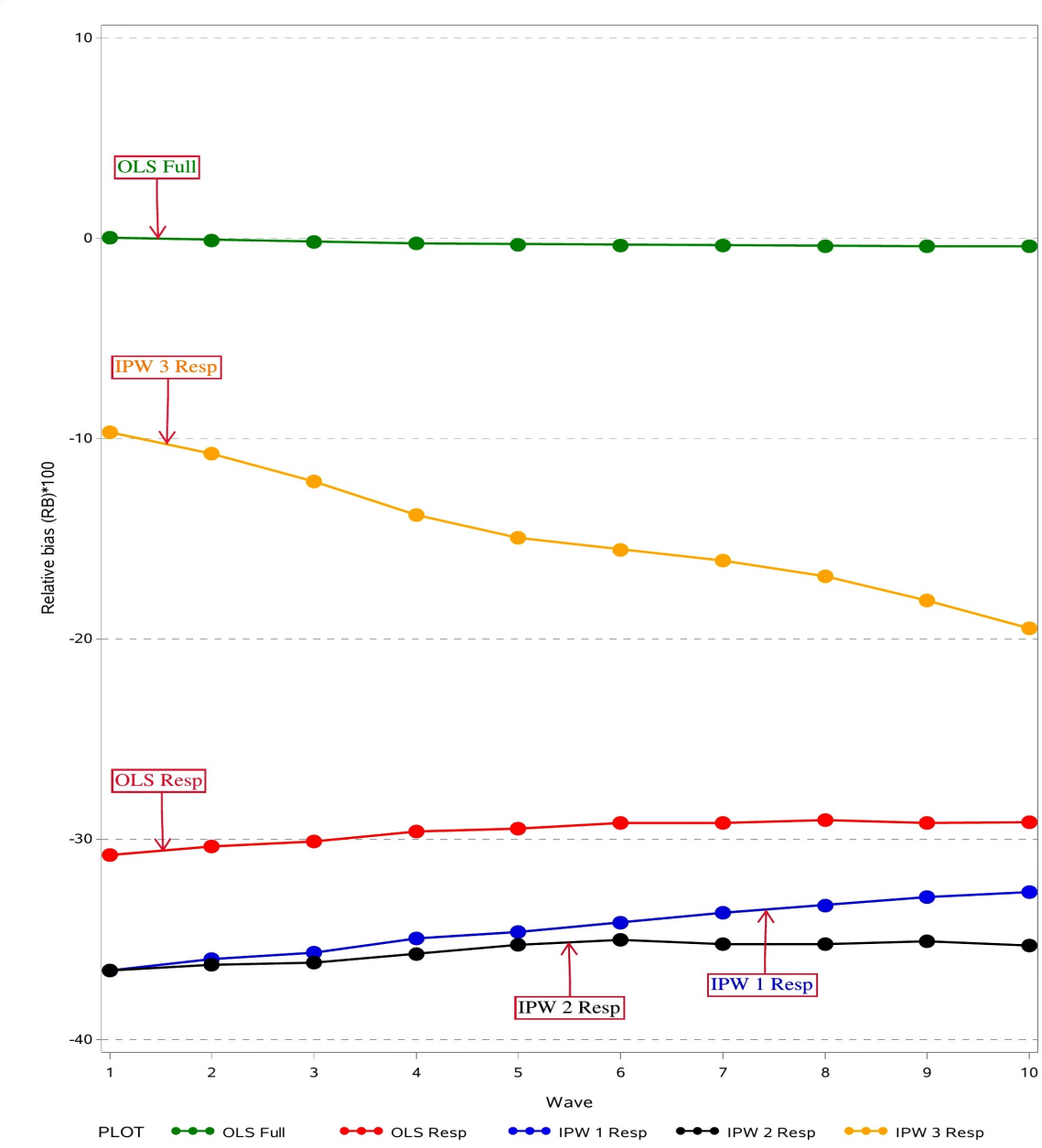


Figure 9: Fade-away effect for the OLS and IPW estimators in Scenario D ($\kappa = \gamma = \rho = \phi = 0.90$) for fix non-response parameters $\alpha = -4.50, \beta = 3.20$ and attrition parameters $\alpha^{\star} = 0.90, \beta^{\star} = 0.90$. For more detail about IPW realistic and unrealistic cases see Table 4.

Note: The vertical axis represents the percent relative bias of the estimators, while the horizontal axis represents the wave of the panel. Number of observations in the sample are $n = 1,000$, which are replicated $R = 100$ times. The bias of the OLS estimator under the Full-Sample and under the Resp-Sample are marked in green and in red colors respectively, while the bias of the IPW estimator under the Resp-Sample is marked in blue color.

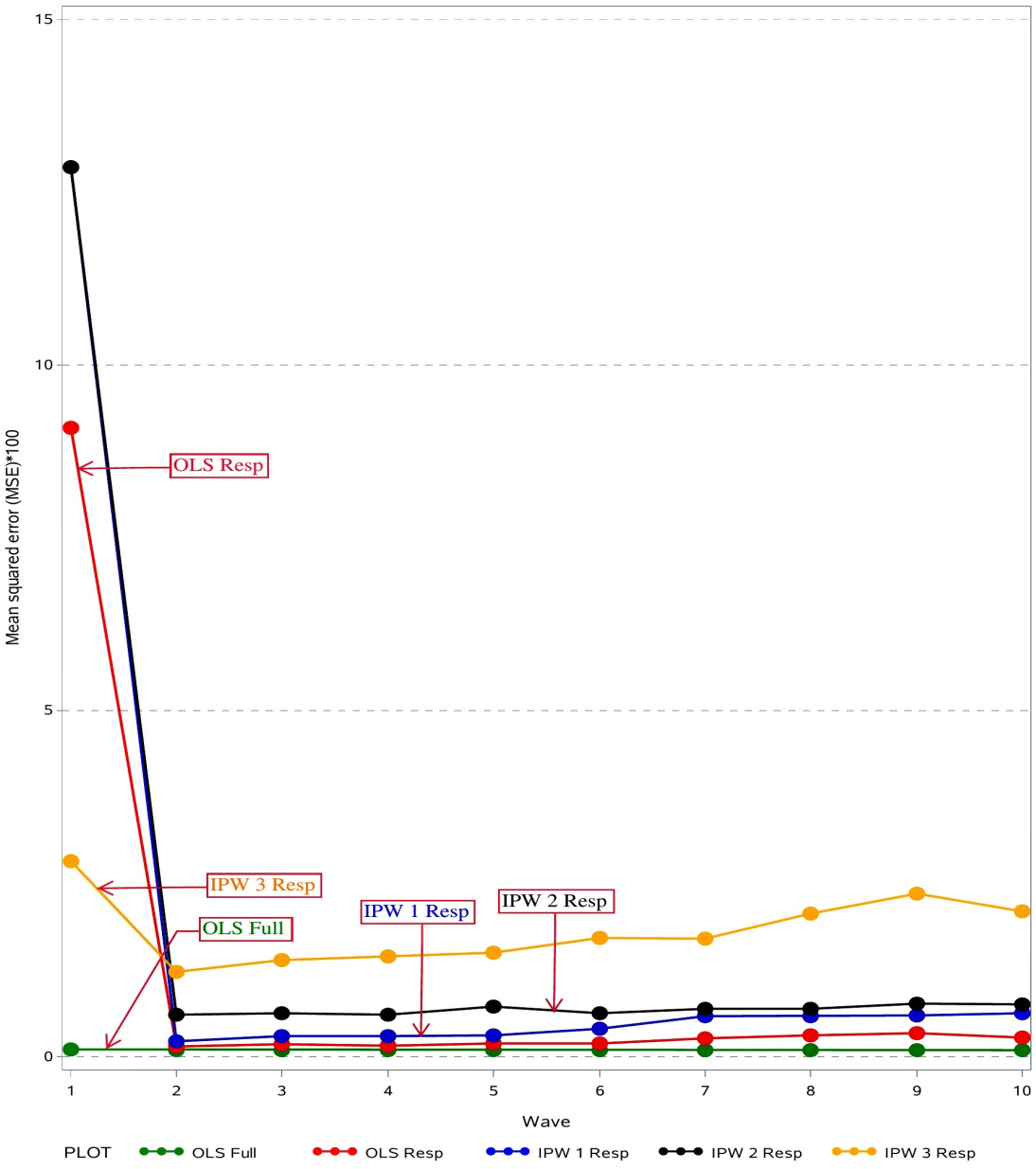


Figure 10: MSE comparison of the OLS and IPW estimators in Scenario A ($\kappa = \gamma = \rho = \phi = 0.10$) for fix non-response parameters $\alpha = -4.50, \beta = 3.20$ and attrition parameters $\alpha^{\star} = 0.90, \beta^{\star} = 0.90$. For more detail about IPW realistic and unrealistic cases see Table 4.

Note: The vertical axis shows the MSE of the estimators, while the horizontal axis shows the wave of the panel. Number of observations in the sample are $n = 1{,}000$, which are replicated $R = 100$ times. The percent MSE of the OLS estimator under the Full-Sample and under the Resp-Sample are marked in green and in red colors respectively, while the percent MSE of the weighted OLS estimator under the Resp-Sample is highlighted in blue color.

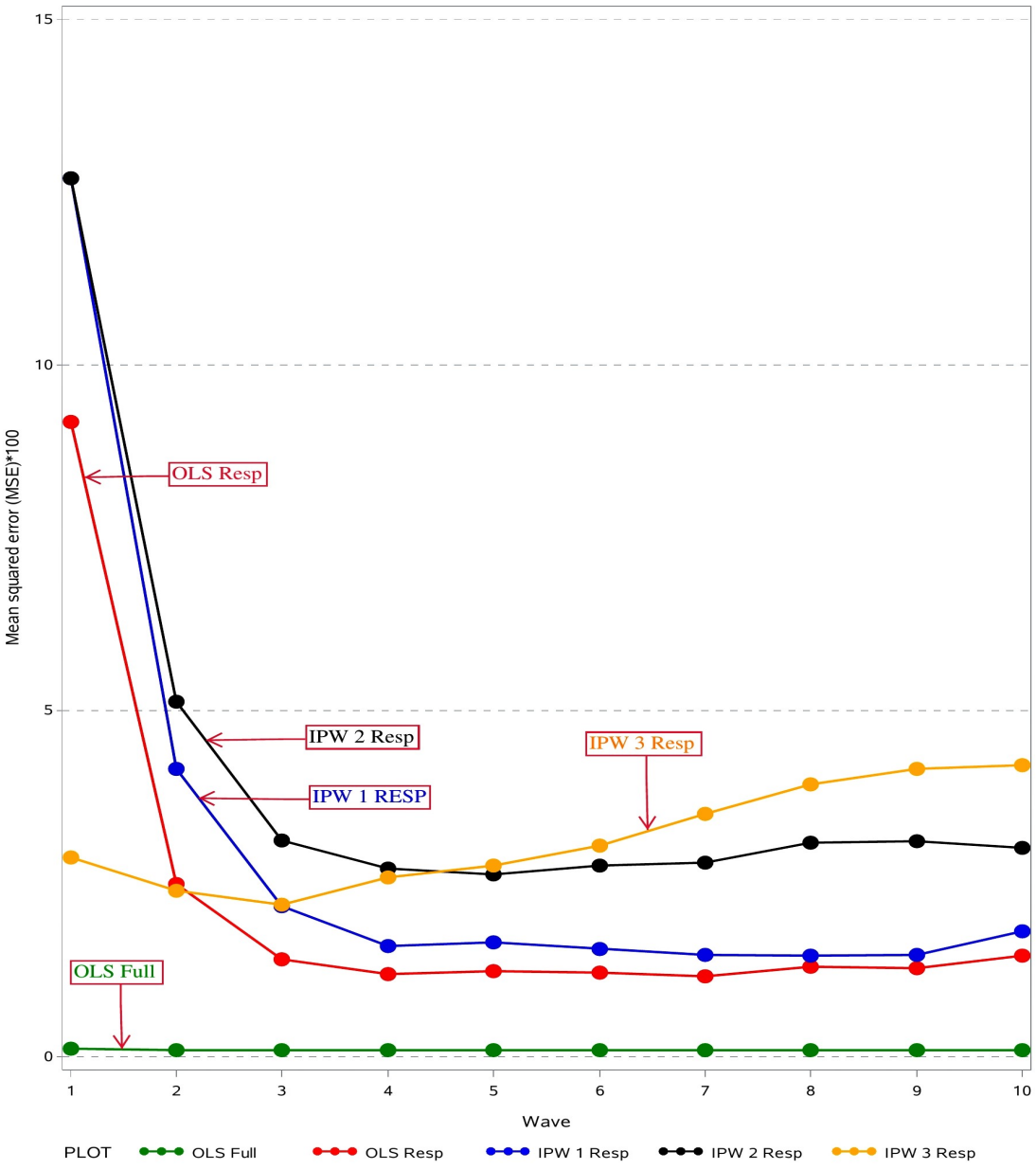


Figure 11: MSE comparison of the OLS and IPW estimators in Scenario B ($\kappa = \gamma = \rho = \phi = 0.50$) for fix non-response parameters $\alpha = -4.50, \beta = 3.20$ and attrition parameters $\alpha^{\star} = 0.90, \beta^{\star} = 0.90$. For more detail about IPW realistic and unrealistic cases see Table 4.

Note: The vertical axis shows the MSE of the estimators, while the horizontal axis shows the wave of the panel. Number of observations in the sample are $n = 1,000$, which are replicated $R = 100$ times. The percent MSE of the OLS estimator under the Full-Sample and under the Resp-Sample are marked in green and in red colors respectively, while the percent MSE of the weighted OLS estimator under the Resp-Sample is highlighted in blue color.

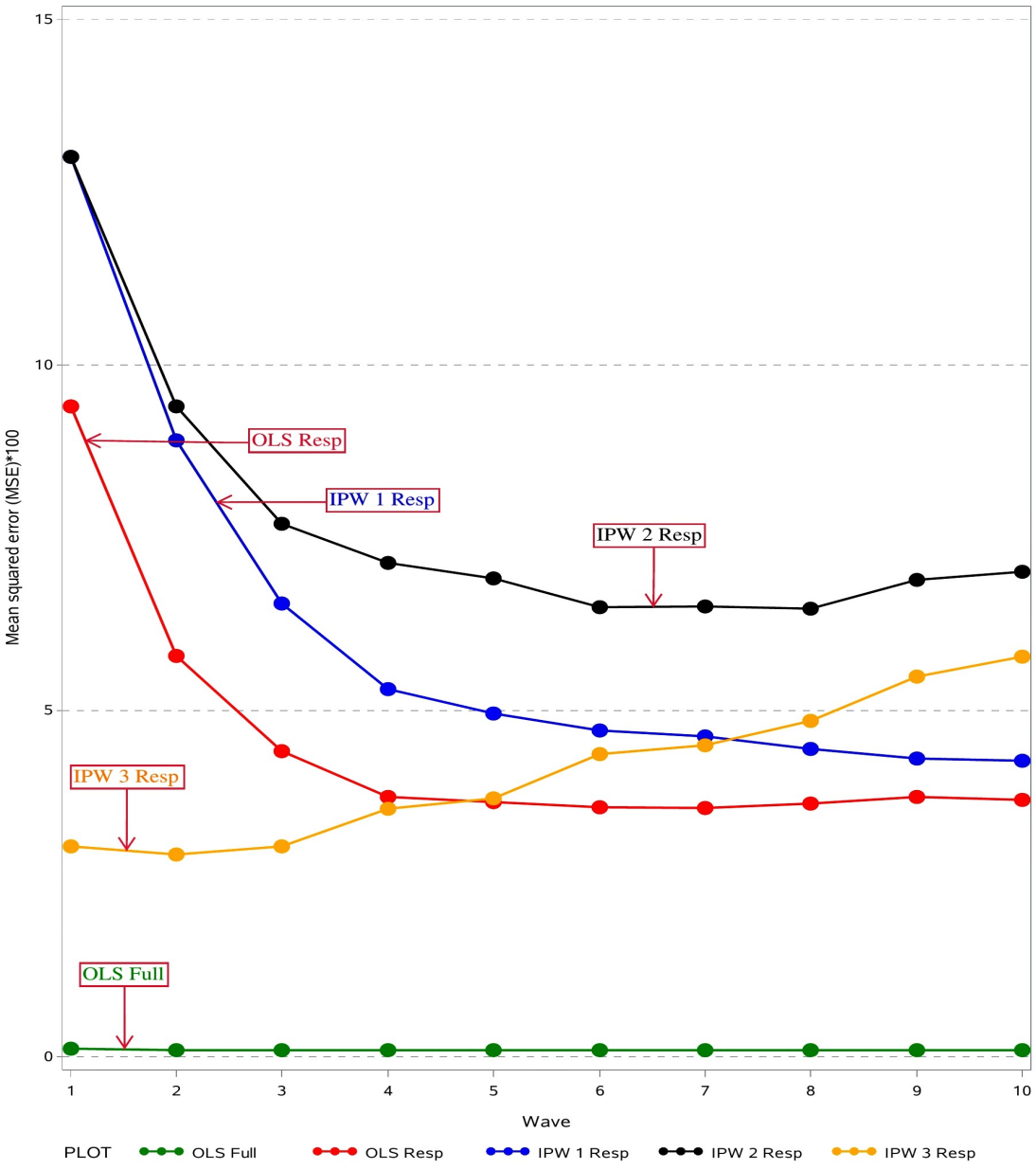


Figure 12: MSE comparison of the OLS and IPW estimators in Scenario C ($\kappa = \gamma = \rho = \phi = 0.70$) for fix non-response parameters $\alpha = -4.50, \beta = 3.20$ and attrition parameters $\alpha^{\star} = 0.90, \beta^{\star} = 0.90$. For more detail about IPW realistic and unrealistic cases see Table 4.

Note: The vertical axis shows the MSE of the estimators, while the horizontal axis shows the wave of the panel. Number of observations in the sample are $n = 1,000$, which are replicated $R = 100$ times. The percent MSE of the OLS estimator under the Full-Sample and under the Resp-Sample are marked in green and in red colors respectively, while the percent MSE of the weighted OLS estimator under the Resp-Sample is highlighted in blue color.

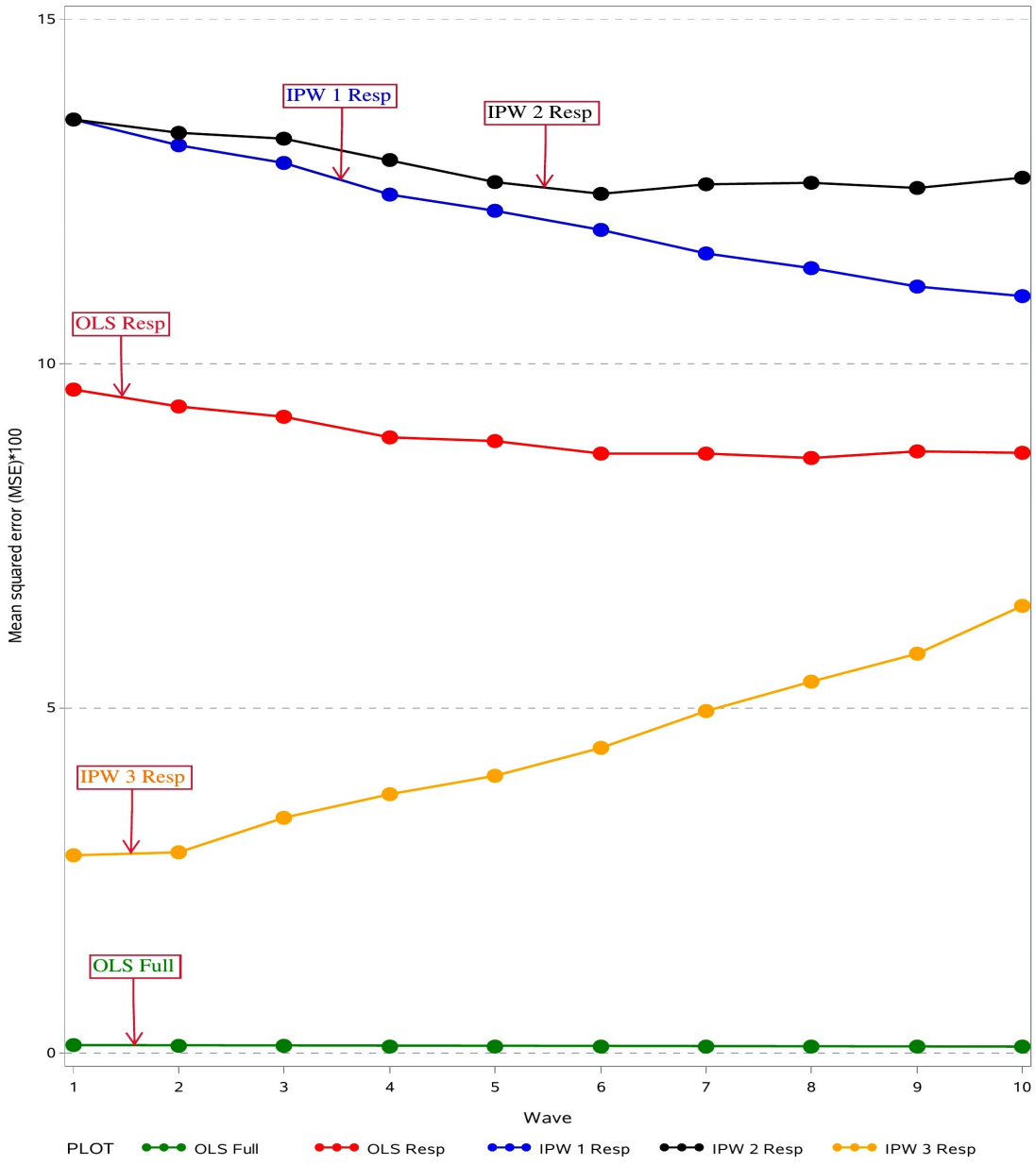


Figure 13: MSE comparison of the OLS and IPW estimators in Scenario D ($\kappa = \gamma = \rho = \phi = 0.90$) for fix non-response parameters $\alpha = -4.50, \beta = 3.20$ and attrition parameters $\alpha^{\star} = 0.90, \beta^{\star} = 0.90$. For more detail about IPW realistic and unrealistic cases see Table 4.

Note: The vertical axis shows the MSE of the estimators, while the horizontal axis shows the wave of the panel. Number of observations in the sample are $n = 1,000$, which are replicated $R = 100$ times. The percent MSE of the OLS estimator under the Full-Sample and under the Resp-Sample are marked in green and in red colors respectively, while the percent MSE of the weighted OLS estimator under the Resp-Sample is highlighted in blue color.

3.6. Fade-away effect for the panel model estimators

In the previous sections of this chapter, we produced very interesting results on the fade-away effect in the case of cross-sectional estimates. The effect was checked for different model stability of the covariate and the error term (Scenario A-G), with and without any attrition pattern. As we know that under the strict exogeneity assumption model parameters can be consistently estimated by OLS. However, if this assumption is violated then the OLS regression suffers from endogeneity problem, and we know that when there are endogeneity issues OLS provides biased and inconsistent parameter estimates. For an in-depth discussion concerning endogeneity issues, see, e.g., Antonakis et al. (2014). Also, the OLS regression doesn't control for the individual effects which may affect the estimates. Therefore, to obtain consistent estimators, we use panel data models that allow for the individual effects that capture the unobserved/time-invariant effects, which may or may not be correlated with the observed model covariates. We estimate two linear panel models: a random effects (RE) model and a fixed effects (FE) model (details are given in Subsection 2.3.3 of Chapter 2).

As mentioned in Section 3.2 of this chapter a necessary condition for the fade-away effect is to assume that non-response at the Resp-Sample is non-ignorable for the estimation of the population parameters. Therefore, for the estimation of initial non-response, we use a binary logit model which is defined by:

$$P(R_{i,1}=1|Y_{i,1}) = \frac{\exp(\alpha+\beta Y_{i,1})}{[1+\exp(\alpha+\beta Y_{i,1})]}, \tag{3.9}$$

To introduce an initial non-response of substantial size at the start of the Resp-Sample, we choose $\alpha=-4.50$ and $\beta=3.20$ from which results in a response rate of 35%.

Further, for panel attrition after wave 1 ($t>1$) we use the following attrition model which is defined by:

$$P(R_{i,t}=1|Y_{i,t},R_{i,1}=1,R_{i,2}=1,R_{i,3}=1,\ldots,R_{i,t-1}=1) = \frac{\exp(\alpha^\star+\beta^\star Y_{i,t})}{[1+\exp(\alpha^\star+\beta^\star Y_{i,t})]}, \tag{3.10}$$

where $t = 2, 3, 4, \ldots, 10$.
For attrition parameters $\alpha^{\star} = 0.90$ and $\beta^{\star} = 0.90$ the model result in an attrition rate of about 5-10% depending on the stability of the covariates and the error term. In order to obtain more stable results on the fade-away effect, we have done the analysis using $R = 100$ Monte Carlo replications.

Contrary, to the cross-sectional OLS and IPW estimators where the simulation database changes from wave to wave, here we use panel estimators of different lengths. The length consists of the number of all panel waves which enter the panel estimators from the simulation database. Therefore, to check the size of bias of the panel estimators and its fade-away effect under different panel lengths we start as follows: First, we obtained the bias of the panel estimator for the first 2 waves longitudinal data. We represent this by length 2. We then "merge" the data in wave 3 to the first 2 waves longitudinal data (length 2) and obtained the bias of the estimators based on 3 waves longitudinal data. We represent this by length 3. This process is repeated until longitudinal length 10 when the whole database from wave 1 to wave 10 is covered.

The properties of theses estimators in terms of bias and MSE under Scenario A-D are given in Table 34 to Table 37 in Section A.3 of Appendix A. Using the results of theses tables we plot the results in the following figures. In Figure 14 to Figure 17 we plot the bias of the pooled OLS estimator, the RE model estimator, and the FE Within model estimator over a different length of the panel. The vertical line shows the bias of the estimator, while the horizontal line shows the length of the panel. As the Full-Sample consists of information from both the respondents and the non-respondents, there is virtually no bias under this sample. This is true for all the considered estimators under Scenario A-D which are displayed in Figure 14 to Figure 17. Here it can be seen that there is no bias under Full-Sample of: the pooled OLS estimator (colored green marked with letter "Pooled Full"), the RE model estimator (colored orange marked with letter "Ranone Full"), and the Within estimator (colored blue marked with letter "Fixone Full"). While the corresponding Resp-Sample consists of information only on the respondents, therefore, there exists some small/large biases depending on: the magnitude of residual variance σ^2, the stability of the covariates and the error term, and the type of the estimator used.

Further, the estimation results in the figures show that the size of the bias of the pooled OLS estimator is quite large, compared to the size of the bias of the

RE model estimator and the FE Within estimator. However, these results are not surprising because the pooled OLS is simply an OLS regression run on panel data without considering the individual heterogeneity. If the appropriate model is either a RE or FE specification, then the pooled OLS estimator obtained by ignoring the panel structure of the data is usually inconsistent. Therefore, we switch to panel estimators.

Before going to discuss the fade-away effect of the panel model estimators, we give a short overview of the previously discussed results of the cross-sectional estimates which are discussed in Section 3.4 of this chapter. It has been observed that the results on the fade-away effect of the OLS estimator depend on the size of the permanent/transient components of the covariates and the error term. This has been checked for low stability, moderate stability and high stability which are plotted in Figure 1 to Figure 2. From there we can see that if the size of the permanent components is small then regression estimates having large initial biases converge to the true regression coefficients without any bias. However, for medium size permanent components, there is always some kind of permanent bias present in estimates in following panel waves. While if the size of the permanent components is large then their distribution remains stable over time and the distorting effects of initial non-response stay permanent. However, this doesn't hold for the transient components which swing into a steady-state distribution of the Markov chain. However, the use of the Within estimator is very beneficial in removing the bias of the permanent components. This is because the Within estimator eliminates the effect of permanent components from the model by subtracting the time mean from each variable in the model. The final transformed model is then consistently estimated by OLS.

Turning back to the fade-away effect of the panel estimators, we see that as the length of the panel increases the size of the initial non-response bias is expected to fade-away over longer panel lengths. However, the speed of the fade-away effect varies considerably among different estimators in Scenario A-D. For example, consider the effect of Scenario-A (low stability: $\kappa = \gamma = \rho = \phi = 0.10$) in Figure 14. We see that in the case when the stability of the covariates and error term is low, the size of the initial bias of the pooled OLS estimator under the Resp-Sample is -10.76% which faded-away to -2.67% at length 10. For easiness, we denote the bias under this sample by color red marked with the letter "Pooled Resp". The size of the initial bias of the RE model estimator is -10.01% which faded-away step by step in successive

panel length and is only -1.86% at length 10. We denote the bias under this sample by color black marked with the letter "Ranone Resp". For the Within estimator, the size of initial bias was estimated only -3.27% which melts down to -0.39% (less than 1%) at length 10. We denote the bias under this sample by color orange marked with the letter "Fixone Resp".

Comparing the bias of the estimators we get the smallest bias of the Within estimator which can be further reduced by increasing the size of the permanent components. Therefore, we consider the effect in Scenario B (medium stability: $\kappa = \gamma = \rho = \phi = 0.50$) which are plotted in Figure 15. Looking to the figure we see that for medium size stabilities of the permanent components, the size of the non-response/attrition biases of the pooled estimator is quite large than the non-response/attrition biases of the estimator in the previous Scenario A. This also holds for the non-response/attrition biases of the RE model estimator. For the pooled OLS estimator the size initial bias is estimated -22.08% which reduces to -12.67% at length 10. Due to large attrition biases, the speed of the fade-away is slower than what it was in Scenario A. Similarly, for the RE model estimator, the distorting effects of initial non-response reduce from -13.80% to -3.12% at length 10. However, if we look at the size of the initial bias of the Within estimator, it is estimated only -1% and for that reason, there is no fade-away effect present over longer panel lengths. Scenario C (high stability: $\kappa = \gamma = \rho = \phi = 0.70$) repeats Scenario B, with more stable results. The results of this scenario are documented in Figure 16.

Finally, Figure 17 reports the fade-away effect of the estimators in Scenario D (high stability: $\kappa = \gamma = \rho = \phi = 0.90$). In most extreme scenario the fade-away effect of the pooled OLS estimator completely disappeared. This doesn't hold for the fade-away effect of the RE model estimator which gradually swing into the steady-state distribution. On the contrary, in spite of high permanent stability in Scenario D ($\kappa = \gamma = \rho = \phi = 0.90$), there is no bias under the FE Within estimator. This is due to the fact that the FE Within estimator is based on OLS regression of individual changes, so the effect of the individual FE is canceled out by differences at the individual level. Therefore, if non-response is related to permanent components then under the FE Within estimator such distorting effect of the initial non-response is eliminated by taking differencing at the individual level. Therefore, the distorting effect of initial non-response bias melts down to zero at the second wave when the difference estimator is used. Hence, the use of the FE Within estimator plays

an important role in reducing the effect of non-response based on the permanent components and is therefore robust against non-response based on the permanent components.

Besides, to the estimation of non-response/attrition biases of the estimators, we also estimate the MSE's of these estimators under Scenario A-D which are displayed in Figure 18 to Figure 21, respectively. The vertical line of the figures shows the MSE of the estimator in percent, while the horizontal line shows the length of the panel. Note that the MSE's of the estimates are multiplied by 100 because they are very small. Looking at the graphical representations of the MSE's results of all the estimators under the Full-Sample is very small (close to zero), this is because there is no bias under the Full-Sample but a variance. However, estimators under the Resp-Sample have some small/large MSE's results depending on the size of permanent/transient components and the type of the estimator used. As we know that the MSE of an estimator depends on two quantities: the variance of the estimator and the squared bias of the estimator. Earlier it has been shown that the bias (negative bias) of the pooled OLS estimator is much larger than that of the RE model estimator and the Within estimator over different panel lengths. Therefore, the square of the large negative/positive biases of the pooled OLS estimator leads to large MSE's results of the estimator. Similarly, the RE model has the second larger MSE's over various panel lengths. Finally, due to low biases under the Within estimator over longer panel lengths, it has, therefore, the smallest MSE's.

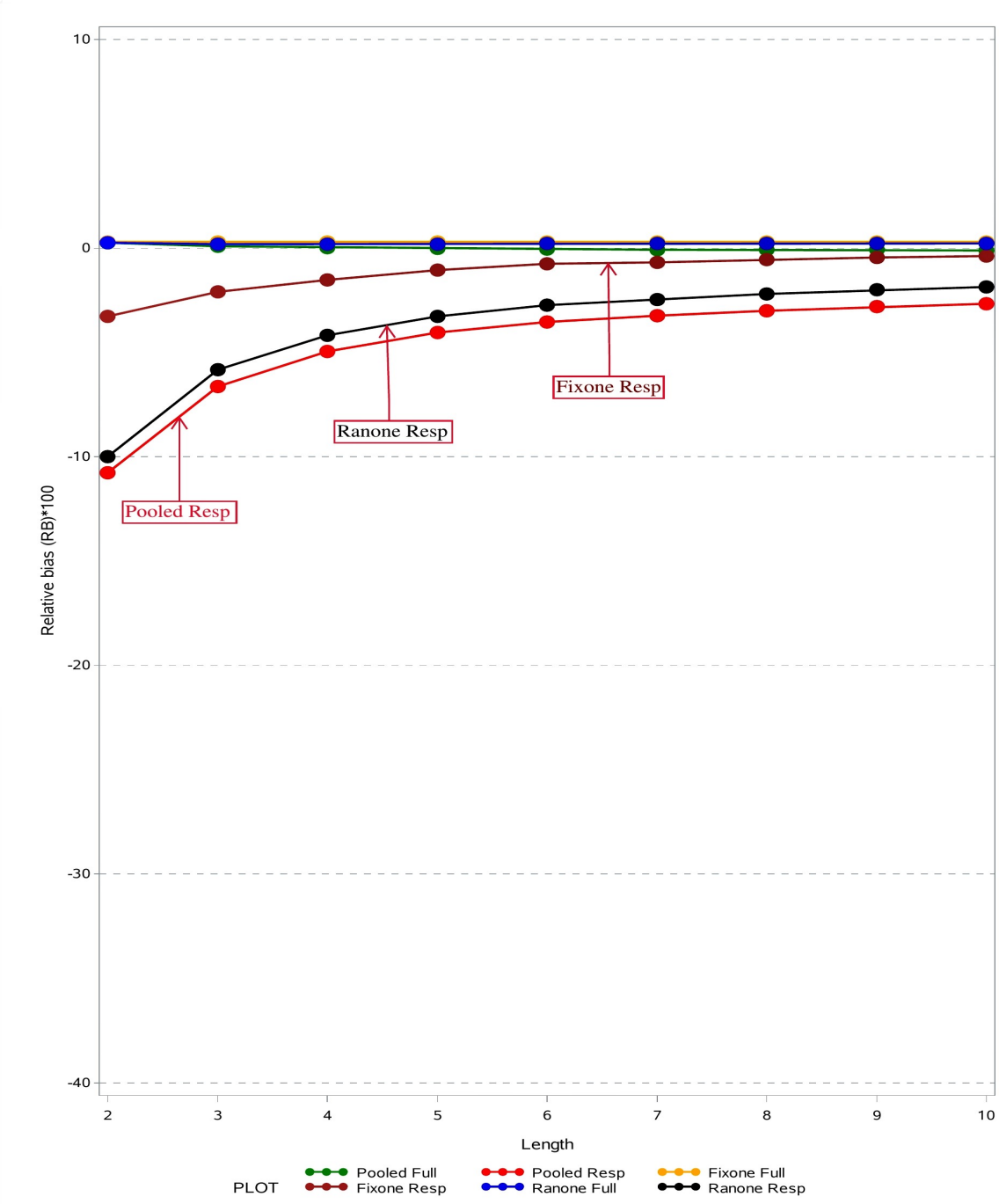


Figure 14: Graphical display of the bias of the panel model estimators in Scenario A ($\kappa = \gamma = \rho = \phi = 0.10$). Non-response rate for $\alpha = -4.50$ and $\beta = 3.20$ is 35%. Attrition rate for $\alpha^{\star} = 0.90$ and $\beta^{\star} = 0.90$ is 10%.

Note: The vertical axis represents the relative bias of the estimators in percent, while the horizontal axis represents the length of the panel. Number of observations in the sample are $n = 1{,}000$, which are replicated $R = 100$ times.

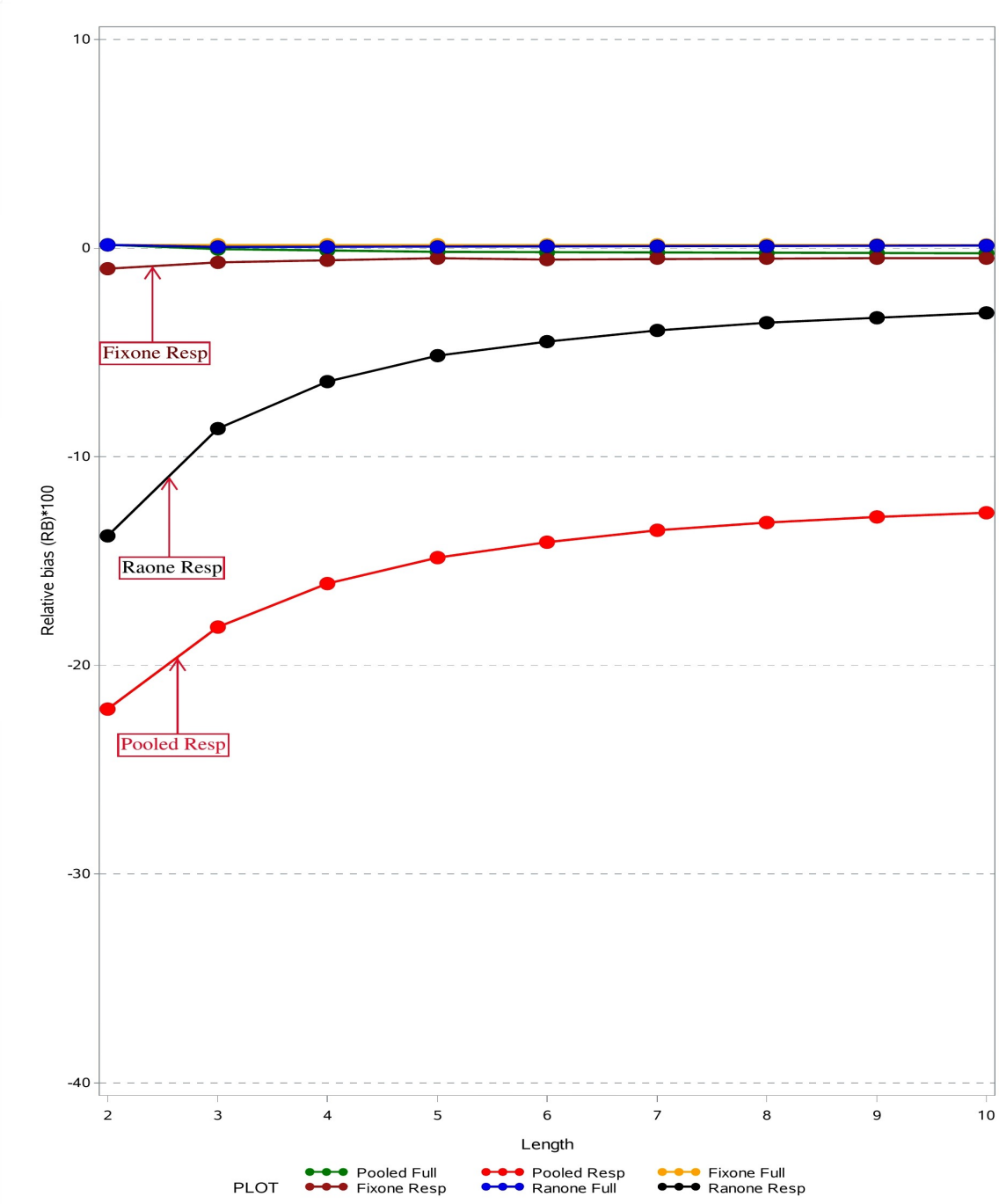


Figure 15: Graphical display of the bias of the panel model estimators in Scenario B ($\kappa = \gamma = \rho = \phi = 0.50$). Non-response rate for $\alpha = -4.50$ and $\beta = 3.20$ is 35%. Attrition rate for $\alpha^{\star} = 0.90$ and $\beta^{\star} = 0.90$ is 10%.

Note: The vertical axis represents the relative bias of the estimators in percent, while the horizontal axis represents the length of the panel. Number of observations in the sample are $n = 1{,}000$, which are replicated $R = 100$ times.

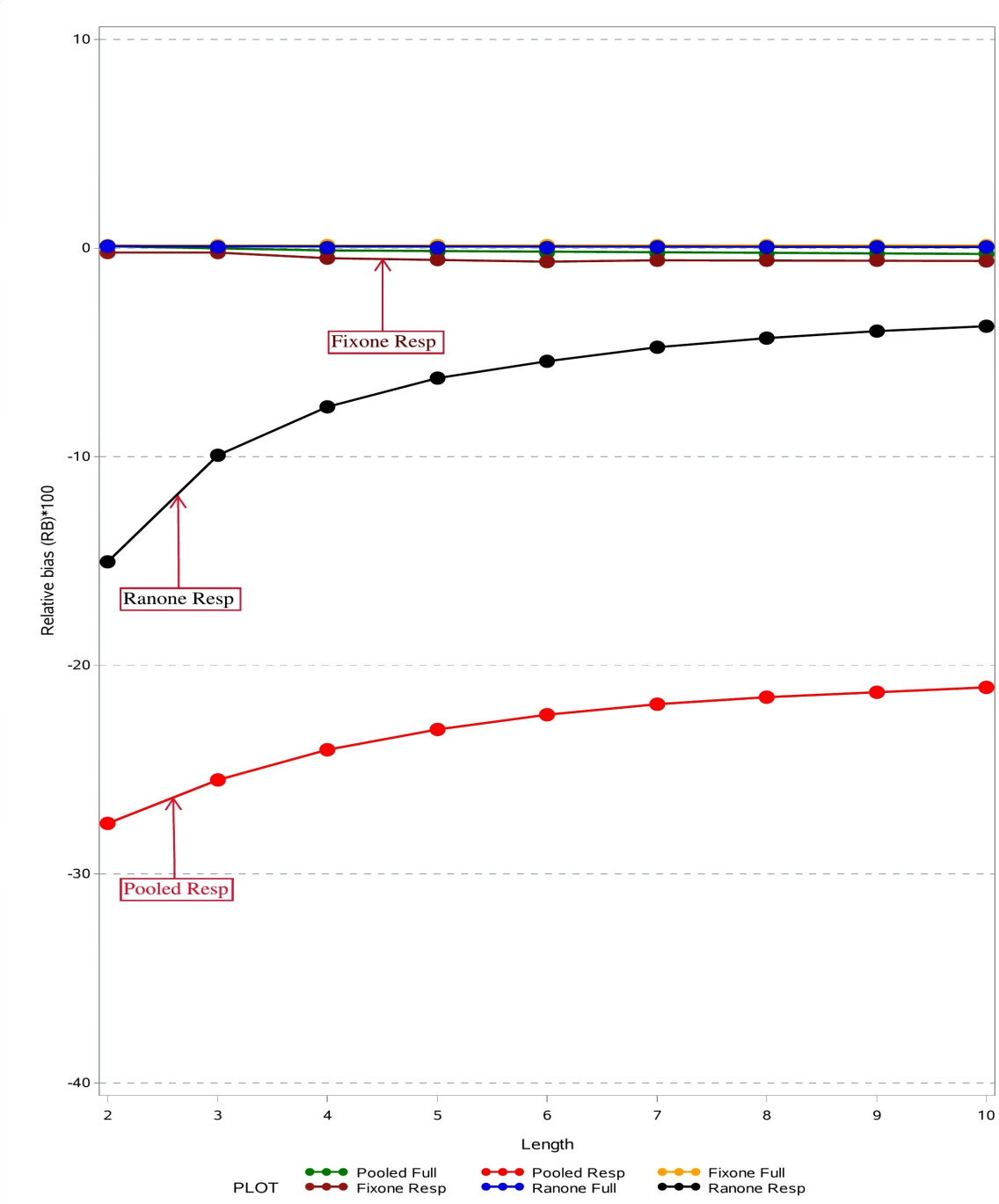


Figure 16: Graphical display of the bias of the panel model estimators in Scenario C ($\kappa = \gamma = \rho = \phi = 0.70$). Non-response rate for $\alpha = -4.50$ and $\beta = 3.20$ is 35%. Attrition rate for $\alpha^{\star} = 0.90$ and $\beta^{\star} = 0.90$ is 10%.

Note: The vertical axis represents the relative bias of the estimators in percent, while the horizontal axis represents the length of the panel. Number of observations in the sample are $n = 1{,}000$, which are replicated $R = 100$ times.

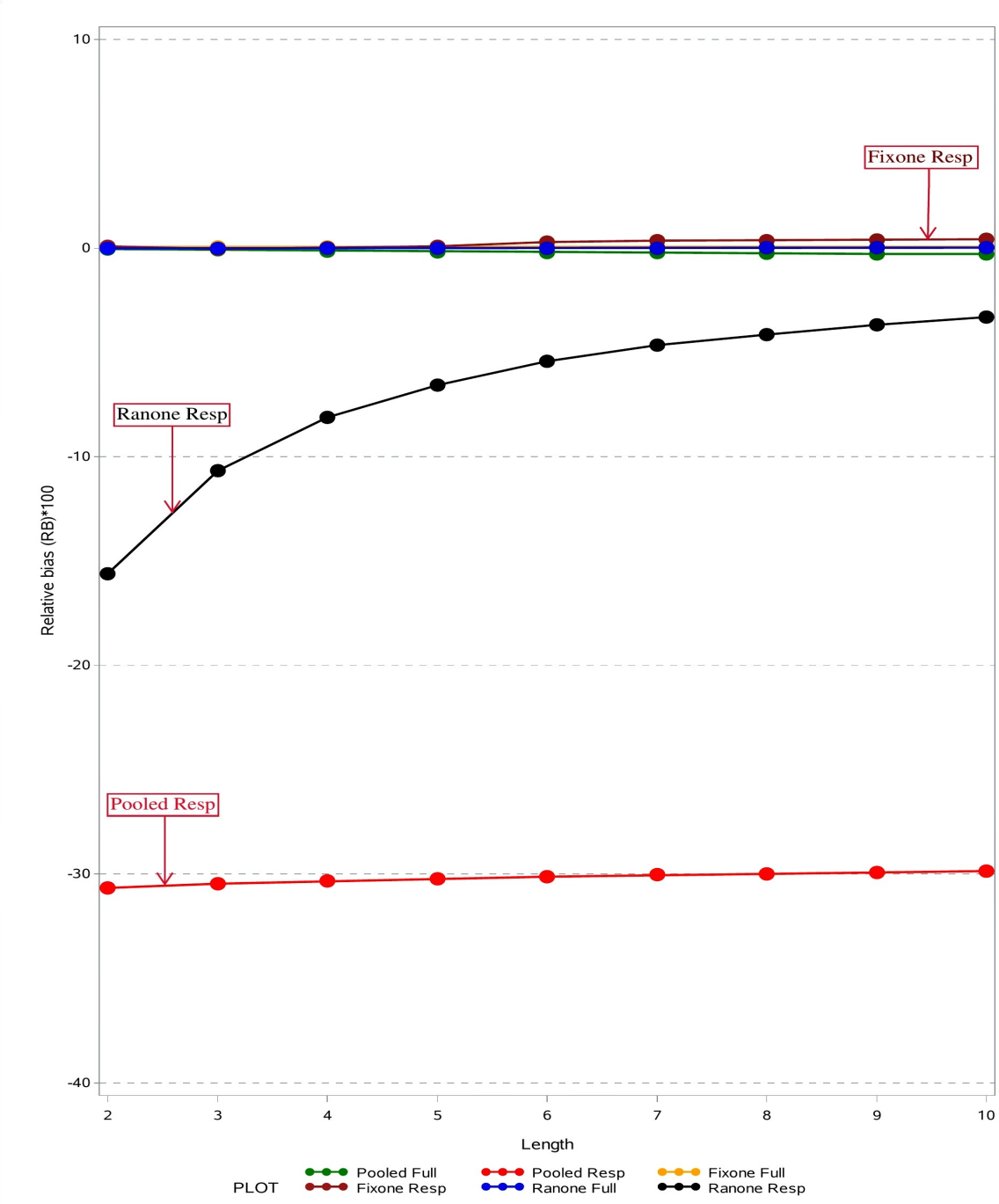


Figure 17: Graphical display of the bias of the panel model estimators in Scenario D ($\kappa = \gamma = \rho = \phi = 0.90$). Non-response rate for $\alpha = -4.50$ and $\beta = 3.20$ is 35%. Attrition rate for $\alpha^{\star} = 0.90$ and $\beta^{\star} = 0.90$ is 10%.

Note: The vertical axis represents the relative bias of the estimators in percent, while the horizontal axis represents the length of the panel. Number of observations in the sample are $n = 1{,}000$, which are replicated $R = 100$ times.

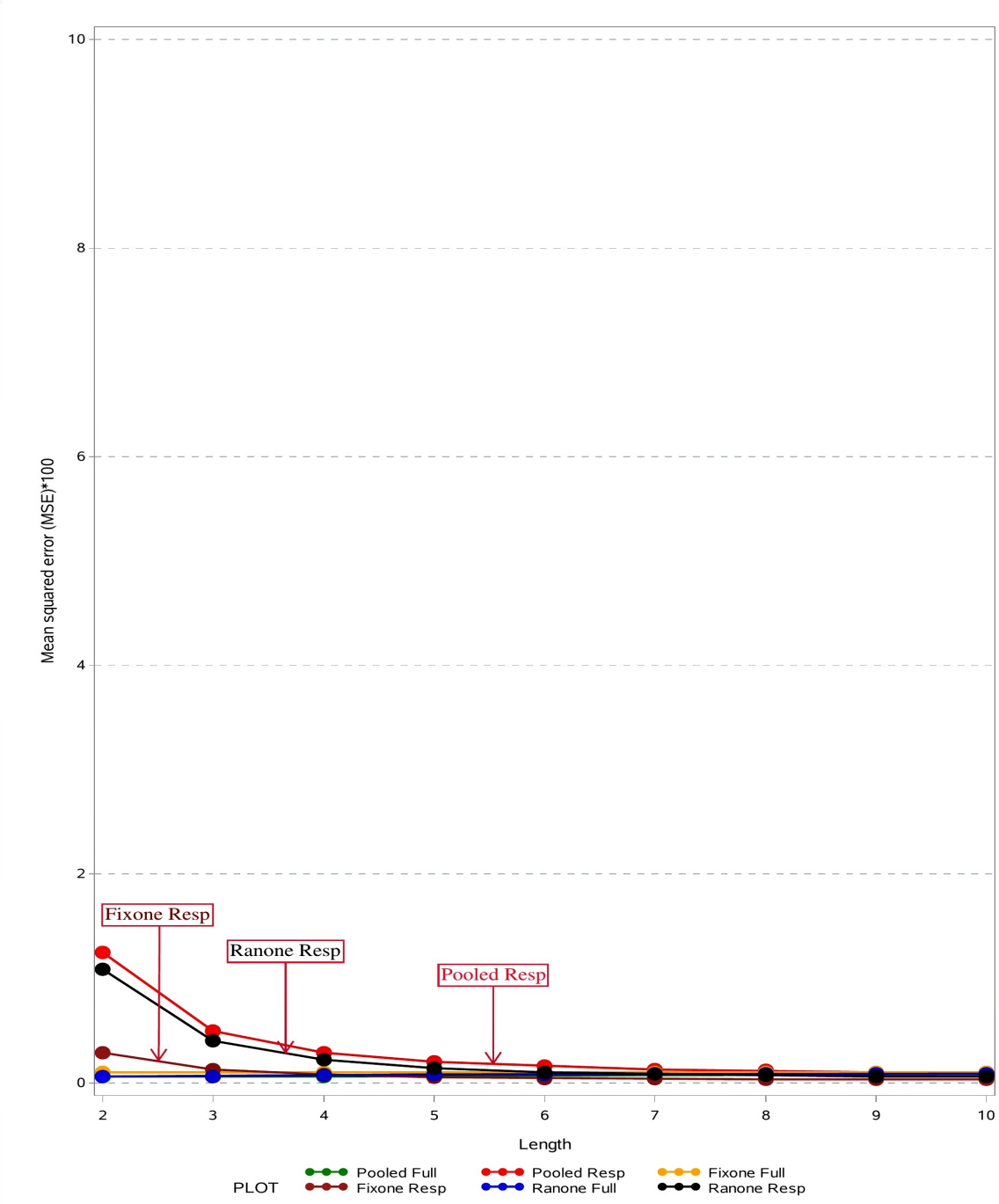


Figure 18: MSE comparison of the panel model estimators in Scenario A ($\kappa = \gamma = \rho = \phi = 0.10$). Non-response rate for $\alpha = -4.50$ and $\beta = 3.20$ is 35%. Attrition rate for $\alpha^{\star} = 0.90$ and $\beta^{\star} = 0.90$ is 10%.

Note: The vertical axis represents the MSE of the estimators in percent, while the horizontal axis represents the length of the panel. Number of observations in the sample are $n = 1{,}000$, which are replicated $R = 100$ times.

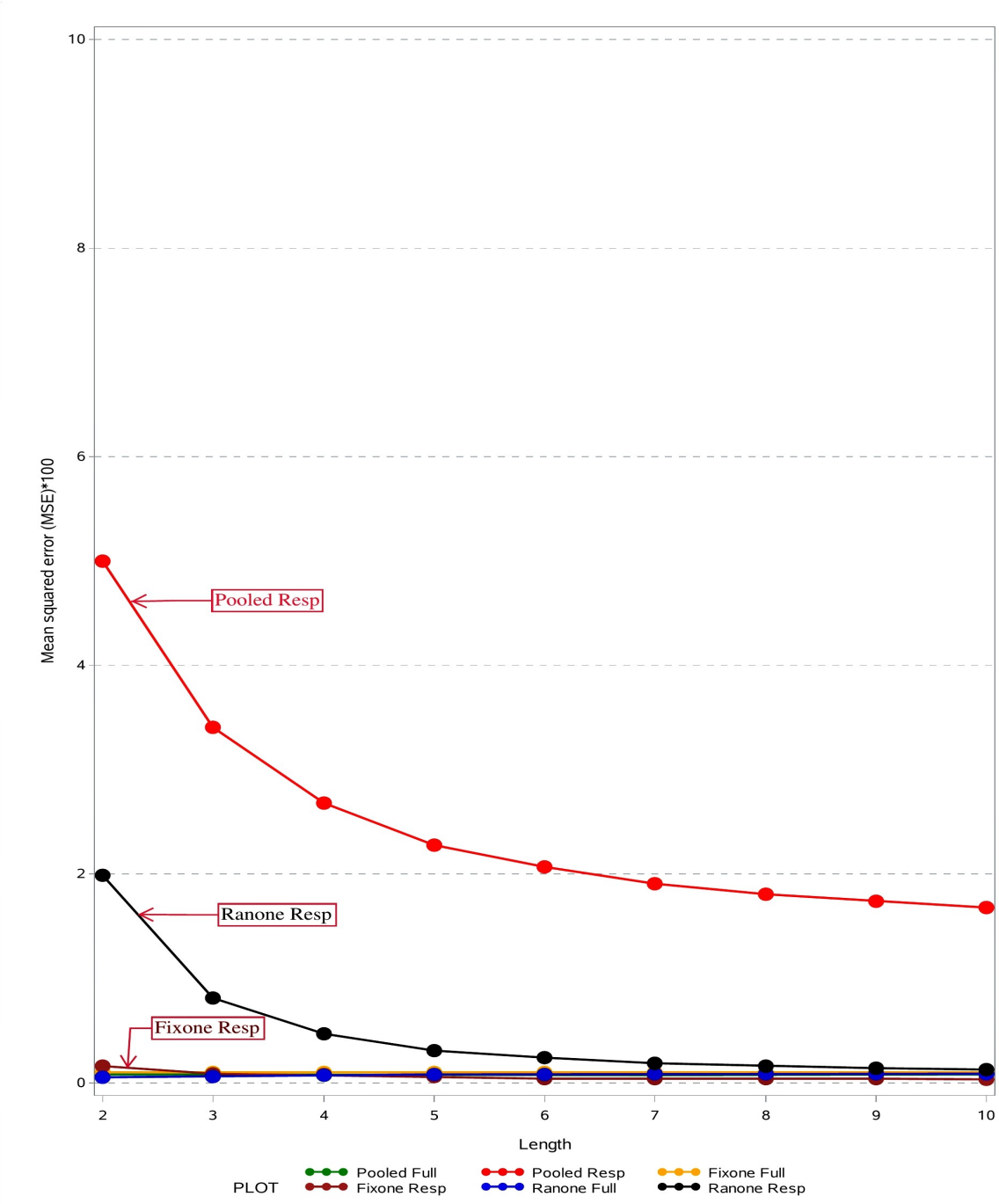


Figure 19: MSE comparison of the panel model estimators in Scenario B ($\kappa = \gamma = \rho = \phi = 0.50$). Non-response rate for $\alpha = -4.50$ and $\beta = 3.20$ is 35%. Attrition rate for $\alpha^{\star} = 0.90$ and $\beta^{\star} = 0.90$ is 10%.

Note: The vertical axis represents the MSE of the estimators in percent, while the horizontal axis represents the length of the panel. Number of observations in the sample are $n = 1{,}000$, which are replicated $R = 100$ times.

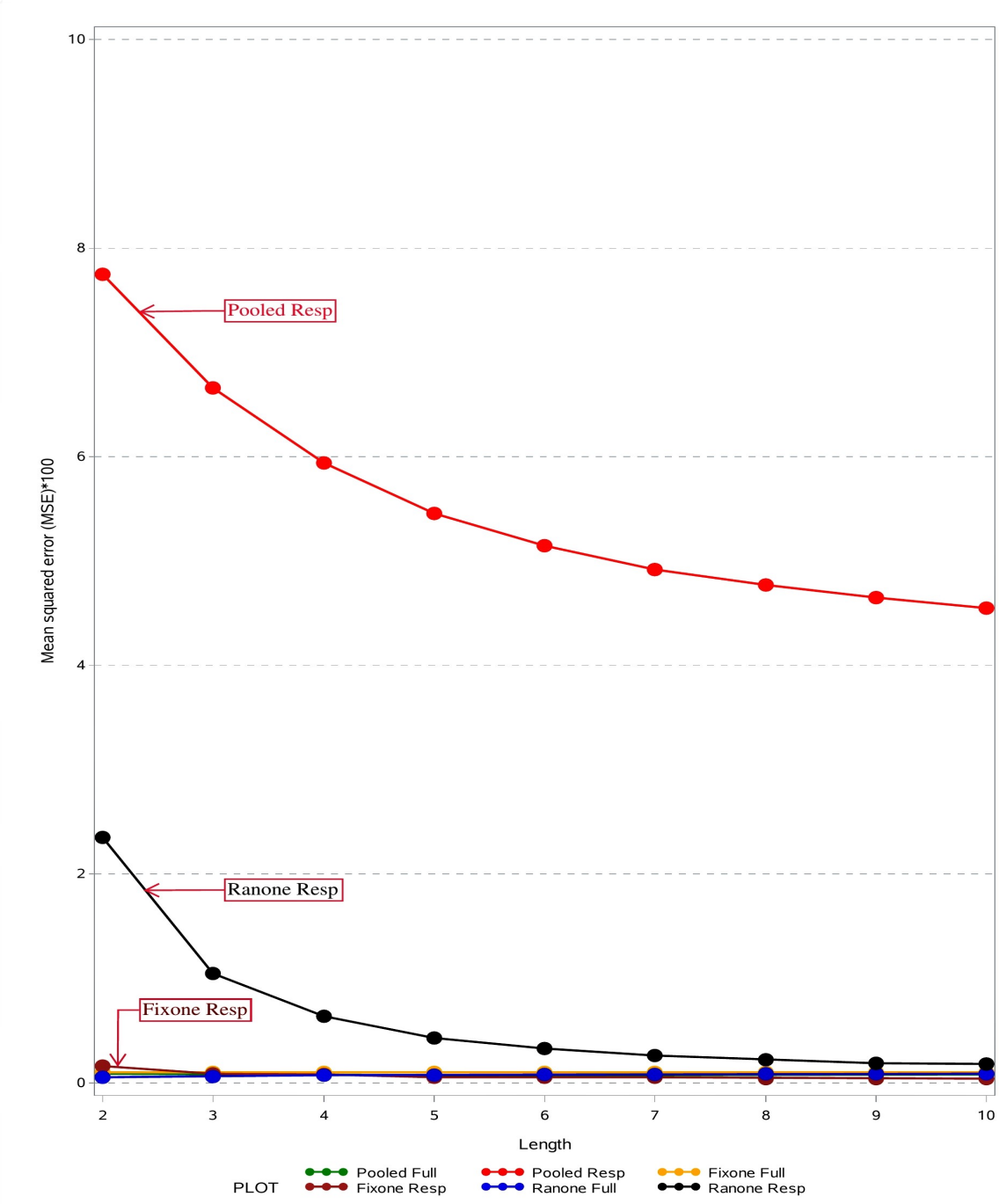


Figure 20: MSE comparison of the panel model estimators in Scenario C ($\kappa = \gamma = \rho = \phi = 0.70$). Non-response rate for $\alpha = -4.50$ and $\beta = 3.20$ is 35%. Attrition rate for $\alpha^{\star} = 0.90$ and $\beta^{\star} = 0.90$ is 10%.

Note: The vertical axis represents the MSE of the estimators in percent, while the horizontal axis represents the length of the panel. Number of observations in the sample are $n = 1{,}000$, which are replicated $R = 100$ times.

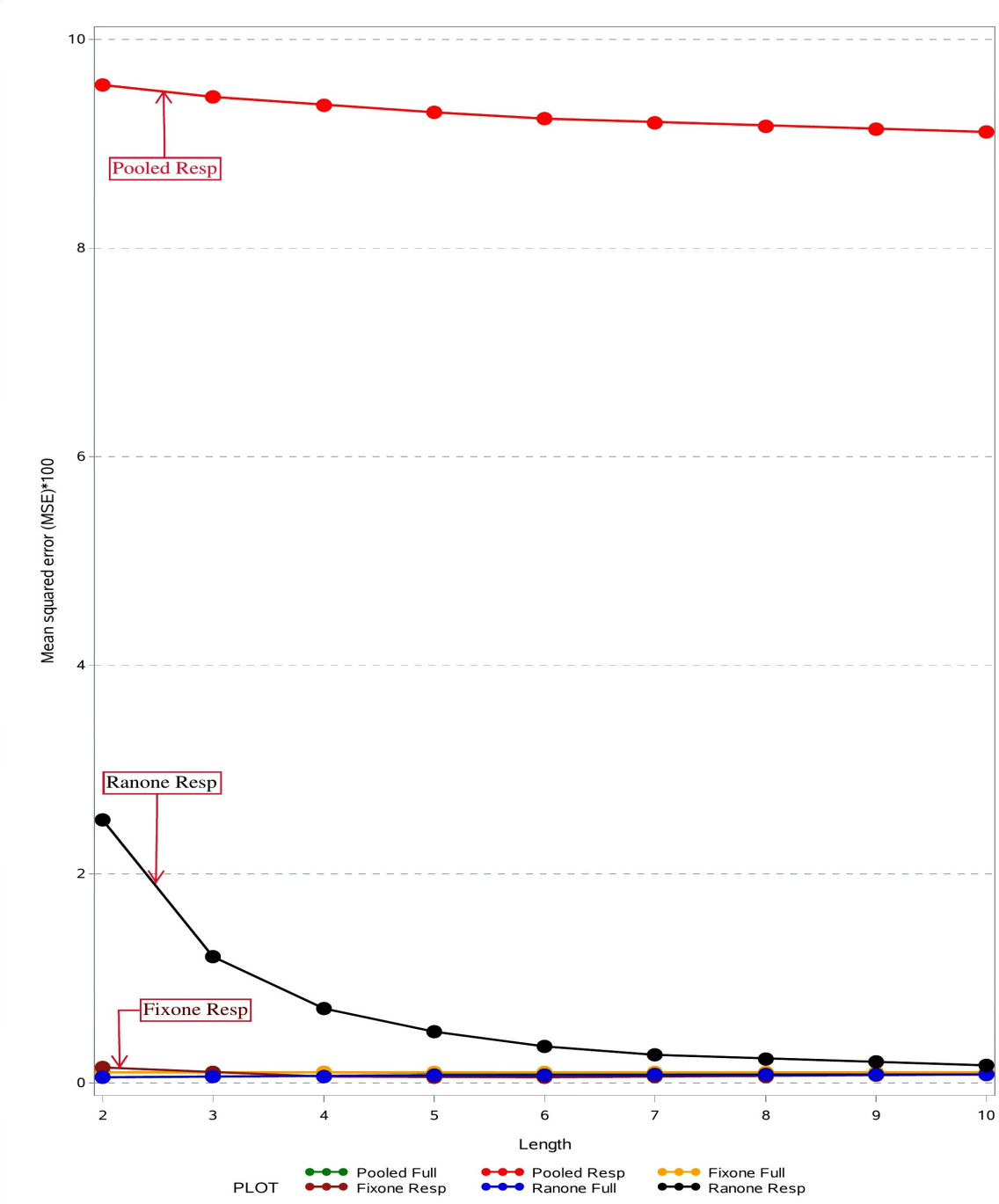


Figure 21: MSE comparison of the panel model estimators in Scenario D ($\kappa = \gamma = \rho = \phi = 0.90$). Non-response rate for $\alpha = -4.50$ and $\beta = 3.20$ is 35%. Attrition rate for $\alpha^{\star} = 0.90$ and $\beta^{\star} = 0.90$ is 10%.

Note: The vertical axis represents the MSE of the estimators in percent, while the horizontal axis represents the length of the panel. Number of observations in the sample are $n = 1{,}000$, which are replicated $R = 100$ times.

3.7. Fade-away effect for the weighted and un-weighted estimates of ordered logit model

In the previous sections of this chapter conducted a Monte Carlo simulation study to investigate the behaviour of cross-sectional and panel estimators in the context of linear regression models. For the cross-sectional case, we compared the properties of the un-weighted OLS estimator with several weighting approaches (IPW estimators). While for panel model estimators we used the pooled OLS regression estimator, the RE estimator, and the FE Within estimator and their properties in terms of bias and mean squared error are checked under the non-ignorable non-response at the start of the Resp-Sample. However, in the context of panel estimators, we don't use any weighting approach and would, therefore, be a possible topic for future research work. Therefore, this section is devoted to the estimation of the non-linear logit model, especially we will use the standard approach of latent variable formulation of the ordered logit model. This approach is more thoroughly discussed in Subsection 2.3.4 of Chapter 2. The second motivation of this section is to compare the un-weighted estimates of the ordered logit model with several weighting approaches (theses weighting approaches are presented in Table 5 of this section).

For the generation of an observed ordered variable for the continuous latent model we consider the following continuous latent model over time t:

$$Y_{i,t} = a_t + b_t X_{i,t} + e_{i,t}, \quad \text{for} \quad t = 1, 2, 3, \ldots 10. \tag{3.11}$$

where $Y_{i,t}$ is the unobserved latent response variable, but it generates an observed response variable D having J categories for the i^{th} subject through the following censoring mechanism:

$$D_{i,t} = \begin{cases} 0, & \text{if } -\infty < Y_{i,t} \leq \tau_0 \\ 1, & \text{if } \tau_0 < Y_{i,t} \leq \tau_1 \\ 2, & \text{if } \tau_1 < Y_{i,t} \leq \tau_2 \\ \ldots & \\ J, & \text{if } \tau_{J-1} < Y_{i,t} \leq \infty, \end{cases} \tag{3.12}$$

where τ's are the unknown cut-off points namely the so-called model thresholds parameters which divide the range of $Y_{i,t}$ into disjoint and exhaustive intervals such that $\tau_0 < \tau_1, \ldots, \tau_{J-1}$. Note, that the model in equation (3.12) doesn't include the constant a_t, because one can't estimate simultaneously the constant of the linear predictor as well as the thresholds parameters. This identification problem can be solved by removing either the constant of the linear predictor or setting the first threshold equal to zero.

For the ordinal response variable $D_{i,t}$ having J categories, the ordered logit model in logit form can be expressed as follows:

$$\begin{aligned} logit[P(D_{i,t} \leq j|x_{i,t})] &= log\left(\frac{P(D_{i,t} \leq j|x_{i,t})}{1 - P(D_{i,t} \leq j|x_{i,t})}\right) \\ &= \tau_j + b_t X_{i,t}, \end{aligned} \tag{3.13}$$

where $P(D_{i,t} \leq j|x_{i,t})$ is the cumulative probability of being at or below than category j, given a covariate $X_{i,t}$. τ_j are the thresholds parameters for $j = 0, 1, 2, \ldots, J$, and b_t is the logit coefficient over time.

For the data generation, we used the same simulation setup as was described in Section 3.3, i.e., we simulate a sample of size 1,000 units from the model over 100 Monte Carlo replications. Further, for the generation of non-response at the start of the Resp-Sample we use a logit model under the assumption that non-response depends on the variable of interest $D_{i,1}$ in the initial wave. This is defined by:

$$P(R_{i,1} = 1|D_{i,1}) = \frac{\exp(\alpha + \beta D_{i,1})}{[1 + \exp(\alpha + \beta D_{i,1})]}, \tag{3.14}$$

We select $\alpha = -4.50$ and $\beta = 2.00$ which results in a non-response rate of size 23%. While for panel attrition after wave 1 ($t > 1$) we use the logit regression to model the attrition probability which is defined by:

$$P(R_{i,t} = 1|D_{i,t}, R_{i,1} = 1, R_{i,2} = 1, R_{i,3} = 1, \ldots, R_{i,t-1} = 1) = \frac{\exp(\alpha^\star + \beta^\star D_{i,t})}{[1 + \exp(\alpha^\star + \beta^\star D_{i,t})]}, \tag{3.15}$$

where $t = 2, 3, 4, \ldots, 10$.

For attrition parameters $\alpha^\star = 0.01$ and $\beta^\star = 0.70$ the model results in an attrition rate of approximately 5-10% depending on the stability of the permanent and transient

components of the covariates and the error term.

Then we will compare the estimation results of the un-weighted ordered logit (UOL) model with two weighting approaches: (i) a realistic weighting scenario; (ii) and an unrealistic weighting scenario. The realistic scenario consists of two versions: in the first case (WOL 1), we correct for both the initial non-response and attrition through weighting, while in the second case (WOL 2) we control only for the initial non-response through weighting but we don't control for attrition. Under this aspect, the initial weights are constructed by using the information on the covariate $X_{i,t=1}$ which is supposed to be known for the respondents and the non-respondents. Here we take $X_{i,t=1}$ as a predictor for the unknown value of $D_{i,t=1}$. The weights in later panel waves are updated by using the information of the lagged dependent variable $D_{i,t-1}$ as an explanatory variable for attrition. In the unrealistic case (WOL 3), we use the true information on $D_{i,1}$ for the construction of non-response weighting. Similarly, for the estimation of attrition weights, we use the information on lagged dependent variable $D_{i,t-1}$ in the attrition model (similar to attrition weights in the realistic case (WOL 1)). These weighting scenarios are more clearly displayed in the following table (Table 5).

Table 5: Weighted and un-weighted estimators in ordered logit model.

	Initial non-response	Attrition
Non-response/attrition	First wave	Later waves
UOL 1 (no weighting)	-	no attrition
UOL 2 (no weighting)	-	-
WOL 1 (realistic case)	X_1	D_{t-1}
WOL 2 (realistic case)	X_1	-
WOL 3 (unrealistic case)	D_1	D_{t-1}

To judge the merits of the weighted and the un-weighted estimators we compared its properties in terms of bias and MSE. For this, Table 38 to Table 41 in Section A.4 of Appendix A contains the bias and the MSE of the estimators under Scenario A-D. As the main theme of the thesis revolves around the fade-away of bias, therefore, the bias of the estimators over various panel waves are plotted in 22 to Figure 25, respectively. The vertical axis of the figures shows the percent relative bias of the

estimator, while the horizontal axis shows the wave of the panel. The points (as shown in colored dots) of the graph display the bias in wave t, where $t = 1, 2, 3, \ldots, 10$, implied by our regression model. The green solid line marked with the letter "UOL 1 Full" corresponds to the bias of the un-weighted estimator of the ordered logit model under the Full-Sample. The red solid line marked with the letter "UOL 1 Resp" corresponds to the bias of the un-weighted estimator of the ordered logit model under the Resp-Sample. While the blue solid line marked with the letter "UOL 2 Resp" corresponds to the bias of the un-weighted estimator of under the Resp-Sample (no attrition is used in this scenario). For the realistic weighting scenarios, the black solid line marked with the letter "WOL 1 Resp" corresponds to the bias of the weighted estimator (sequential weights with attrition) under the Resp-Sample, while, the orange solid line marked with the letter "WOL 2 Resp" corresponds to the bias of the weighted estimator (only initial non-response weights with attrition) under the Resp-Sample. Similarly, for the more unrealistic weighting case, the pink solid line marked with the letter "WOL 3 Resp" corresponds to the bias of the estimator under the Resp-Sample.

By comparing the fade-away effect of the various estimators we see a considerable variation in the strength of the fade-away effect. However, the strength of the fade-away effect of the un-weighted estimator is higher than the strength of the fade-away effect in all weighting scenarios including the unrealistic one. This is because under the realistic cases we used a wrong weighting model for the probability of response. These correction methods work well if the model for response is correctly specified, otherwise, the weighting can make the results even more biased. In the unrealistic case with full information on $D_{i,1}$ the bias vanishes completely in wave 1. However, such gain from weighting is completely disappeared quite soon after wave 1 depending on the size of permanent components. For illustration, consider the fade-away effect in Scenario A (low stability: $\kappa = \gamma = \rho = \phi = 0.10$) which is plotted in Figure 22. It can be shown from Figure 22 that the initial non-response biases of the weighted and un-weighted estimators do fade-away rapidly in the same fashion in later panel waves. Also in the unrealistic case, the regression coefficient at the initial wave is estimated with zero bias, this happens because we use the true information on $D_{i,1}$ at the start of the panel. Although, in later waves, coefficients are estimated with some bias. Moreover, the speed of convergence to a steady-state distribution is fast, the main reason for this fast turn-over is that the size of the

permanent components is very small.

Therefore, we consider the fade-away effect in medium size stability scenario of permanent components (Scenario B: $\kappa = \gamma = \rho = \phi = 0.50$) which are plotted in Figure 23. It can be seen from the figure that in the case of no attrition the size of the initial bias of the un-weighted estimator (colored red marked with the letter "UOL 1 Resp") is -15.39% which reduces to -4.10% in wave 10. While in the case of attrition the initial bias of the un-weighted estimator (colored blue marked with the letter "UOL 2 Resp") is -15.39% which increases to -20.04% in wave 10. All the weighting estimators including the unrealistic one perform worse than the un-weighted estimator. For the realistic case, the effect of initial bias increases from -15.79% (colored orange marked with the letter "WOL 1 Resp") to 25.15% in wave 10. Similarly in the unrealistic case, the size of initial bias is 0.92% (colored pink marked with the letter "WOL 3 Resp") increases to 25.34% in wave 10. This is clear from all the figures that biases of the un-weighted estimates are smaller than the biases of the weighted estimators. This holds for the more extreme stability scenarios (Scenario C-D), where weighting makes the results even more biased (see Figure 24 and Figure 25 for detail).

Here it should be noted that the sign and direction of the bias changes in later waves. This is due to the fact attrition goes into the same direction e.g., consider Figure 26. In Figure 26, we compare the distribution on the state space of the Full-Sample and the Resp-Samples in different panel waves. The x-axis of the figure shows the different categories of the observed response variable. While the y-axis shows the percent total frequency of the Full-Sample and the Resp-Samples participants who participate in i^{th} state in wave t $(t = 0, 1, 2, \ldots, 10)$. In the figure, there are 11 different colored lines for each wave $(t = 0, 1, 2, \ldots, 10)$ starting from wave 0 to wave 10, where $t = 0$ refers to the Full-Sample at the start of the panel (wave 0). Further, each line has six points showing the percent frequencies in i^{th} state in wave t.

It is visible from Figure 26 that there is an under-representation for the lower categories (1 to 3) of the Resp-samples, while an over-representation for the higher categories (4 to 6) of the Resp-Samples. Moreover, for the lower categories (states), the case numbers in the Resp-Samples become extremely small. So it may happen that the lowest state 1 dies out due to non-response and later attrition. In such an event the ordered logit can't be estimated with the original number of thresholds. This affects the estimation of the other model parameters, namely also the slope

parameters of the variable $X_{i,t}$.

Besides, to the estimation of non-response/attrition biases of the estimators, we also estimate the MSE's of these estimators under Scenario A-D which are presented from Figure 27 to Figure 30, respectively. The vertical line of the figures shows the MSE of the estimator in percent, while the horizontal line shows the length of the panel. Looking at the graphical representations of the MSE's results of all the estimators under the Full-Sample is very small (close to zero), this is because there is no bias under the Full-Sample but a variance. However, estimators under the Resp-Sample have some small/large MSE's results depending on the size of permanent/transient components and the type of the estimator used. Most of the relationships are due to the bias component. Therefore, the square of the large biases of the weighted estimator leads to large MSE's results of the estimator. This also holds for all stability scenario (Scenario A-D). The results in this section are similar to the case of cross-sectional OLS/IPW estimators (see Section 3.5 of this chapter). Overall, from our analysis we conclude that weighting is only beneficial in reducing the bias and MSE of the estimates if the model for response is correctly specified, otherwise, it will make the results even more biased.

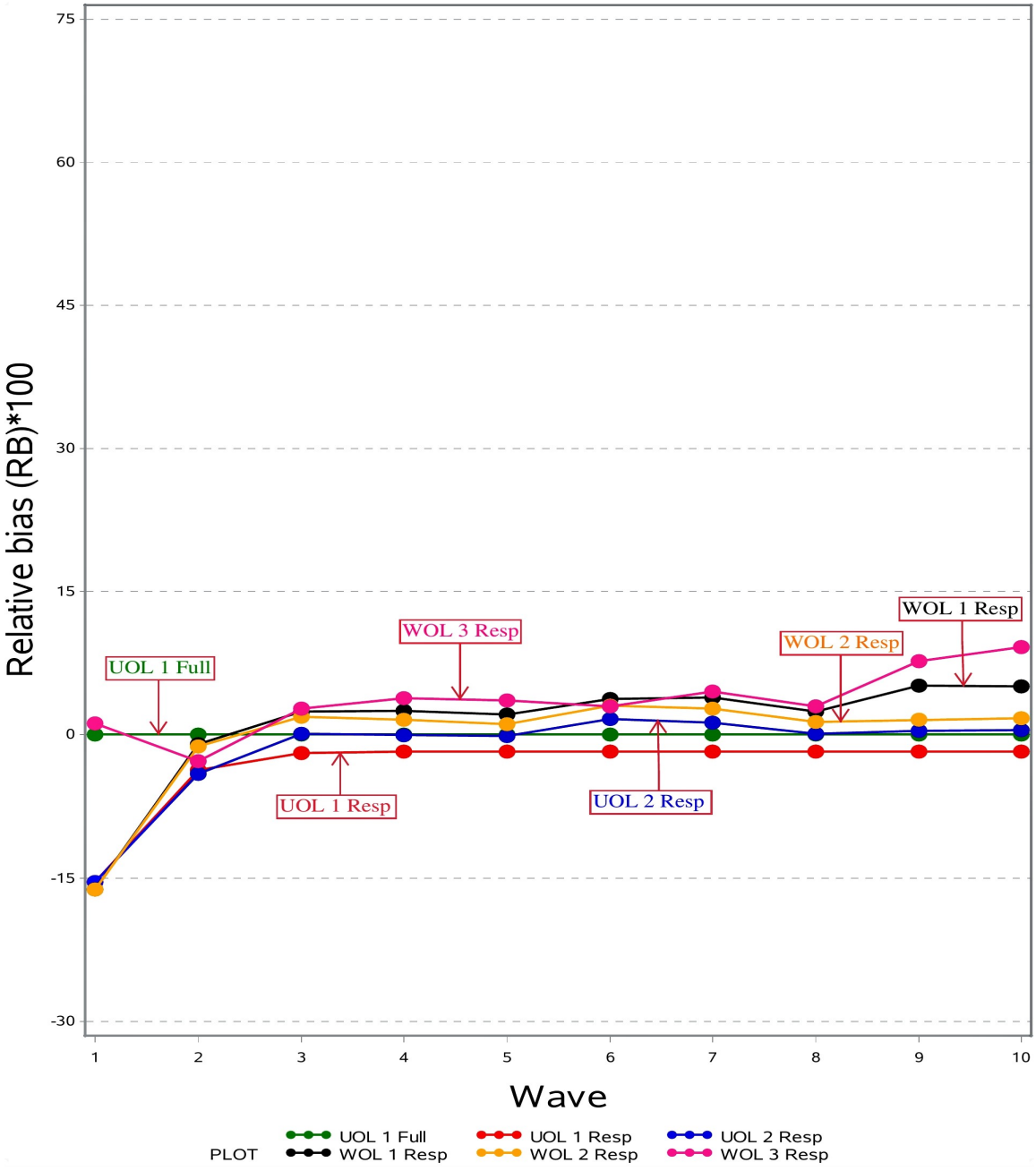


Figure 22: Fade-away effect of the initial non-response bias of the weighted and un-weighted estimates of an ordered logit model in Scenario A ($\kappa = \gamma = \rho = \phi = 0.10$) for fix non-response parameters $\alpha = -4.50, \beta = 2.00$ and attrition parameters $\alpha^{\star} = 0.01, \beta^{\star} = 0.70$. For more detail about the realistic and the unrealistic weighting approaches in the ordinal logit model see Table 5.

Note: The vertical axis represents the percent relative bias of the estimates, while the horizontal axis represents wave of the panel. Number of observations in the sample are $n = 1,000$, which are replicated $R = 100$ times.

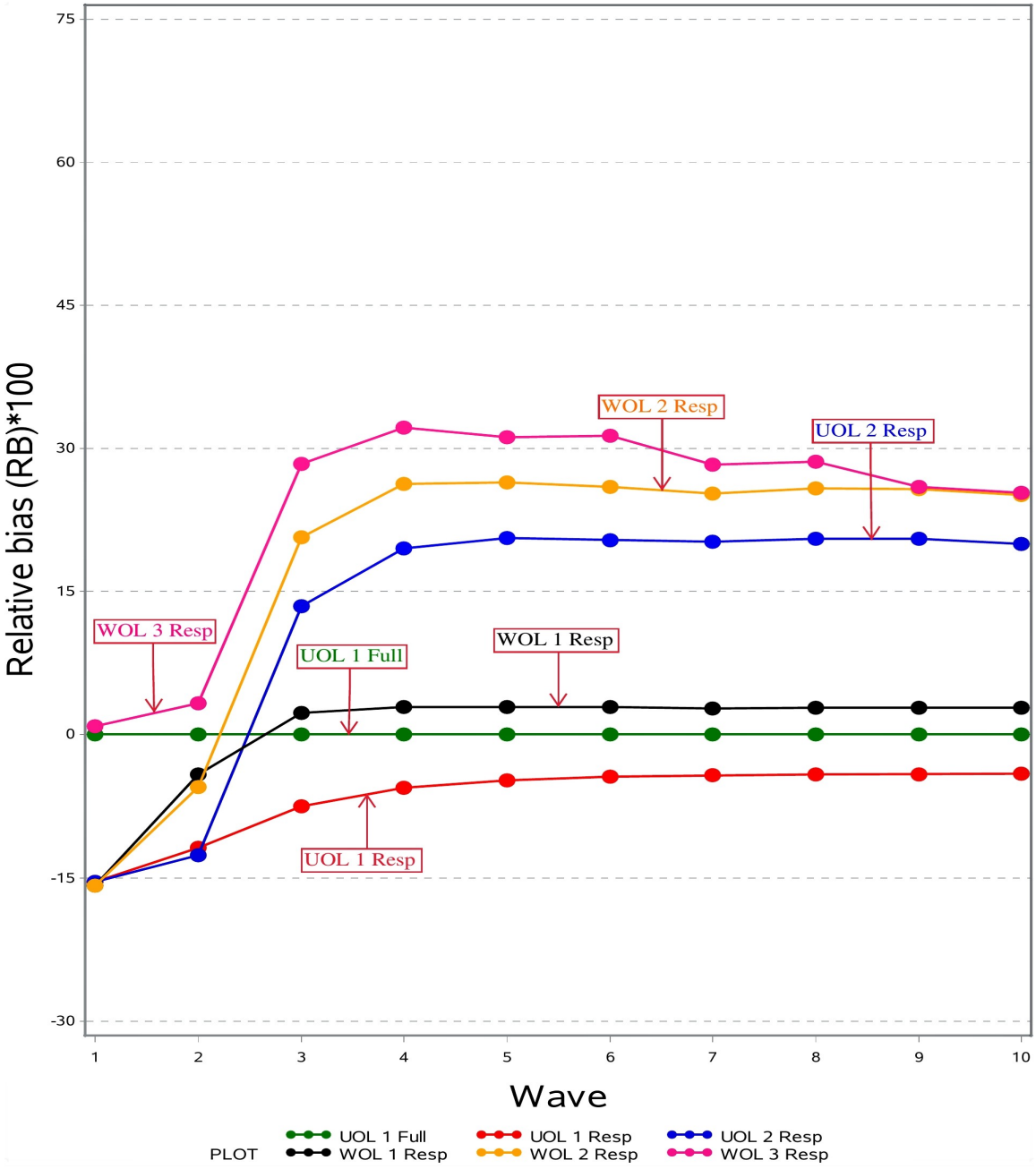


Figure 23: Fade-away effect of the initial non-response bias of the weighted and un-weighted estimates of an ordered logit model in Scenario B ($\kappa = \gamma = \rho = \phi = 0.50$) for fix non-response parameters $\alpha = -4.50, \beta = 2.00$ and attrition parameters $\alpha^{\star} = 0.01, \beta^{\star} = 0.70$. For more detail about the realistic and the unrealistic weighting approaches in the ordinal logit model see Table 5.

Note: The vertical axis represents the percent relative bias of the estimates, while the horizontal axis represents wave of the panel. Number of observations in the sample are $n = 1,000$, which are replicated $R = 100$ times.

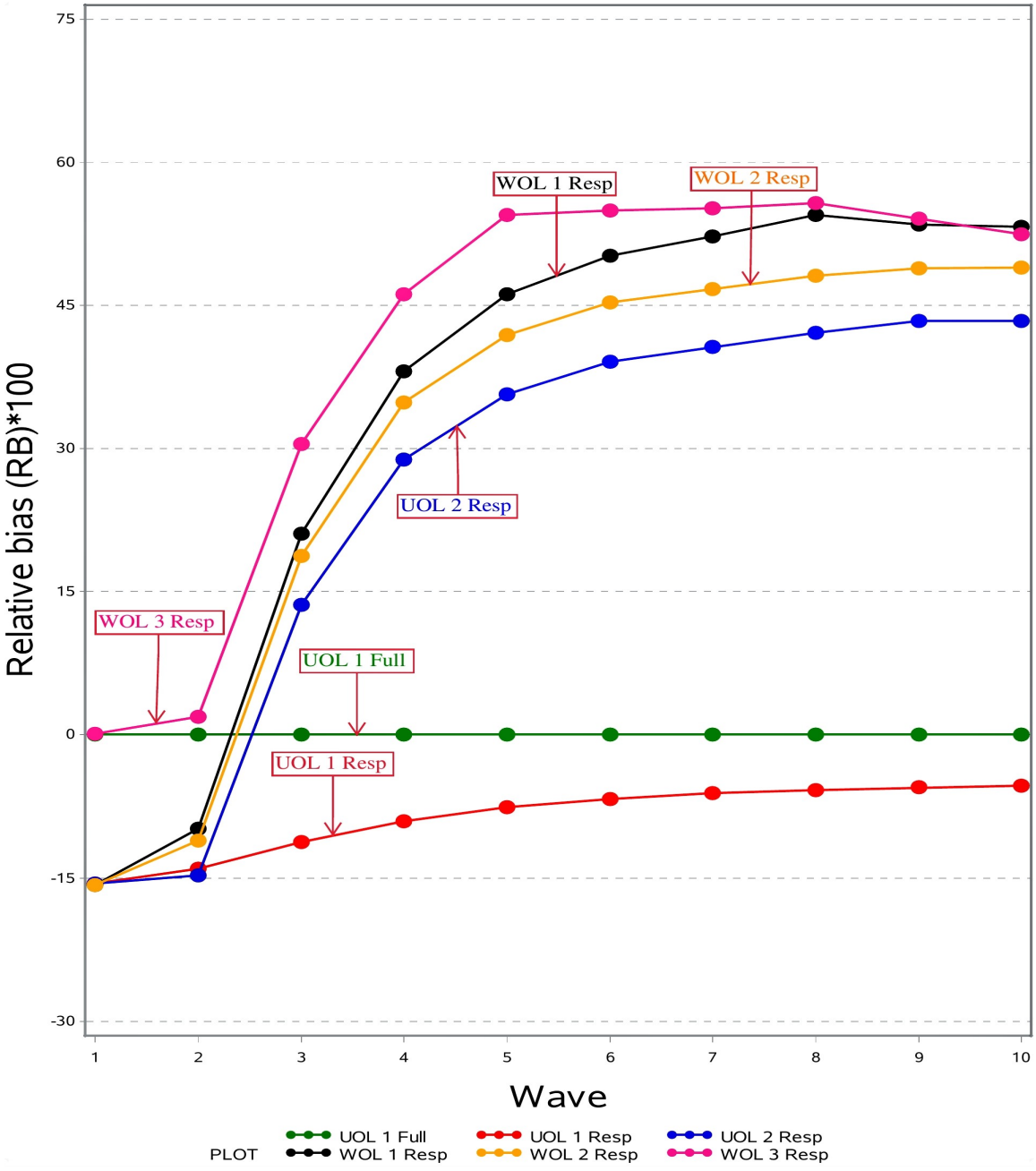


Figure 24: Fade-away effect of the initial non-response bias of the weighted and un-weighted estimates of an ordered logit model in Scenario C ($\kappa = \gamma = \rho = \phi = 0.70$) for fix non-response parameters $\alpha = -4.50, \beta = 2.00$ and attrition parameters $\alpha^{\star} = 0.01, \beta^{\star} = 0.70$. For more detail about the realistic and the unrealistic weighting approaches in the ordinal logit model see Table 5.

Note: The vertical axis represents the percent relative bias of the estimates, while the horizontal axis represents wave of the panel. Number of observations in the sample are $n = 1,000$, which are replicated $R = 100$ times.

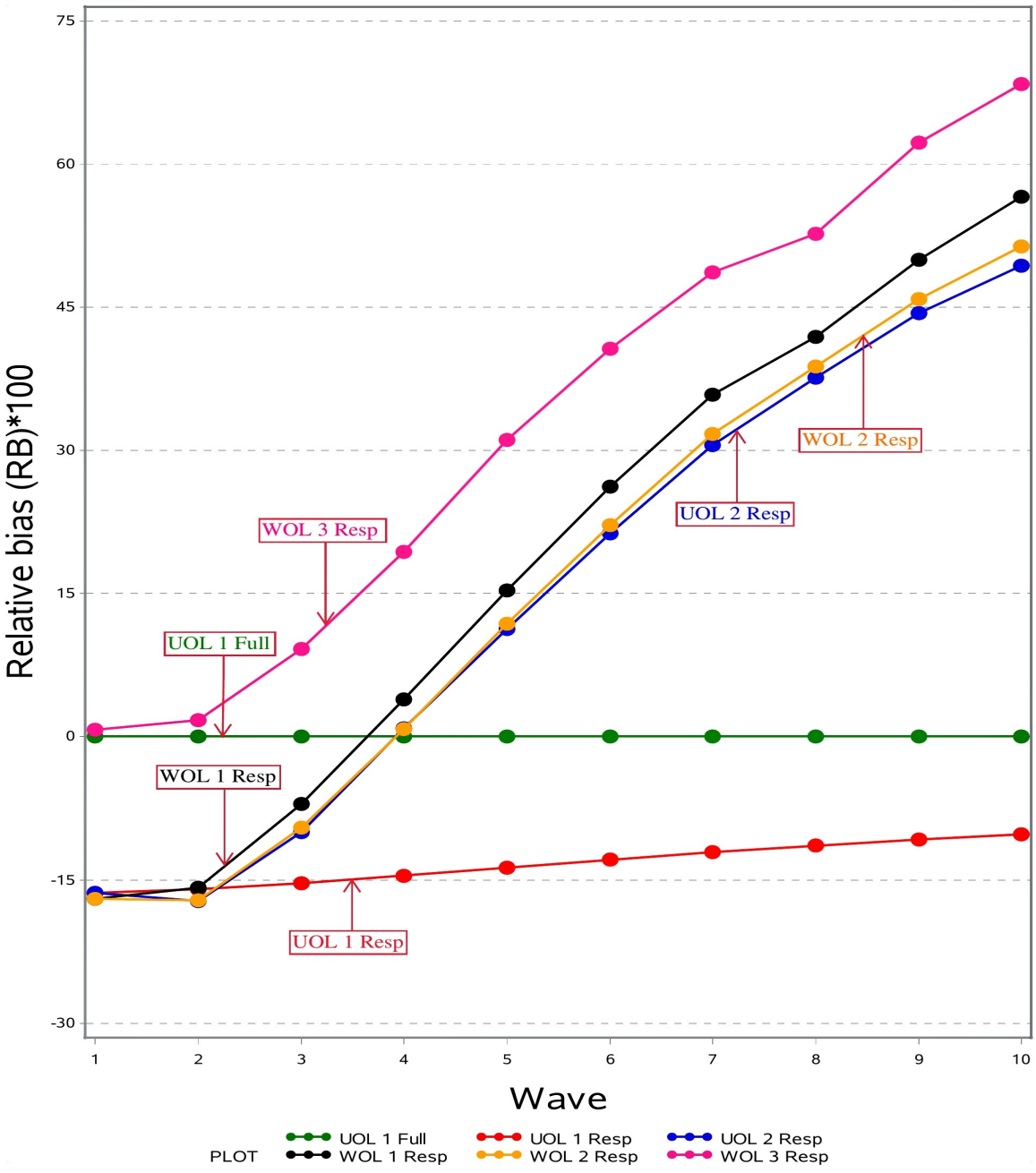


Figure 25: Fade-away effect of the initial non-response bias of the weighted and un-weighted estimates of an ordered logit model in Scenario D ($\kappa = \gamma = \rho = \phi = 0.90$) for fix non-response parameters $\alpha = -4.50, \beta = 2.00$ and attrition parameters $\alpha^{\star} = 0.01, \beta^{\star} = 0.70$. For more detail about the realistic and the unrealistic weighting approaches in the ordinal logit model see Table 5.

Note: The vertical axis represents the percent relative bias of the estimates, while the horizontal axis represents wave of the panel. Number of observations in the sample are $n = 1,000$, which are replicated $R = 100$ times.

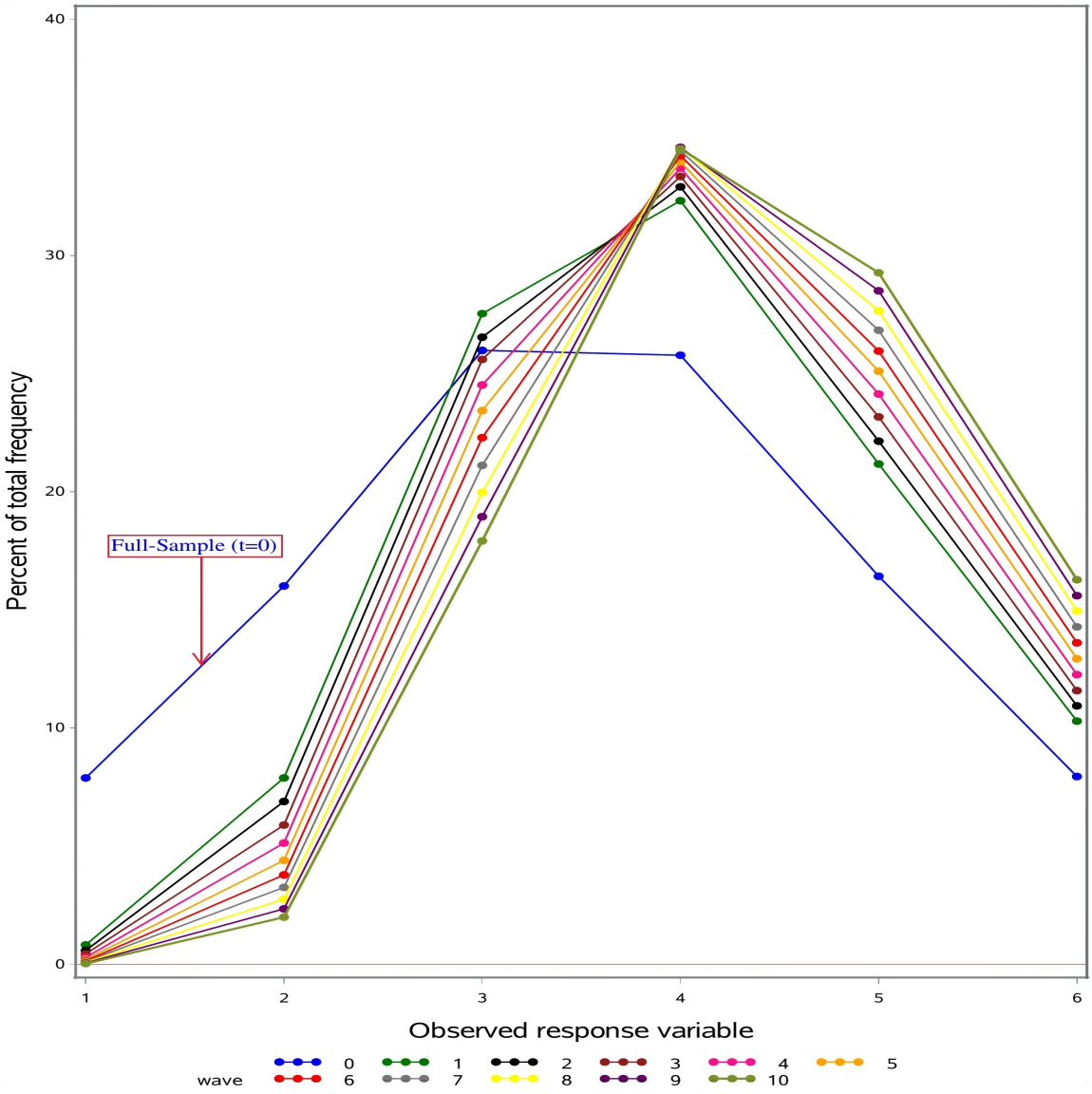


Figure 26: Distribution on states, for fix non-response parameters $\alpha = -4.50, \beta = 2.00$ and attrition parameters $\alpha^\star = 0.01, \beta^\star = 0.70$. Number of Monte Carlo replications $R = 100$ times. Initial non-response rate is 23%.

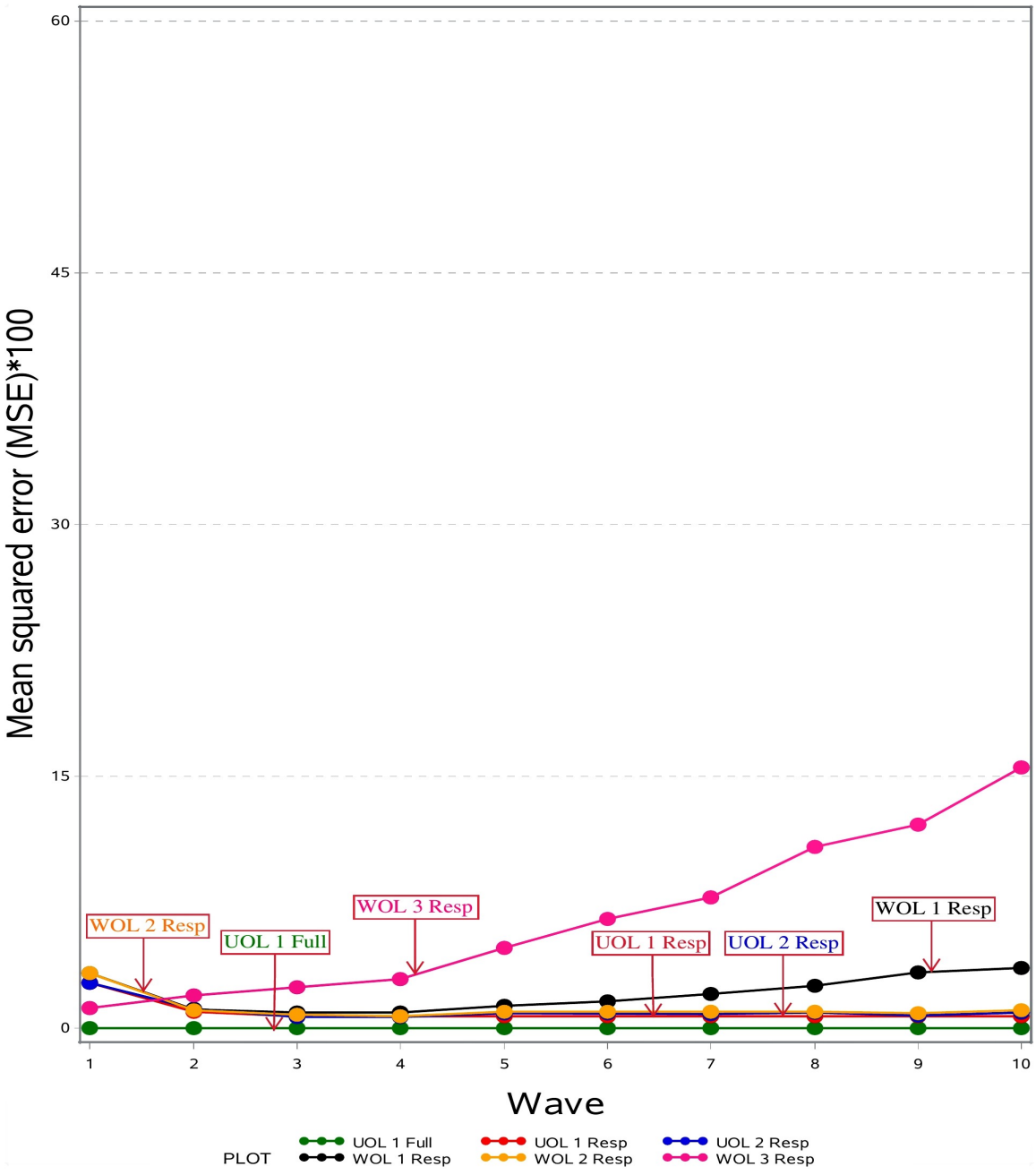


Figure 27: MSE comparison of the weighted and un-weighted estimates of an ordered logit model in Scenario A ($\kappa = \gamma = \rho = \phi = 0.10$) for fix non-response parameters $\alpha = -4.50, \beta = 2.00$ and attrition parameters $\alpha^{\star} = 0.01, \beta^{\star} = 0.70$. For more detail about the realistic and the unrealistic weighting approaches in the ordinal logit model see Table 5.

Note: The vertical axis represents the percent MSE of the estimates, while the horizontal axis represents wave of the panel. Number of observations in the sample are $n = 1{,}000$, which are replicated $R = 100$ times.

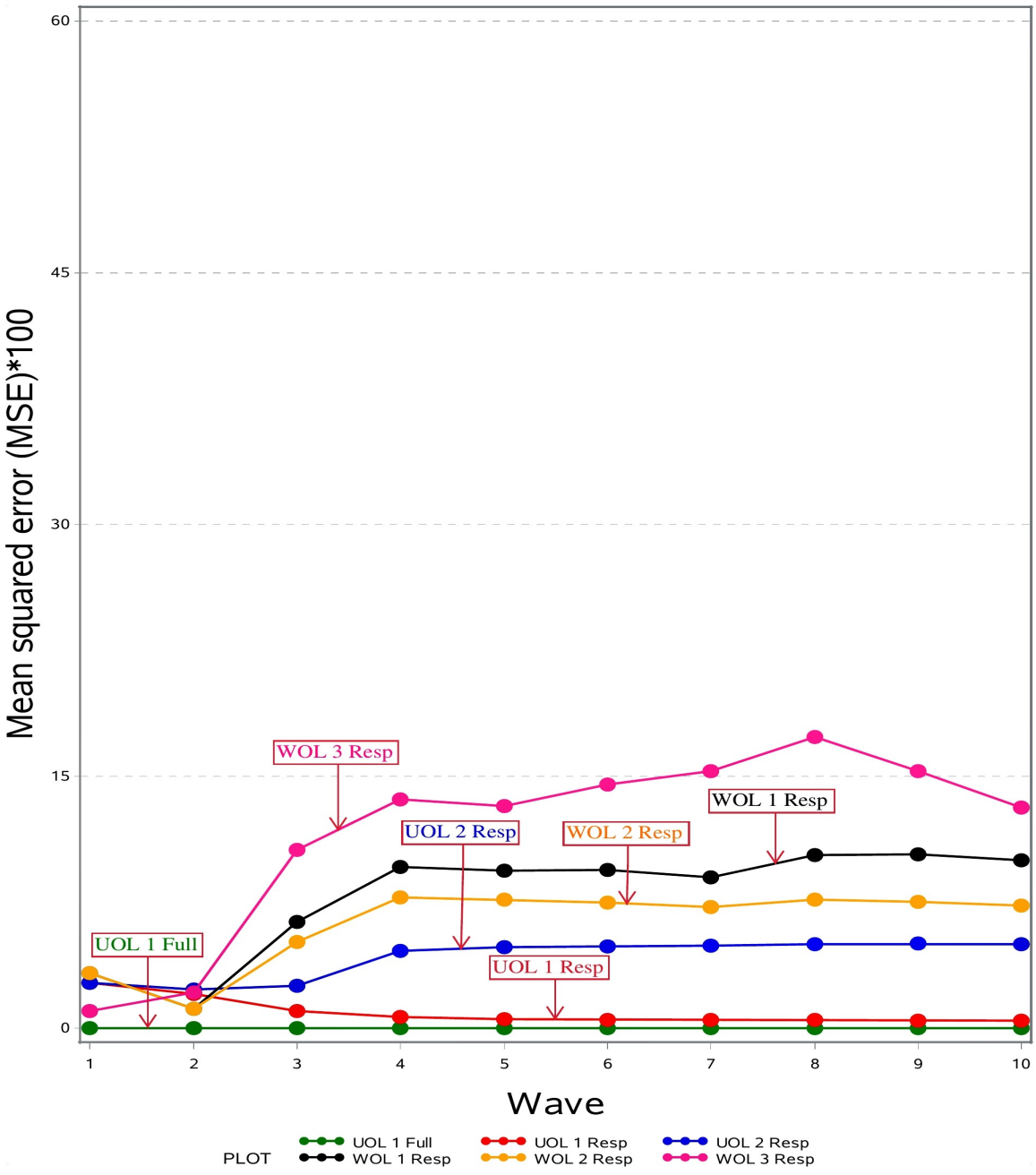


Figure 28: MSE comparison of the weighted and un-weighted estimates of an ordered logit model in Scenario B ($\kappa = \gamma = \rho = \phi = 0.50$) for fix non-response parameters $\alpha = -4.50, \beta = 2.00$ and attrition parameters $\alpha^{\star} = 0.01, \beta^{\star} = 0.70$.For more detail about the realistic and the unrealistic weighting approaches in the ordinal logit model see Table 5.

Note: The vertical axis represents the percent MSE of the estimates, while the horizontal axis represents wave of the panel. Number of observations in the sample are $n = 1{,}000$, which are replicated $R = 100$ times.

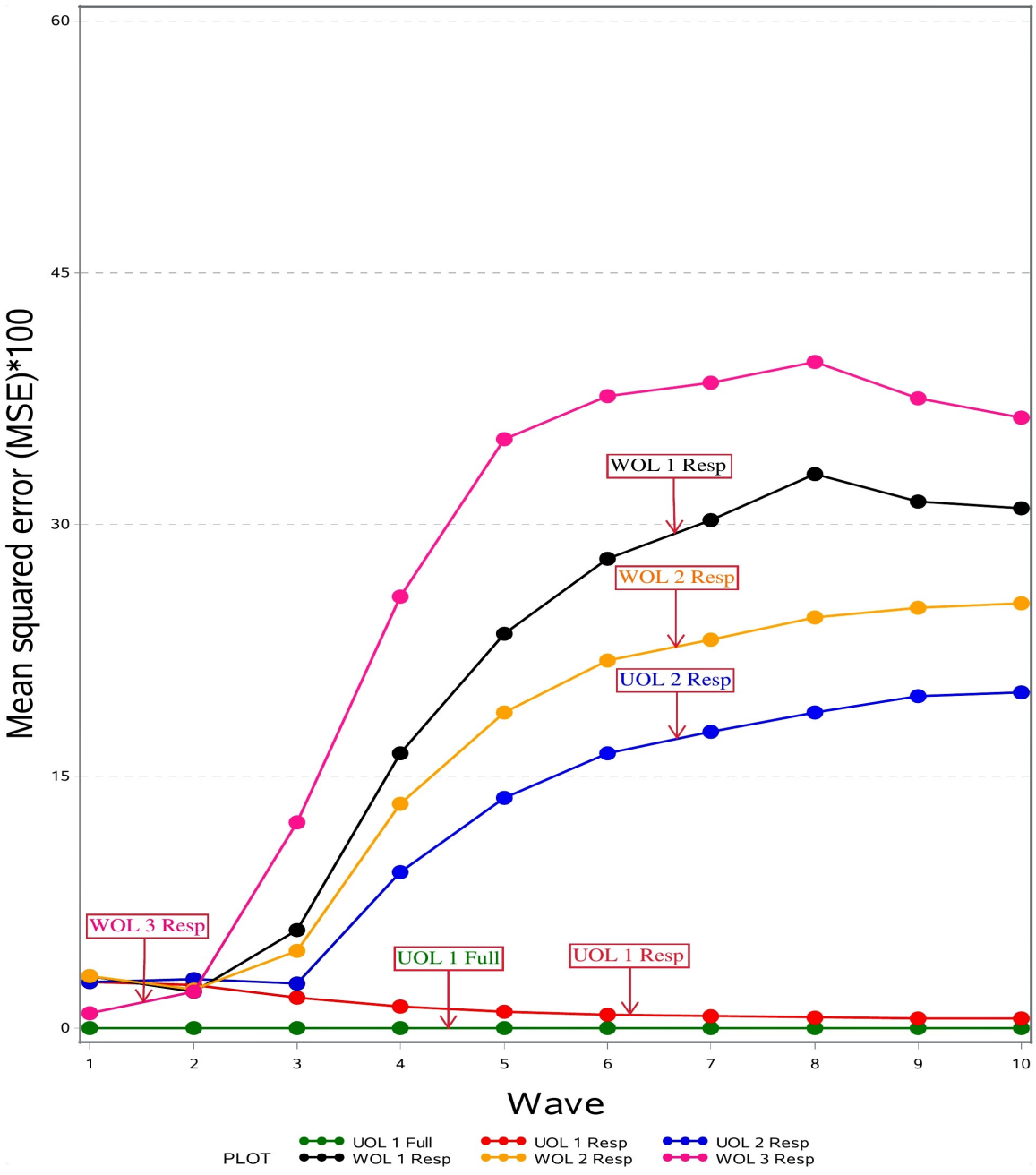


Figure 29: MSE comparison of the weighted and un-weighted estimates of an ordered logit model in Scenario C ($\kappa = \gamma = \rho = \phi = 0.70$) for fix non-response parameters $\alpha = -4.50, \beta = 2.00$ and attrition parameters $\alpha^{\star} = 0.01, \beta^{\star} = 0.70$. For more detail about the realistic and the unrealistic weighting approaches in the ordinal logit model see Table 5.

Note: The vertical axis represents the percent MSE of the estimates, while the horizontal axis represents wave of the panel. Number of observations in the sample are $n = 1{,}000$, which are replicated $R = 100$ times.

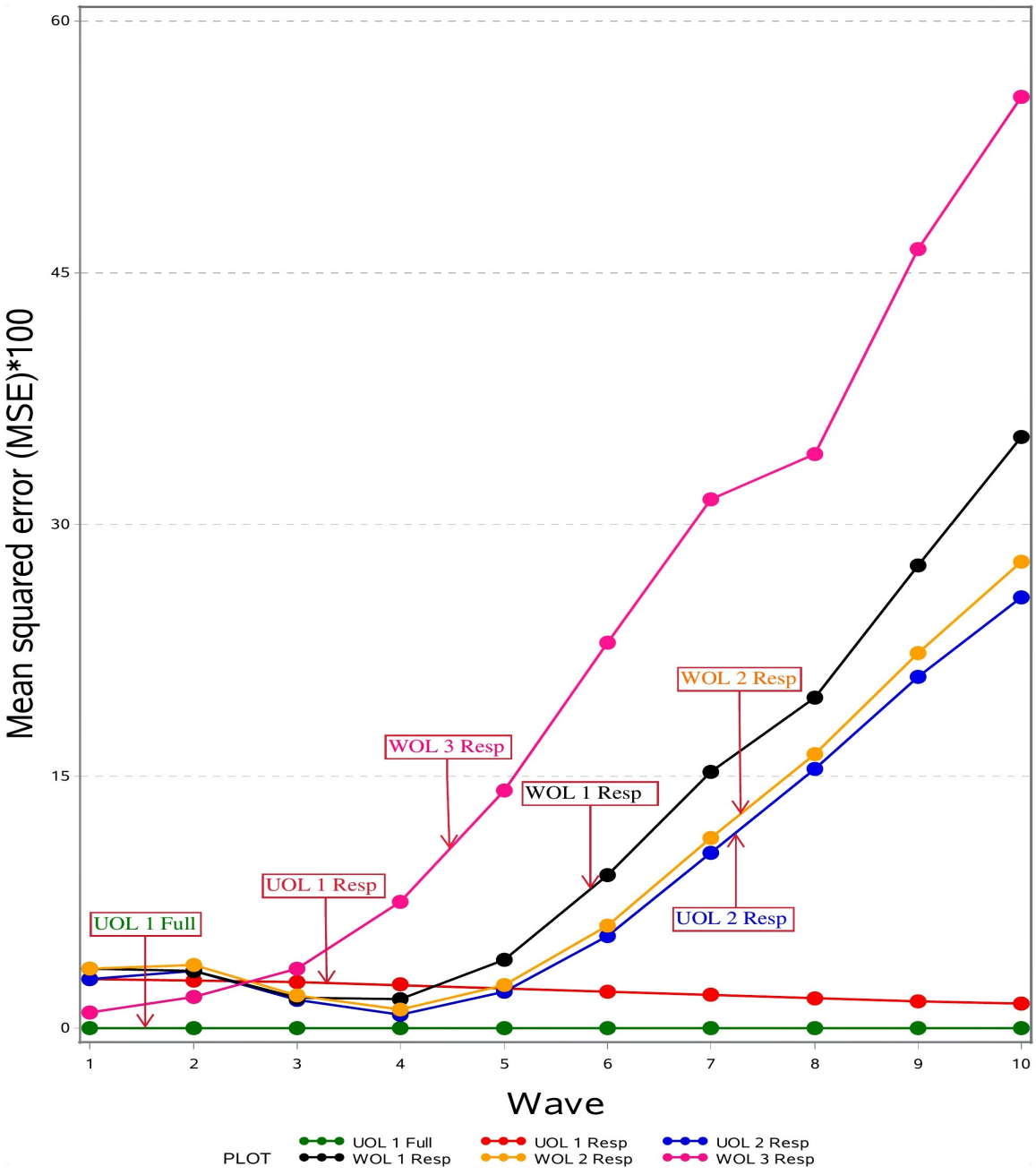


Figure 30: MSE comparison of the weighted and un-weighted estimates of an ordered logit model in Scenario D ($\kappa = \gamma = \rho = \phi = 0.90$) for fix non-response parameters $\alpha = -4.50, \beta = 2.00$ and attrition parameters $\alpha^{\star} = 0.01, \beta^{\star} = 0.70$. For more detail about the realistic and the unrealistic weighting approaches in the ordinal logit model see Table 5.

Note: The vertical axis represents the percent MSE of the estimates, while the horizontal axis represents wave of the panel. Number of observations in the sample are $n = 1{,}000$, which are replicated $R = 100$ times.

Part III

Application to SOEP Income and Life Satisfaction Data

CHAPTER 4

Empirical and Simulation Data

4.1. Motivation: Approach with SOEP

In Chapter 3 of this thesis, we have conducted a simulation study to verify the approximate results of Alho (2015), and investigated the accuracy of the bias approximation in a simulation setting. We checked the size of the fade-away in later panel waves with no analytical bias approximation. The speed of the fade-away effect of the initial non-response bias is then investigated for different stability scenarios of covariates and error terms, with and without any attrition patterns in later panel waves. As the speed of the fade-away depends on the stability of the covariates and the error terms it is important to investigate this effect not only for simulated data but also for real longitudinal data. Therefore, in the application part of this thesis, we switch to real data from the German Socio Economic Panel (SOEP[1]): specifically to income data and life satisfaction scores data of the SOEP. Here we used the following two settings:

- Income data and their explanation via regression.
- Life satisfaction scores data and their explanation by an ordered logit model.

[1]Socio Economic Panel (SOEP), data for years 1984-2016, version 33, SOEP, 2017, doi: 10.5684/soep.v33.

While Alho's approximation was developed for cross-sectional OLS estimates, in a panel more complex estimators are used, for example, estimators that use information from several panel waves and incorporate dependent error terms. For such models and estimators, analytical bias approximation is out of scope. Therefore, simulation will be also used here to study the fade-away effect and its size.

In the first part of this chapter, we will investigate the fade-away effect of initial non-response on the estimation of a wage equation using data from the first 10 panel waves of the SOEP. To examine the fade-away effect, we use three different model settings: a RE model for wages with and without lagged dependent variable, a RE model with auto-correlated errors for wages and a FE wage model. Here we will not use any simulation of the dependent variable and the covariates. The only part which is simulated is the endogenous drop-out of observations at the start of the panel. Furthermore, we will switch to a model with multiple covariates. In order to demonstrate the fade-away effect, we will gradually extend the database from 1 to 10 panel waves. This covers the length 1984 to 1993 of the Sub-sample A-B of the SOEP. Contrary to the previous work in Chapter 3, where the dependent variable and the covariate are simulated, here we use real data from the SOEP. The estimations are done with **SAS** and with the procedure: **PROC REG**, **PROC PANEL**, and **PROC HPMIXED**, respectively.

The second part of this chapter is to explore the effect of initial-response on the estimation of a model which explains life satisfaction scores by using SOEP data from year 2000 to 2010 of the Sub-sample F. To examine the fade-away effect, we use two models: the ordered logit model for cross-sectional data and the RE model for longitudinal data. In order to demonstrate the fade-away effect we gradually extend the database from 1 to 11 panel waves. This covers 11 panel waves of the SOEP starting from the year 2000 to 2010. The estimations are done with SAS and with the procedure: **PROC HPLOGISTIC**, and **PROC GLIMMIX**, respectively.

4.2. Application to SOEP income data

4.2.1. The models

If we observe the wages of labor, it is not only linked with education, but there are many other variables that affect wages. The reasons for variation in wages may be

due to work experience, occupation, gender, education, region, and demographics, etc. There exists an extensive literature on the analysis of wage regression. The early work of the Mincer (1974) on wage explanation is the most widely used tool in empirical work. Mincer modeled the log of earnings as a function of years of education, years of experience and a quadratic function of experience. So in order to evaluate the relationship among the economic variables, we use the following wage equation.

$$W_{i,t} = b_0 + b_1 Age_{i,t} + b_2 Edu_{i,t} + b_3 G_i + b_4 D_{i,t} + b_5 F_{3i,t} + b_6 F_{4i,t} + b_7 F_{5i,t} + b_8 Tenure_{i,t} + e_{i,t}, \tag{4.1}$$

where $W_{i,t}$ represents the log of the hourly wage of individual i at time t, $Age_{i,t}$ and $Edu_{i,t}$ are the variables representing individual age and years of education. Here it is important to mention that we exclude age-squared, years of experience and a quadratic function of experience from the model due to multicollinearity. G_i is a binary variable for gender. $D_{i,t}$ is a dummy for single, which is one if the individual is single and zero otherwise. Similarly, $F_{3i,t}$, $F_{4i,t}$ and $F_{5i,t}$ are the firm size dummies for the categories: (3) 20 to 199 employees (4) 200 to 1,999 employees and (5) at least 2,000 employees respectively. The reference categories are small firms with up to 19 employees. $Tenure_{i,t}$ is the length of individual job tenure with a firm. Finally, $e_{i,t}$ is the time-varying error term of the model that determines the unknown factors that affect $W_{i,t}$, where t refers to the selected time of interview $t = 1, 2, 3, \ldots, 10$.

In the context of the longitudinal structure of the data set, the estimation of the model (4.1) also includes FE with respect to time and individual RE with respect to individuals. The inclusion of fixed time effects accounts for the yearly changes that are the same for all the individual's life inflation. The individual RE account for the unobserved heterogeneity that is constant over time but is different for each individual. So in order to control for unobserved heterogeneity in the panel, we use the following panel models. By rewriting the error structure of Equation (4.1) as:

$$e_{i,t} = v_i + \eta_{i,t}, \tag{4.2}$$

where v_i is the time-invariant intercept of individual i which may or may not be correlated with the other observed covariates of the model, and $\eta_{i,t}$ is the idiosyncratic

error term of individual i at time t. Then the relationship between the log hourly wage $W_{i,t}$ of the individual i at time t and the set of observed covariates together with an error components model $e_{i,t} = v_i + \eta_{i,t}$, can be modeled by the following linear regression:

$$W_{i,t} = b_0 + b_1 Age_{i,t} + b_2 Edu_{i,t} + b_3 G_i + b_4 D_{i,t} + b_5 F_{3i,t} + b_6 F_{4i,t} + b_7 F_{5i,t} + b_8 Tenure_{i,t} + v_i + \eta_{i,t}, \tag{4.3}$$

Then depending on the nature of v_i, and $\eta_{i,t}$ in the panel, we distinguished the following panel models:

- Random effects (RE) model: It assumes that v_i is the time-invariant random intercept associated with each person i, which is assumed to be normally distributed $v_i \overset{iid}{\sim} N(0, \sigma_v^2)$, and is orthogonal to the observed covariates $X_{it,k}$ of the model and with the idiosyncratic error term $\eta_{i,t}$. Where k denotes the number of covariates. Further, the idiosyncratic error term $\eta_{i,t}$ is assumed to be normally distributed $\eta_{i,t} \overset{iid}{\sim} N(0, \sigma_\eta^2)$ and orthogonal to the observable explanatory variables $X_{it,k}$ in the model. Mathematically, the model in equation (4.3) is said to be RE model if the following orthogonality condition is satisfied:

  $$Cov(v_i, X_{it,k}) = 0, \quad \text{for all} \quad k, i \quad \text{and} \quad t. \tag{4.4}$$

  As long as the regressors are uncorrelated with the individual effects v_i and the error term $\eta_{i,t}$, we can get unbiased, consistent, and efficient parameter estimates by using generalized least squares (GLS) for fixed values of σ_v^2 and σ_η^2. However, in practice the values of σ_v^2 and σ_η^2 are unknown so that GLS is not feasible, then in such a case σ_v^2 and σ_η^2 can be estimated by using the method called feasible generalized least squares (FGLS).

- Fixed effects (FE) model: The RE model is based on the orthogonality assumption, consistency involves that the unobserved effects v_i should be uncorrelated with the observed covariates $X_{it,k}$ included in the model. However, if v_i is

correlated with the observed model covairtaes $X_{it,k}$, such that:

$$Cov(v_i, X_{it,k}) \neq 0, \quad \text{for all} \quad k, i \quad \text{and} \quad t. \tag{4.5}$$

Then regression parameters can be more efficiently estimated by using the FE model. For more detail about RE and RE models see Subsection 2.3.3 of Chapter 2.

- Panel model with auto-correlated errors:

$$W_{i,t} = b_0 + b_1 Age_{i,t} + b_2 Edu_{i,t} + b_3 G_i + b_4 D_{i,t} + b_5 F_{3i,t} + b_6 F_{4i,t} + b_7 F_{5i,t} + b_8 Tenure_{i,t} + \eta_{i,t}, \tag{4.6}$$

The within individual errors $\eta_{i,t}$ are assumed to follow a first order auto-regressive AR(1) process given by:

$$\eta_{i,t} = \phi \eta_{i,(t-1)} + \varepsilon_{i,t} \tag{4.7}$$

where, $\varepsilon_{i,t}$ is the fresh error term which is assumed to be $\varepsilon_{i,t} \overset{iid}{\sim} N(0, \sigma_\varepsilon^2)$ and ϕ is the auto-regressive coefficient.

- RE model with lagged dependent variable $W_{i,(t-1)}$: To examine the fade-away effect in a RE model with lagged dependent variable $W_{i,(t-1)}$, we replace equation (4.6) in the sense that it contains one period lagged values of the dependent variable. Dynamic panel models are useful when the outcome variable depends on its own past periods. For example, one motivational example in this context is the present income of a person which heavily depends on its previous income. Therefore log income is a good predictor for the present income and should be included in the model to keep non-response bias small. Therefore, by including the lagged dependent variable as a predictor variable in equation (4.6), the economic model becomes

$$W_{i,t} = \delta W_{i,(t-1)} + b_0 + b_1 Age_{i,t} + b_2 Edu_{i,t} + b_3 G_i + b_4 D_{i,t} + b_5 F_{3i,t} + b_6 F_{4i,t} + b_7 F_{5i,t} + b_8 Tenure_{i,t} + v_i + \eta_{i,t}, \tag{4.8}$$

where, $W_{i,(t-1)}$ is the lagged dependent variable (LDV) at time $t-1$ with an unknown slope coefficient δ.

4.2.2. Data and descriptive statistics

The SOEP is a longitudinal survey that began in 1984 with initially 12,000 participants. This number increased periodically by refreshment samples and now more than 20,000 individuals are interviewed each year from more than 10,000 households, including Germans and immigrants[2]. The SOEP provides rich information on a wide range on various socio-economic and demographic characteristics, household composition, labor market participation, education, health, job characteristics, and satisfaction levels, etc. Here we use individual-level data from the first 10 waves of the SOEP for the year 1984 to 1993. The sample consists of all full time and part-time workers who are in the labor force (both men and women between age 18 to 65). Further, we restrict our sample to only Germans and foreigners in the Sub-sample A and B (Sub-sample A consists of "residents in the Federal Republic of Germany (FRG)" and Sub-sample B consists of "foreigners households in the Federal Republic of Germany (FRG)").

The dependent variable is the log of the hourly wage. Hourly wages are calculated by dividing the gross annual labor income by the annual work hours of individuals. This study excluded missing, imputed, zero or less than zero wages from the analysis. The final sample after these reduction steps results in an unbalanced panel with almost 35,082 observations from 4,467 persons of the 10 panel waves. Based on the above-described data, Table 6 provides summary statistics of the main variables used in the analysis.

[2]The SOEP contains various sub-samples (Sub-sample A-M), which are sampled with different rates. Initially, the SOEP survey was started in West Germany in 1984 with two Sub-samples A and B. Sub-sample A consists of German residents of West Germany, which covered 4,528 households in 1984, while Sub-sample B is a sample of foreigners households in West Germany consisting of Turkish, Yugoslavian, Greek, Spanish or Italian household heads. This Sub-sample B is also started in 1984 with 1,393 households. For more detailed information about the SOEP samples see (Wagner et al., 2007) and Grabka (2012).

Table 6: Summary statistics for the estimation sample using data from 1984 to 1993 of individuals aged between 18-65.

Variable	Mean	Median	SD	Min	Max
Hourly wage (in Euro)	10.96	9.21	13.70	0.11	628.64
Age	41.00	41.00	10.73	18.00	65.00
Years of education (Edu)	11.06	10.50	2.53	7.00	18.00
Male (G)	0.61	1.00	0.49	0.00	1.00
Single (D)	0.16	0.00	0.37	0.00	1.00
Firm size 20-199 (F_3)	0.25	0.00	0.43	0.00	1.00
Firm size 200-1999 (F_4)	0.22	0.42	0.41	0.00	1.00
Firm size 2,000+ (F_5)	0.26	0.00	0.44	0.00	1.00
Tenure	11.59	9.60	9.43	0.00	49.90
Observations			35,082		

4.2.3. The design of the simulation study

In order to investigate the fade-away effect for the distributional differences between the distributions of the Full and the Resp samples, we use a simulation approach. We restrict our analysis to all those individuals who continuously participate in the first 10 panel waves of the SOEP for the year 1984 to 1993. We further exclude all those individuals who leave or temporary dropout in any panel wave. There are about 3,575 persons belonging to the "Full-Sample".

We then artificially introduce initial non-response in wave 1 under the assumption that non-response at the start of the panel is endogenous:

$$P(R_{i,1} = 1|W_{i,1}) = \frac{\exp(\alpha + \beta W_{i,1})}{[1 + \exp(\alpha + \beta W_{i,1})]}, \tag{4.9}$$

where $W_{i,1}$ is the log hourly wages of the individuals in wave 1, and α and β are the initial non-response parameters whose values are to be chosen such that the resulting response probabilities from equation (4.9) lie in the interval [0, 1]. For $\alpha = -2.0$ and $\beta = 1.50$ we generate a non-response rate of about 30% in the initial wave. In order to get more stable results, we replicated the initial non-response 100 times. Thus, we get 100 "Resp-Samples" of size 2,505 persons for each panel wave. However, the Resp-Samples are not further reduced by panel attrition after wave 1. Note, that

we did not simulate the dependent variable and the covariates of the model as was done in the simulation of Alho's model. An essential condition to demonstrate the fade-away effect is to choose that non-response at the start of the Resp-Samples should be highly selective for response. So for $\alpha = -2.0$ and $\beta = 1.50$ we generate substantial initial non-response in the first wave of Resp-Samples. We demonstrate this in Figure 31. It can be seen that persons with high incomes have a trend to respond with a higher probability than persons with low incomes.

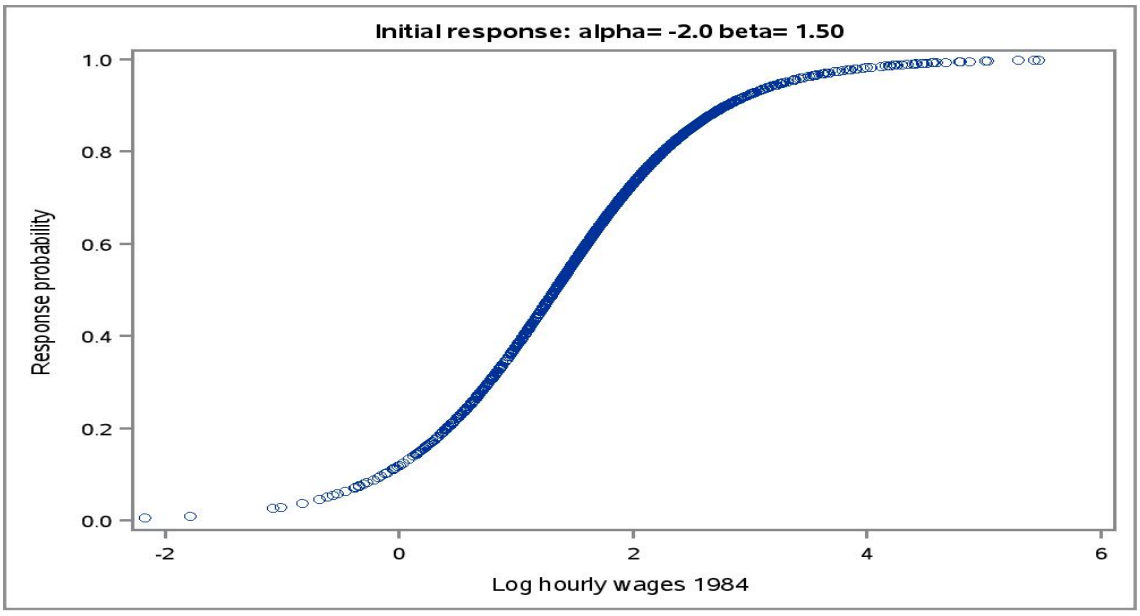


Figure 31: Impact of wages on the response probabilities.
The vertical axis shows the response probability and the horizontal axis shows hourly wages in 1984. The average response rate over $R = 100\%$ Monte Carlo replication is 70%.

For the assessment of the non-response bias, we compared the estimates of the Full and the Resp samples in waves 1 to 10. As we know that the bias of the estimator $\hat{b}$ is the difference between the expected value of the estimator and its true parameter which is being estimated, say b. Mathematically, it can be written as $\text{bias}(\hat{b}) = E(\hat{b}) - b$. As in our simulation approach, the true parameter b is unknown the estimation bias of $\hat{b}$ is estimated by the difference of the Full-Sample and the Resp-Samples estimates. Let $\hat{b}_{Full}$ and $\hat{b}_{Resp}$ be the regression coefficients of the Full and the Resp samples, respectively. Under this respect the bias in wave t is $\text{bias}(\hat{b}_{p,t}) = \hat{b}_{p,t}^{Full} - \hat{b}_{p,t}^{Resp}$, where $t = 1, 2, 3, \ldots, 10$ and the subscript p describes the variable intercept, age, years of education, marital status (single), firm size 20-199,

firm size 200-1999, firm size 2000+, gender (male) and job tenure, respectively.

4.2.4. Discussion of results

4.2.4.1. The cross-sectional results

We first discuss the regression results for the cross-sectional estimates. For the cross-sectional data, the empirical results of the Full-Sample and the results of the simulated Resp-Samples are summarized in Table 42 to Table 43 in Section B.1 of Appendix B. For these estimates, we compute the biases in Table 44 in Section B.1 of Appendix B. To give a better understanding of the fade-away of bias, we plot these results in Figure 32. Figure 32 emphasizes the effect of initial bias fades-away over the life of the panel. However, the speed factor of the fade-away process varies considerably between different slope parameters of the covariates. For example, an initial bias of the estimate of the intercepts (blue) is -0.37 which reduces very fast in the subsequent waves and is about -0.14 in wave 10. Similarly, the effect of singles (brown) reduces from -0.12 to -0.06 in wave 10. In the case of firm sizes (deep pink, orange and black at the top of the Figure) the initial biases fade-away in a geometrical pattern in the subsequent panel waves e.g., the size of initial biases in wave 1 which varies between 0.11 to 0.14 faded-away to 0.05 to 0.07 in wave 10. The effect of gender (deep yellow) is almost stable over the entire panel waves. There is apparently no bias for the cross-sectional estimation of the effect of education (purple), age (red) and tenure (green).

We also investigate the fade-away effect for the distributional differences of the Full and the Resp samples estimates through box-plot diagrams which reflects the variance of the Resp-Samples results over replications of the initial non-response. We placed these diagrams in Figure 33 to Figure 35. The vertical axis displays the magnitude of non-response bias($\hat{b}_{p,t}$) of the OLS estimates, while the horizontal axis displays the panel waves 1 to 10. The filled circles of the plots show the median. The lower and upper ends of the boxes are the lower and upper quartiles, and the vertical lines are used to indicate the spread and shape of the tails of the distribution. The little white circles outside the boxes indicate outliers in the data. The plots also display a line indicating a zero bias. Interestingly, the zero line is never meet and also the boxes don't cover the zero line. What can be seen, however, is that the centers

of the boxes move towards the zero bias line, but will not reach it. This advocate for a stable error component of wages which is related to initial non-response. This is in line with the theoretical results of Alho (2015). However, the persistent biases of the estimate of age, years of education and tenure are very small in the absolute numbers of Figure 33. Figure 33 also displays that after 6 waves there is no further fade-away of the initial bias.

We also consider the estimation of a dynamic model for cross-sectional data with lagged log hourly wage $W_{i,t-1}$ as an explanatory variable. We estimate regression parameters under the Full and for the Resp samples. The bias of the estimates is then obtained by the difference between the Full and the Resp samples estimates. We present these results in Table 45 to Table 47 in Section B.1 of Appendix B, which are graphically displayed in Figure 36. Figure 36 reveals a surprising result. The slope coefficient of the lagged hourly wage is estimated with a large positive bias while on the opposite side the intercept is reversely overestimated, resulting in a negative bias. There is an apparent trade-off between the two biases. Also, these biases substantially diminished until wave 10.

What the reason for this behaviour ? Figure 37 displays a scatter plot of W_2 (log hourly wage in 1985) and W_1 (log hourly wage in 1984). Besides the scatter plot one finds the regression line (red) and non-parametric LOESS estimation (blue) for the conditional expectation of W_2 for given values of W_1. The LOESS regression line strongly suggests that for wages larger than $W_1 = 3$ there is a different slope coefficient for the impact of W_1 on W_2 compared to lower values of $W_1 = 3$. Thus a model with only one slope coefficient, which is the case here, is misspecified. As the non-response drops-off persons with low wages with a higher probability, we expect a lower estimate for the slope coefficient in the Resp-Samples. This is displayed in Figure 38 which shows the scatter plot of W_1 on W_2 in the Resp-Samples together with the estimated regression line and the LOESS curve. Here the regression line lies below the LOESS line, while in the Full-Sample the regression line lies above the LOESS line. Consequently, the bias of the regression analysis of the slope coefficient of W_1 is positive. On the other hand, as persons with low incomes have a higher probability for non-response the general level of the income becomes larger in the Resp-Samples, resulting in a negative bias of the intercept.

Thus the bias in the estimation of the slope coefficient of W_1 results from the misspecification of the regression model. It is interesting to see that such biases also

fade-away quite fast.

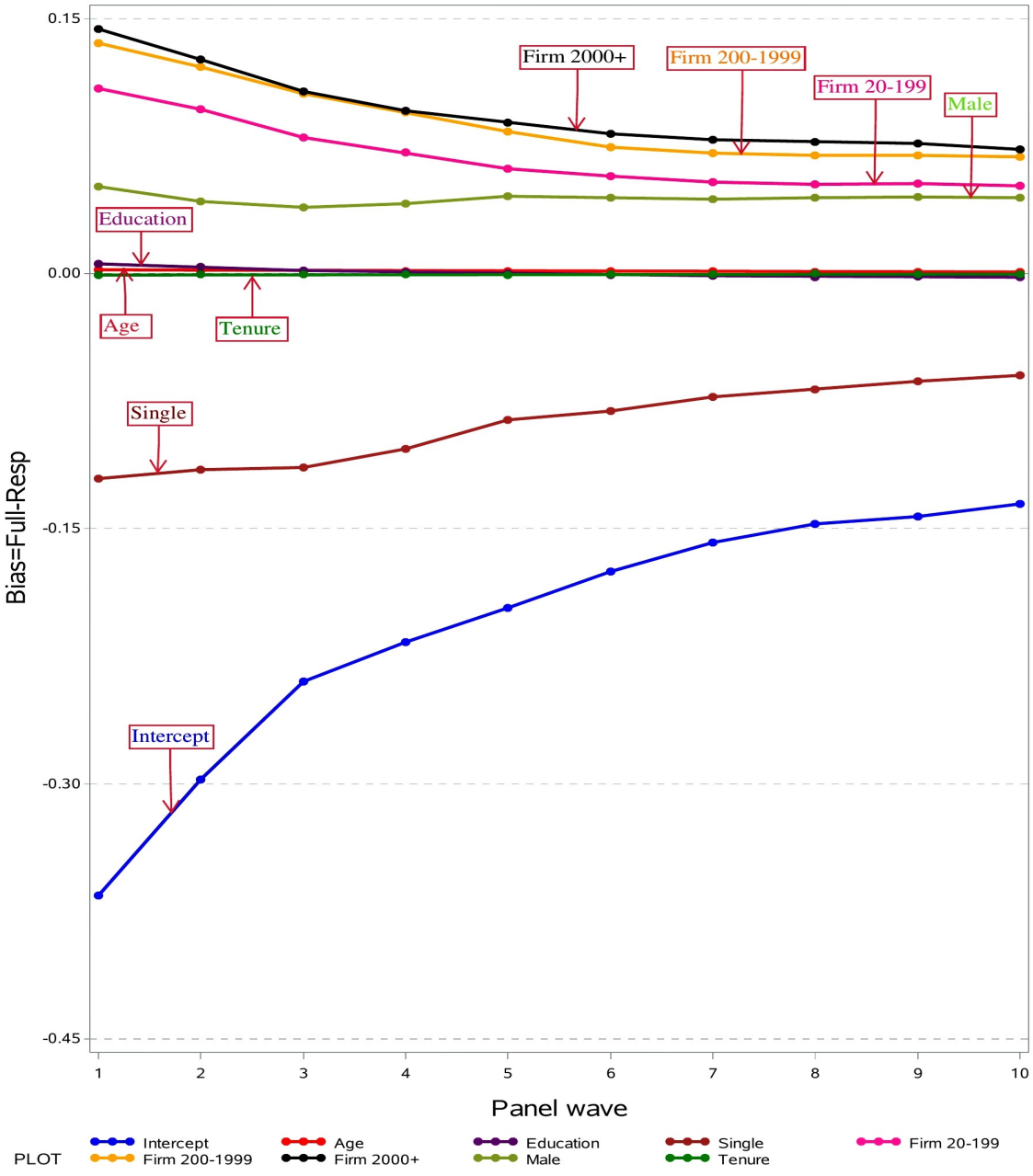
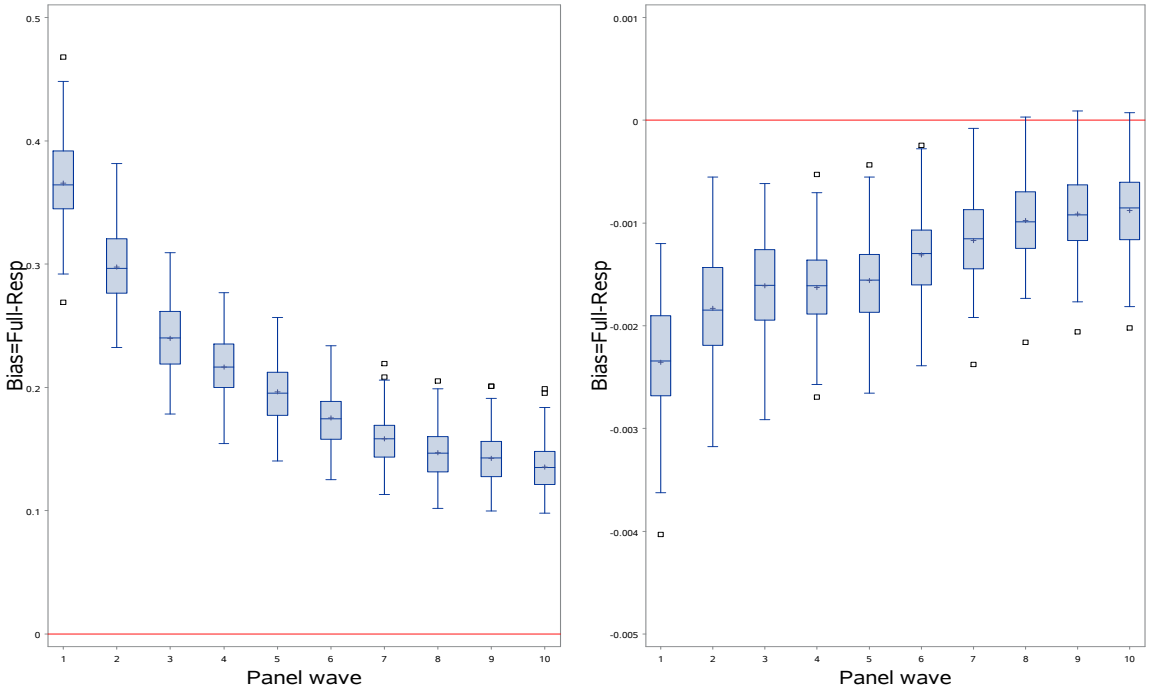


Figure 32: Graphical display of the fade-away of bias of the OLS estimator, with SOEP data and artificial initial non-response.

Note: The vertical axis displays the bias, while the horizontal axis displays panel waves. Number of Monte Carlo replications is $R = 100$. Initial non-response rate is 30%. The bias in each panel wave is obtained by the difference of the regression coefficients in the two samples i.e., $\text{bias}(\hat{b}_{p,t}) = \hat{b}_{p,t}^{Full} - \hat{b}_{p,t}^{Resp}$, where p denotes the estimated coefficients of the intercept, age, education, single, firm size, gender, and tenure, respectively. The points on the graph as highlighted in different colors represent the biases in the OLS estimates in certain panel waves.

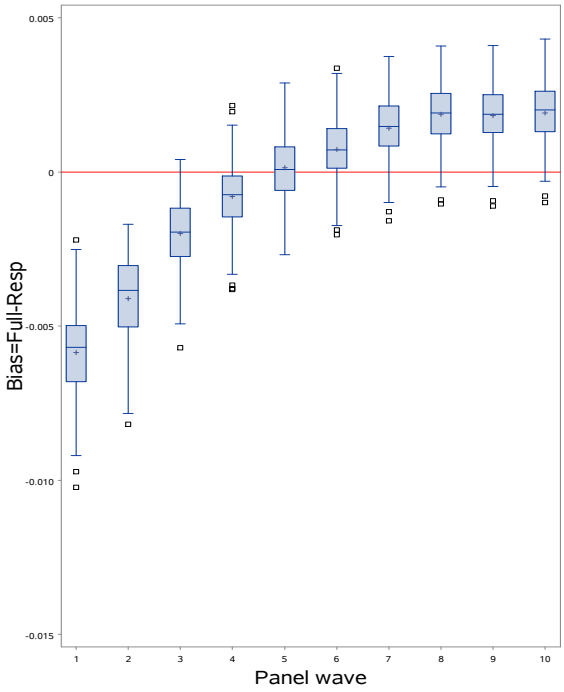
(a) Box plot for intercept.

(b) Box plot for age.



(c) Box plot for years of education (*Edu*).

(d) Box plot for job tenure.



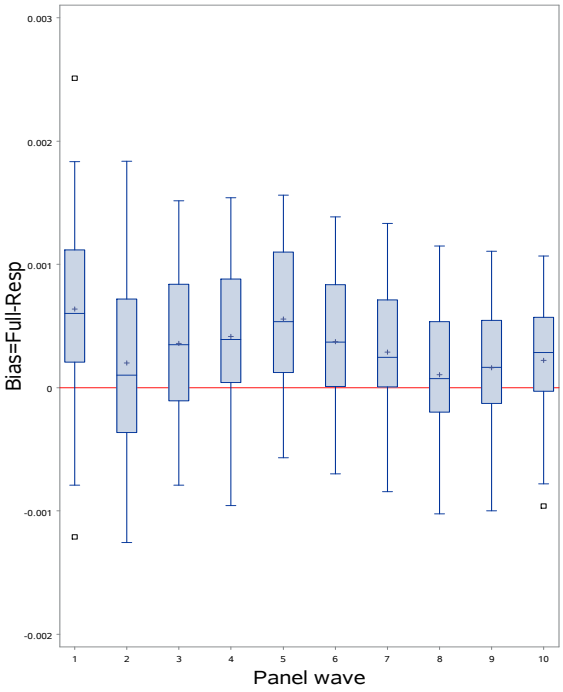
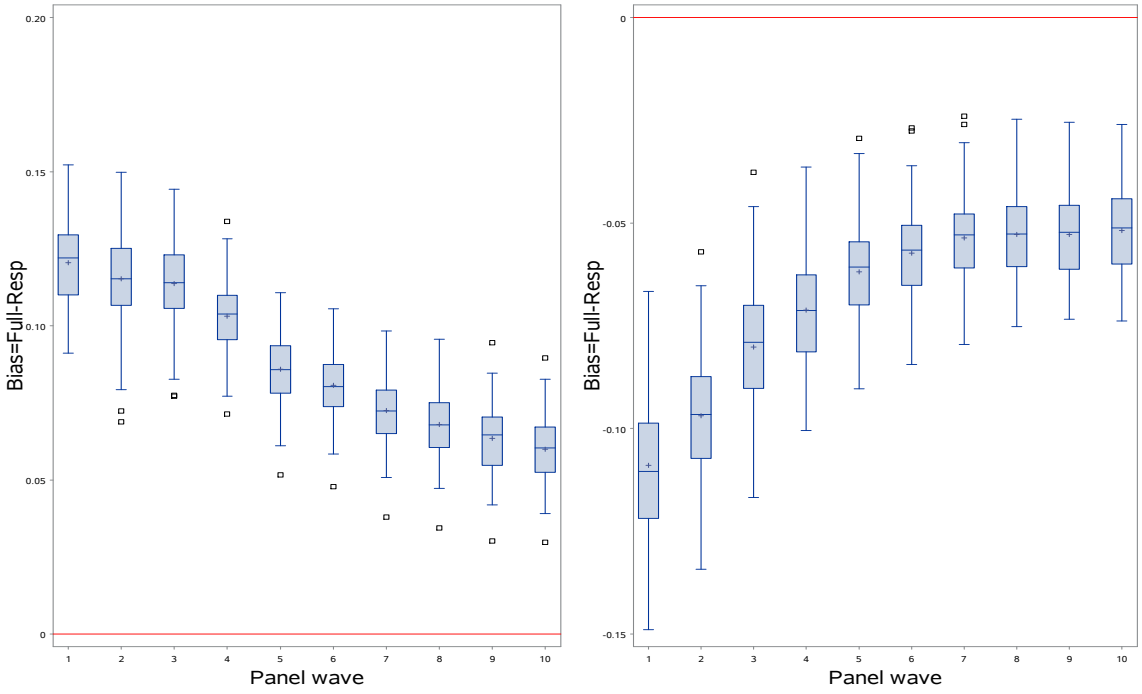


Figure 33: Box plots for the difference of estimated model parameters in the Full and the Resp samples, for 100 Monte Carlo replications of the non-response experiment.

(a) Box plot for single (D).

(b) Box plot for firm size 20-199 (F_3).



(c) Box plot for firm size 200-1999 (F_4).

(d) Box plot for firm size 2000+ (F_5).

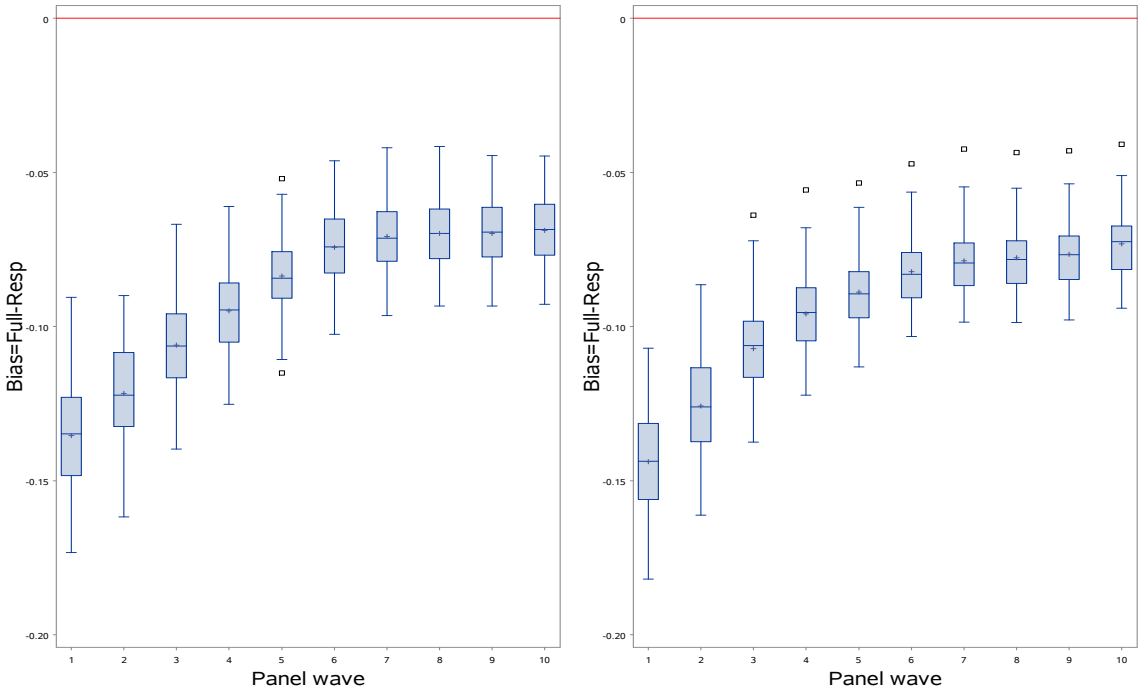


Figure 34: Box plots for the difference of estimated model parameters in the Full and the Resp samples, for 100 Monte Carlo replications of the non-response experiment.

(a) Box plot for male (G).

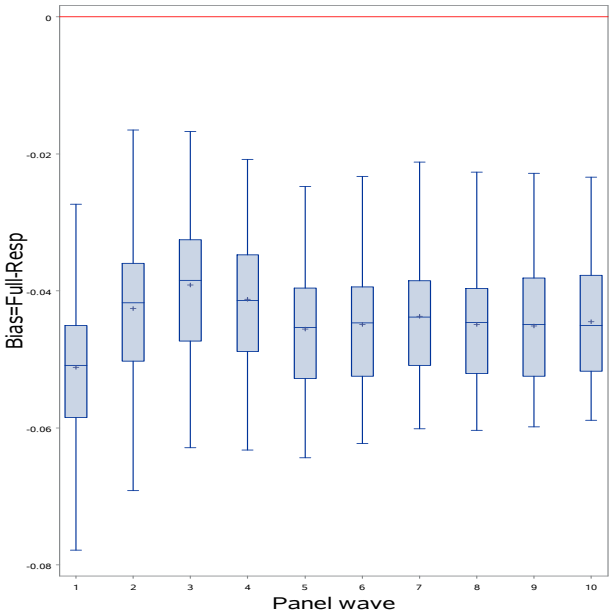


Figure 35: Box plots for the difference of estimated model parameters in the Full and the Resp samples, for 100 Monte Carlo replications of the non-response experiment.

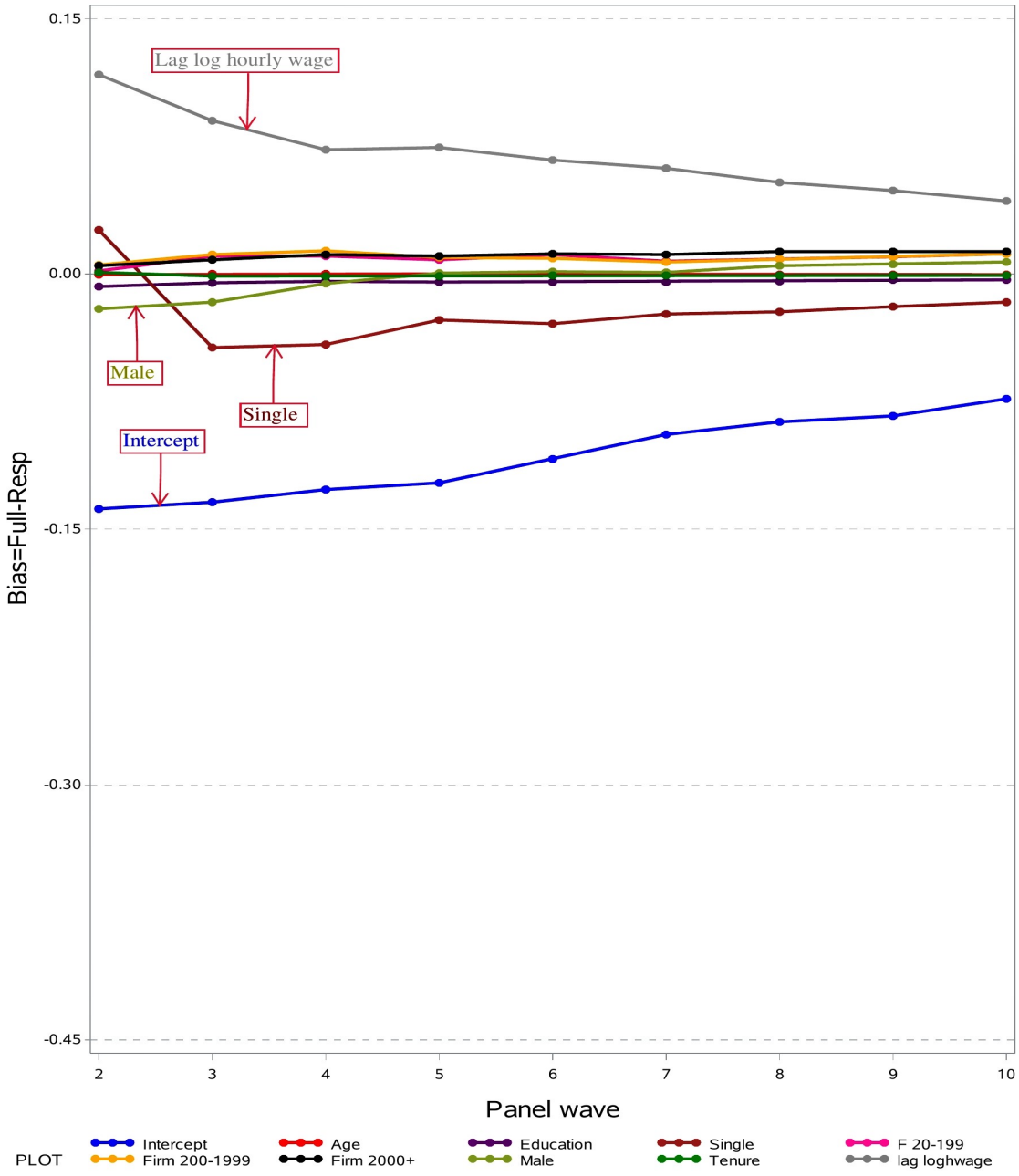


Figure 36: Graphical display of the fade-away of bias of the cross-sectional OLS estimator with lagged $W_{i,t-1}$ using SOEP data and artificial initial non-response.

Note: The vertical axis displays the bias, while the horizontal axis displays panel waves. Number of Monte Carlo replications is $R = 100$. Initial non-response rate is 30%. The bias in each panel wave is obtained by the difference of the regression coefficients in the two samples i.e., $\text{bias}(\hat{b}_{p,t}) = \hat{b}_{p,t}^{Full} - \hat{b}_{p,t}^{Resp}$, where p denotes the estimated coefficients of the intercept, age, education, single, firm size, gender, lagged log hourly wage and tenure, respectively. The points on the graph as highlighted in different colors represent the biases in the estimates in certain panel waves.

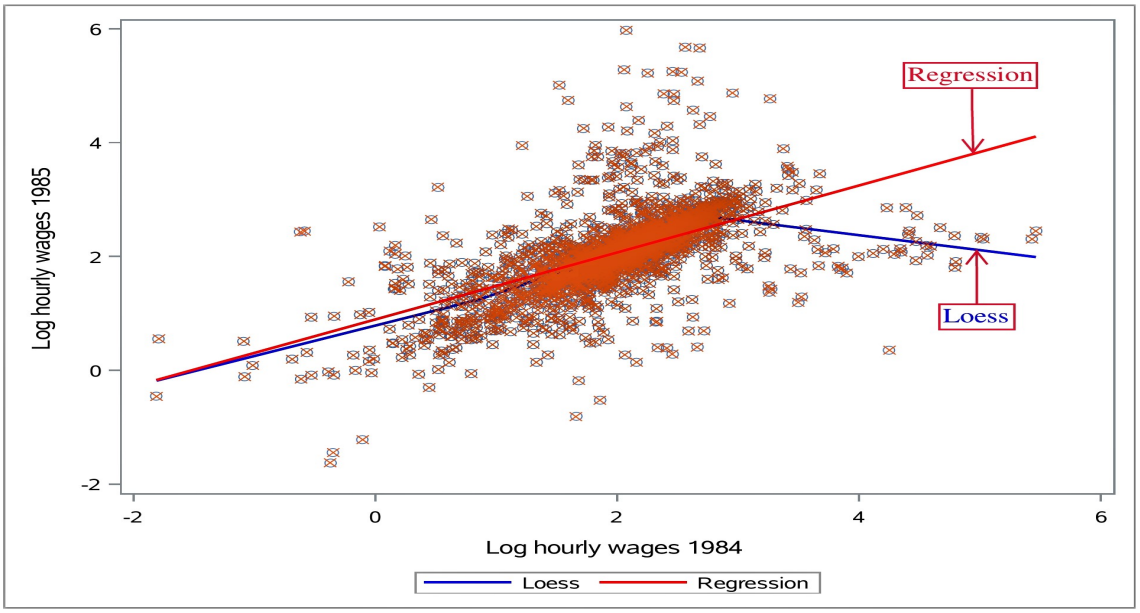


Figure 37: Display a scatter plot of log hourly wage in 1985 and log hourly wage in 1984 in the Full-Sample. The solid line in red color represents the regression line, while the solid line in blue color represents the LOESS line.

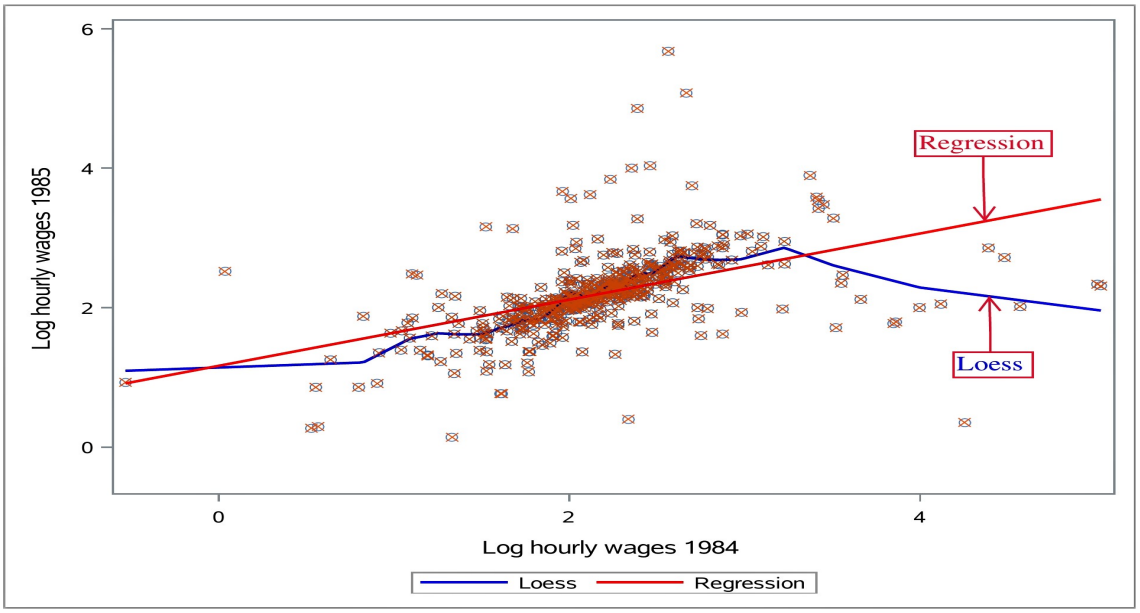


Figure 38: Display a scatter plot of log hourly wage in 1985 and log hourly wage in 1984 in the Resp-Samples. The solid line in red color represents the regression line, while the solid line in blue color represents the LOESS line.

4.2.4.2. Regularity conditions for the fade-away effect

The fact of a non-response bias for the coefficient of the lagged hourly wage is surprising, since non-response is completely explained by the hourly wage in 1984 and the lagged wage is used as a control variable in the regression model. Therefore we investigate this setting in more detail. We assume that the variable of interest W_t, here the log hourly wage, follows a Markov chain with a finite state space $E = \{1, 2, 3, \ldots, I\}$:

$$P(W_t = j | W_1 = i_1, W_2 = i_2, \ldots, W_{t-1} = i) = P(W_t = j | W_{t-1} = i), \tag{4.10}$$

Let R_1 be the response indicator at wave 1, such that $R_1 = 1$ indicates response and $R_1 = 0$ indicates non-response at wave 1. Then the distribution on the income states at wave t in the Resp-Samples is $P(W_t | W_{t-1}, X_t, R = 1)$. The fade-away hypothesis assumes that $P(W_t | W_{t-1}, X_t, R = 1)$ converges to $P(W_t | W_{t-1}, X_t)$ for large t, where X_t is a set of covariates at time t, and W_{t-1} is the lagged dependent variable at time $t-1$. Now we have

$$\begin{aligned}
P(W_t | W_{t-1}, X_t, R = 1) &= \frac{P(W_t, W_{t-1}, X_t, R = 1)}{P(W_t, X_t, R = 1)} \\
&= P(W_t, R = 1 | W_{t-1}, X_t) . \frac{P(W_{t-1}, X_t)}{P(W_t, X_t, R = 1)} \\
&= P(W_t, R = 1 | W_{t-1}, X_t) . \frac{1}{P(R = 1 | W_{t-1}, X_t)} \\
&\stackrel{!}{=} P(W_t | W_{t-1}, X_t) . \frac{P(R = 1 | W_{t-1}, X_t)}{P(R = 1 | W_{t-1}, X_t)} \qquad (4.11) \\
&= P(W_t | W_{t-1}, X_t) \qquad (4.12)
\end{aligned}$$

Equation (4.11) results from the assumption of conditional independence of W_t and R.

$$P(W_t, R = 1 | W_{t-1}, X_t) = P(W_t | W_{t-1}, X_t) . P(R = 1 | W_{t-1}, X_t) \tag{4.13}$$

For $t = 2$, we have

$$P(W_2, R = 1 | W_1, X_2) = P(W_2 | W_1, X_2) . P(R = 1 | W_1) \tag{4.14}$$

where the last term $P(R=1|W_1)$ is the starting distribution for the Resp-Samples at wave 1.
For $t > 2$ we have:

$$P(W_t, R=1|W_{t-1}, X_t) = P(W_t|W_{t-1}, X_t).P(R=1|W_{t-1}, X_t) \tag{4.15}$$

If there is no direct impact of W_1 on W_t, then equation (4.15) holds. This holds, for example, in the case of a first-order Markov chain. However, for Markov chains of higher-order a direct impact of W_1 on W_t may not be excluded.

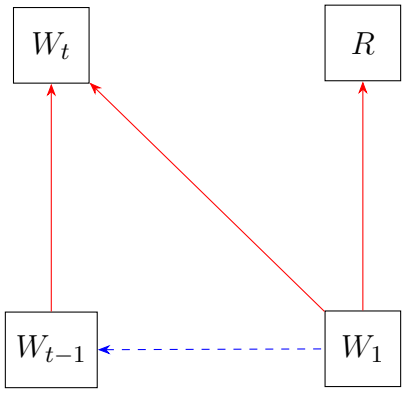


Figure 39: A path diagram

In the setting of our non-response experiment, the response probability depends entirely on the value of W_1. If we keep W_1 fixed, we would conclude that there is no direct impact of R on W_2. As a consequence, we would expect no impact of R on the conditional distribution $P(W_2|W_1, X_2)$. However, as our simulation results have demonstrated, this conclusion is wrong as we observed a severe non-response bias for the estimation of the slope coefficient of W_1. The reason for this bias is the misspecification of the impact of W_1 on W_2. There is apparent non-linearity of the impact of W_1 on W_2. As our model assumes an overall linear relationship between W_1 on W_2, R is informative for the value of W_2. Here, $R=1$ indicates that we have probably smaller values of W_2 than expected by an overall linear relationship. Similarly, $R=0$ indicates that W_2 is well predicted by an overall linear relationship.

The use of weighting in the context of regression analysis is the most commonly used statistical procedure to compensate for non-response and attrition. There are numerous types of weighting adjustments procedures available in the literature which

are used to account for non-response in the context of regression analysis. In this section, and throughout this thesis, our agenda is on the use of a non-response propensity score weighting estimator. This estimator is also called an inverse probability weighted (IPW) estimator. This method is based on the conditional probability of response of each individual in the sample given the set of covariates. According to this method, we first calculate the conditional probability of response. In the second step, we weight the data by the inverse of the response probability of each individual in the sample. Then we analyze the regression using the weighted data. Since in the SOEP simulation approach we have a reasonable estimate of response probability of approximately 0.70, then we use the inverse of this probability as a weighting variable in the cross-sectional OLS estimator as well as in the panel model estimators. We will interpret the fade-away results for the cross-sectional IPW estimator in this subsection, while for panel model estimators it will be interpreted in the next subsection (Subsection 4.2.4.3).

From Figure 40, we notice that the use of inverse probability as weights is very helpful in reducing the effect of initial non-response and its selection effects on the estimates. It is worth to mention that the bias of the un-weighted cross-sectional OLS estimator in Figure 32 is significantly larger than the bias of the IPW estimator in Figure 40. For example, in Figure 32 the bias of the OLS estimator for the effect of firm sizes varies from 15 to 5, while the effect is very small for the IPW estimator in Figure 40 which varies approximately between 2 to 0 only. Note that in the SOEP setting the weighting substantially reduces the bias of the estimates against the un-weighted OLS estimates, which contradicts with the simulation study in Section 3.5 of Chapter 3 (where weighting does not help in reducing the bias of the estimates). There are two reasons for the different effects of IPW in the simulation study and the SOEP data. First, in the SOEP case, we used the correctly specified response propensity model, while in the simulation study we used the wrong weighting model. Second, in the SOEP data, all the relevant variables are included in the model which explains wages, whereas in the simulation study the model contains only one covariate.

As the non-response model and attrition model depend on the income it is reasonable to augment the model by the lagged income. Further, as the lagged income is a reasonable predictor for the present income it may be regarded as a valuable control variable to keep the non-response bias small. The motivation to

include "lagged log hourly wage $W_{i,t-1}$" in the model is to reduce the effect of non-response on the estimated slope coefficients. Therefore, by using the lagged dependent variable as an explanatory variable in the IPW estimator, the estimated regression coefficients are estimated with no bias. However, one has to wait until wave 2 until $W_{i,t-1}$ is available. Since we control for the initial non-response there is no large bias in the first wave of the panel so there is no fade-away effect in follow-up panel waves (see Figure 41 for detail). Further, details about the estimated coefficients and the size of biases of the IPW estimator with and without lagged dependent income variable $W_{i,t-1}$ are given in Table 48 to Table 53 in Section B.1 of Appendix B.

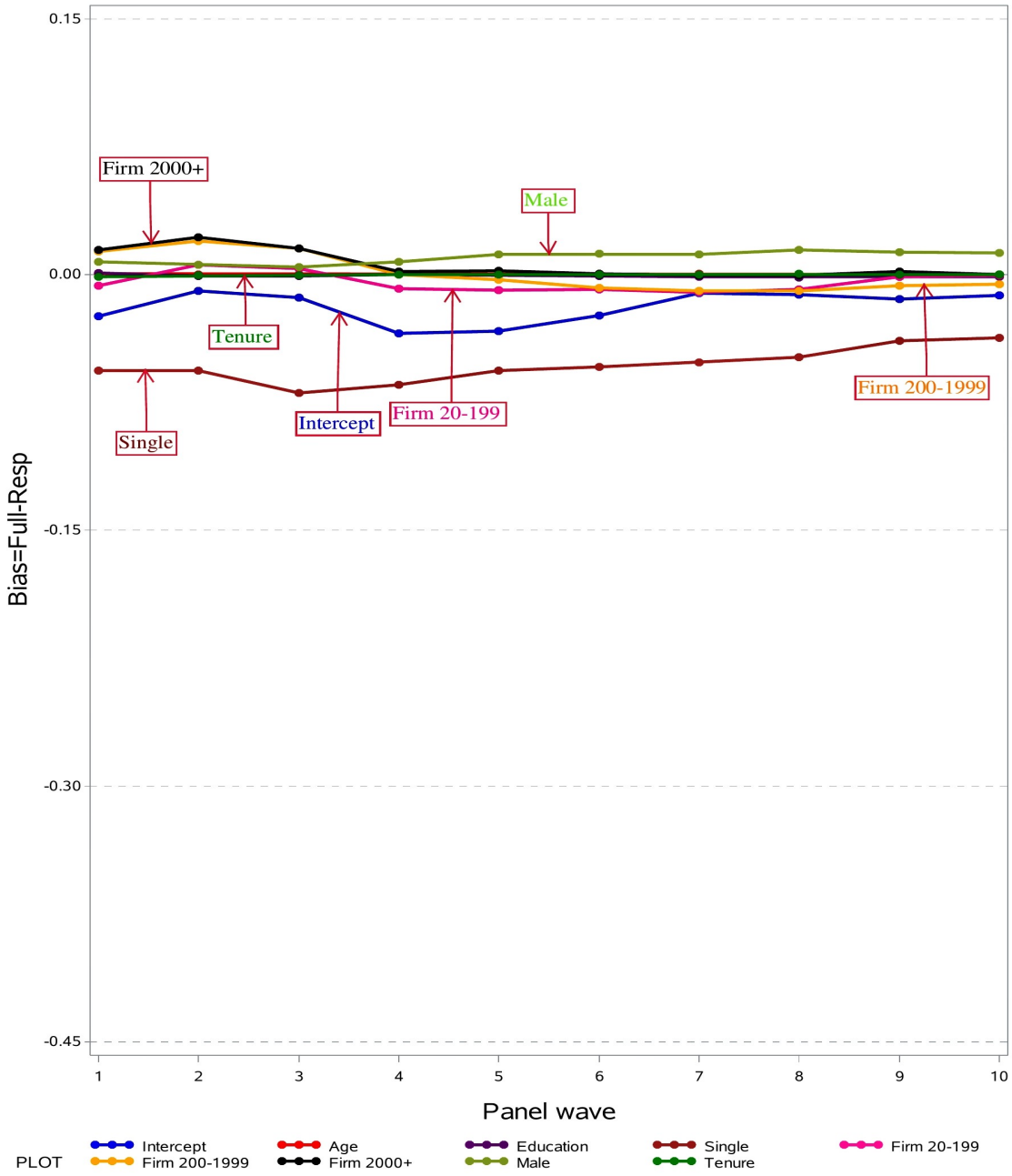


Figure 40: Graphical display of the fade-away of bias of the cross-sectional IPW estimator, with SOEP data and artificial initial non-response.

Note: The vertical axis displays the bias, while the horizontal axis displays panel waves. Number of Monte Carlo replications is $R = 100$. Initial non-response rate is 30%. The bias in each panel wave is obtained by the difference of the regression coefficients in the two samples i.e., $\text{bias}(\hat{b}_{p,t}) = \hat{b}_{p,t}^{Full} - \hat{b}_{p,t}^{Resp}$, where p denotes the estimated coefficients of the intercept, age, education, single, firm size, gender, and tenure, respectively. The points on the graph as highlighted in different colors represent the biases in the OLS estimates in certain panel waves.

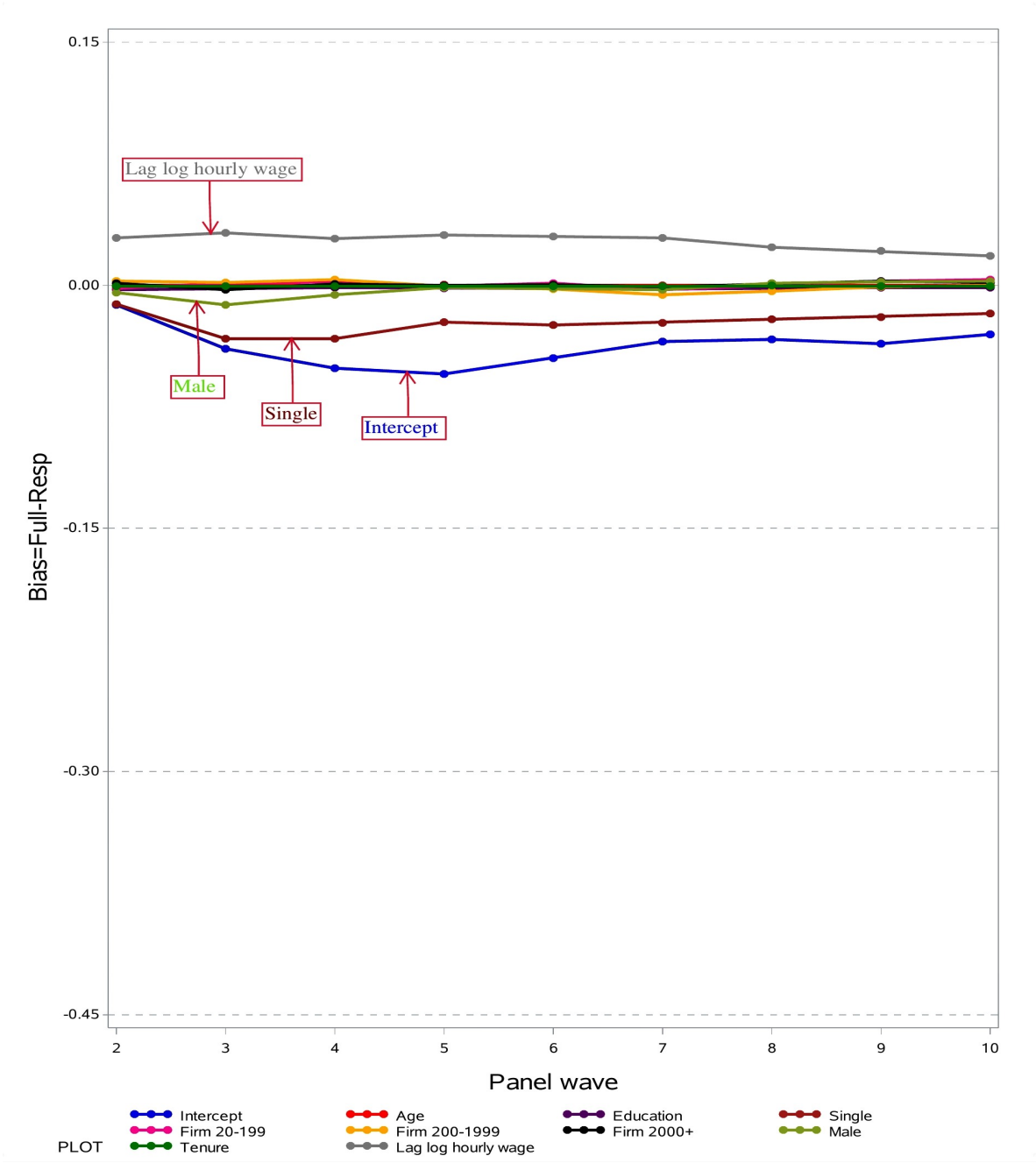


Figure 41: Graphical display of the fade-away of bias of the cross-sectional IPW estimator with lagged $W_{i,t-1}$, using SOEP data and artificial initial non-response.

Note: The vertical axis displays the bias, while the horizontal axis displays panel waves. Number of Monte Carlo replications is $R = 100$. Initial non-response rate is 30%. The bias in each panel wave is obtained by the difference of the regression coefficients in the two samples i.e., $\text{bias}(\hat{b}_{p,t}) = \hat{b}_{p,t}^{Full} - \hat{b}_{p,t}^{Resp}$, where p denotes the estimated coefficients of the intercept, age, education, single, firm size, gender, lagged log hourly wage and tenure, respectively. The points on the graph as highlighted in different colors represent the biases in the OLS estimates in certain panel waves.

4.2.4.3. Longitudinal panel model results

In the previous subsection, we investigated the fade-away effect of the cross-sectional OLS estimator from wave 1 to wave 10 of the SOEP. However, one disadvantage of the regular OLS is that it doesn't control for the individual heterogeneity/unobserved heterogeneity that varies across cross-sections but is constant over time and thus ignoring the panel structure of the data. These unobserved time-invariant variables capture all the unobserved time-invariant factors which affect the wages $W_{i,t}$. Ignoring such factors from the model may sometime results in a heterogeneity bias. Therefore, the purpose of this section is to control for the unobserved heterogeneity in the panel we use different panel models estimators. We estimate three different models: A RE model with and without lagged dependent income $W_{i,t-1}$, a RE model with auto-correlated errors and a FE model, respectively. Contrary to the cross-sectional OLS estimator where the SOEP database changes from wave to wave, here we use panel estimators of different lengths. This is the number of all panel waves which enter to panel estimators.

In order to investigate the fade-away effect for the estimates of these panel models, we proceed as follows: First, we find the bias for each cross-sectional OLS estimate in the year 1984 (wave 1). We denote this by length 1. We then add the year 1985 (wave 2) to the initial year 1984 and find the biases for the panel estimates based on the two year longitudinal length. We denote this by length 2. Similarly, we extend the longitudinal lengths up to length 10 when we have data from the year 1984 to 1993. Thus, we estimate the model parameters of the Full and the Resp-Samples for each longitudinal length (length 1 to length 10). The biases of the estimates are then obtained by the difference of the Full-Sample and the Resp-Samples estimates in a given longitudinal length. We present the estimates and the biases of the estimators (the RE model estimator, the RE model estimator with auto-correlated errors and the FE Within estimator) of different lengths in Table 54 to Table 65, respectively in Section B.1 of Appendix B. For the graphical representation of the fade-away effect we plot the biases of the estimators over a different length in Figure 42 to Figure 45, respectively.

It can be clearly seen from the graph of the RE model estimator in Figure 42, that as the length of the database (longitudinal length) increases the corresponding biases reduces in the subsequent lengths, however after some lengths of time the

biases become stable over the rest of the lengths (length 6 to 10). For example, the effect of singles (color: brown) decreases from -13 to -9 at length 6 and then it remains stable for the rest of the lengths. Similarly, the effect of firm sizes (black, orange and deep pink on the top of Figure 42) reduces in a geometrical fashion up to length 6, and then it becomes stable. Similar to cross-sectional estimation results, there is apparently no bias for the panel estimation of the effect of age (color: red), years of education (color: purple) and tenure (color: green). The same results for the fade-away of the bias also hold for the more standard panel model that is the RE model estimator with auto-correlated errors, this is visualized in Figure 43.

Since the income of a person at time t depends heavily on its previous income at time $t-1$. Also as the non-response and attrition models depend on income it is reasonable to augment the model by the lagged income. Therefore, lagged income is a good predictor for the present income and its inclusion in the wage equation as a covariate may keep the non-response bias small. However, one has to wait until wave 2 until $W_{i,t-1}$ is available. The exclusion of the lagged dependent variable from the model results in an omitted variable bias and increases the occurrence of auto-regression arising from model misspecification. In such a scenario it could be the case that the model estimates may not reliable. Therefore, we consider the estimation of the RE model with lagged dependent variable in order to account for omitted variable bias as well as reduce the occurrence of auto-correlation originating from model misspecification and best fit the model. As expected, by controlling for the lagged hourly wages $W_{i,t-1}$ in the estimator of the RE model the biases of the estimator is very small and due to low bias, there is no fade-away effect present in subsequent lengths. We display this in Figure 44. This is because when we control for the non-response in wave 1, the effect on the different slope coefficients is very small and therefore there is no further reduction of the initial non-response bias in subsequent panel lengths.

If we compare the results for the panel estimators with the cross-sectional OLS estimator results from the previous section (for detail see Figure 36) we notice a much smaller fade-away effect for the longitudinal estimators. This is due to the fact that the panel estimators use information from the first panel waves. Therefore, for panel estimators, the speed of the fade-away effect is weaker than the cross-sectional OLS estimator. Although the longitudinal estimators seem to be more efficient because of the use of a larger database they are prone to be affected by biased data

from the first panel waves. To overcome this dilemma, it might be useful to discard observations from the first panel waves. However, this topic is not discussed in this thesis and is the topic for future research work.

From the simulation results in the cross-sectional case, we know that the size of the fade-away effect depends on the size of the permanent and the transient components. If the size of the permanent component is large then their distribution stays stable and the distorting effects of initial non-response remain permanent, while this doesn't hold for the stability of the transient component which swings into a steady-state distribution. However, as we know that the FE Within estimator is based on OLS regression of individual changes, so the effect of the individual FE is canceled out by differences at the individual level. If non-response is based on the permanent components then under the FE Within estimator such distorting effect of the initial non-response is annihilated by taking differencing at the individual level. Therefore, the distorting effect of initial non-response bias melts down to zero at the second wave when the difference estimator is used. Hence, the use of the FE Within estimator plays an important role in reducing the effect of non-response based on the permanent components and is therefore robust against non-response based on permanent components. We demonstrate the fade-away result of the Within estimator in Figure 45. On the contrary, the speed of the fade-away effect of the Within estimator in 45 is considerably low (except for the years of education) in comparison with the speed of the fade-away effect of the pooled OLS estimator (see Figure 32) and the RE model estimator (see Figure 42). However, one disadvantage of the Within estimator is that one can't estimate the slope coefficients of time-invariant variables, and therefore the effect of gender is dropped out from the model by taking first differences.

Here it is attractive to report the empirical ratio of the permanent error component to the total variance component to get a clue about which of the many stability scenarios of the simulation study fits the empirical needs. Therefore, we checked the ratio of the variance components under the Full and the Resp samples through the estimation of the RE model over different lengths. The variance of components and their ratio are reported in Table 7. It is visible from the table that in both cases the size of the ratio is varying between 40% to 67% which is the medium size. Therefore, medium stability (Scenario B: $\kappa = \gamma = \rho = \phi = 0.50$) is the more realistic one.

Table 7: Variance component estimate for random effects and residual term under RE model.

	Full-Sample			Resp-Sample		
Length	Pid	Residual	Ratio	Pid	Residual	Ratio
2	0.13	0.19	0.67	0.07	0.19	0.40
3	0.12	0.19	0.63	0.08	0.18	0.42
4	0.12	0.18	0.63	0.08	0.17	0.44
5	0.11	0.18	0.62	0.08	0.17	0.45
6	0.11	0.17	0.64	0.08	0.16	0.50
7	0.11	0.17	0.65	0.08	0.15	0.50
8	0.11	0.16	0.67	0.08	0.15	0.52
9	0.11	0.16	0.66	0.08	0.14	0.54
10	0.10	0.16	0.67	0.08	0.14	0.56

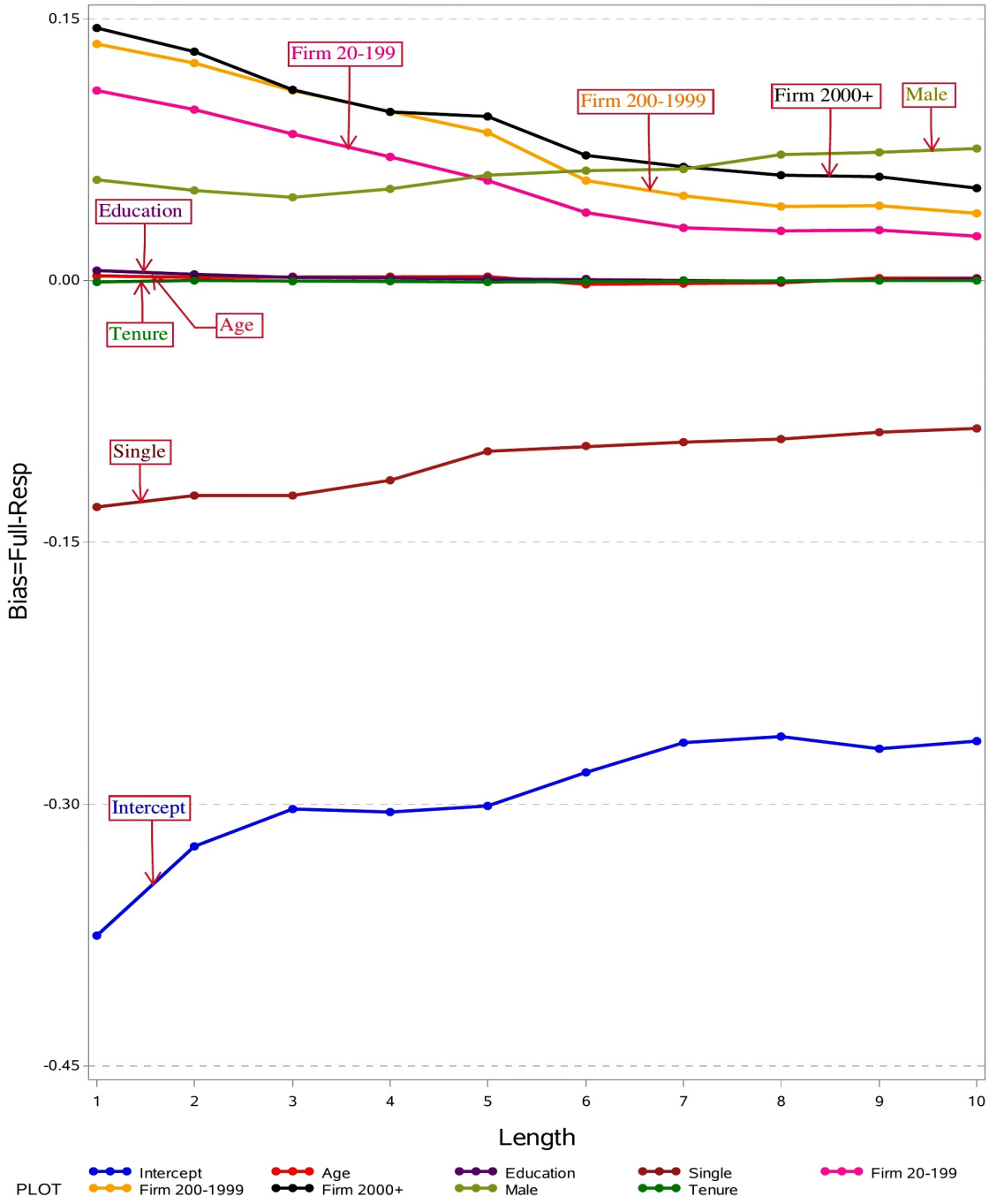


Figure 42: Graphical display of the fade-away of bias of the RE model estimator,with SOEP data and artificial initial non-response.

Note: The vertical axis displays the bias, while the horizontalaxis displays the length of the number of panel waves. Number of Monte Carlo replications is $R = 100$. Initial non-response rate is 30%. The points on the graph as highlighted in different colors represent the biases in the estimates in a certain panel length.

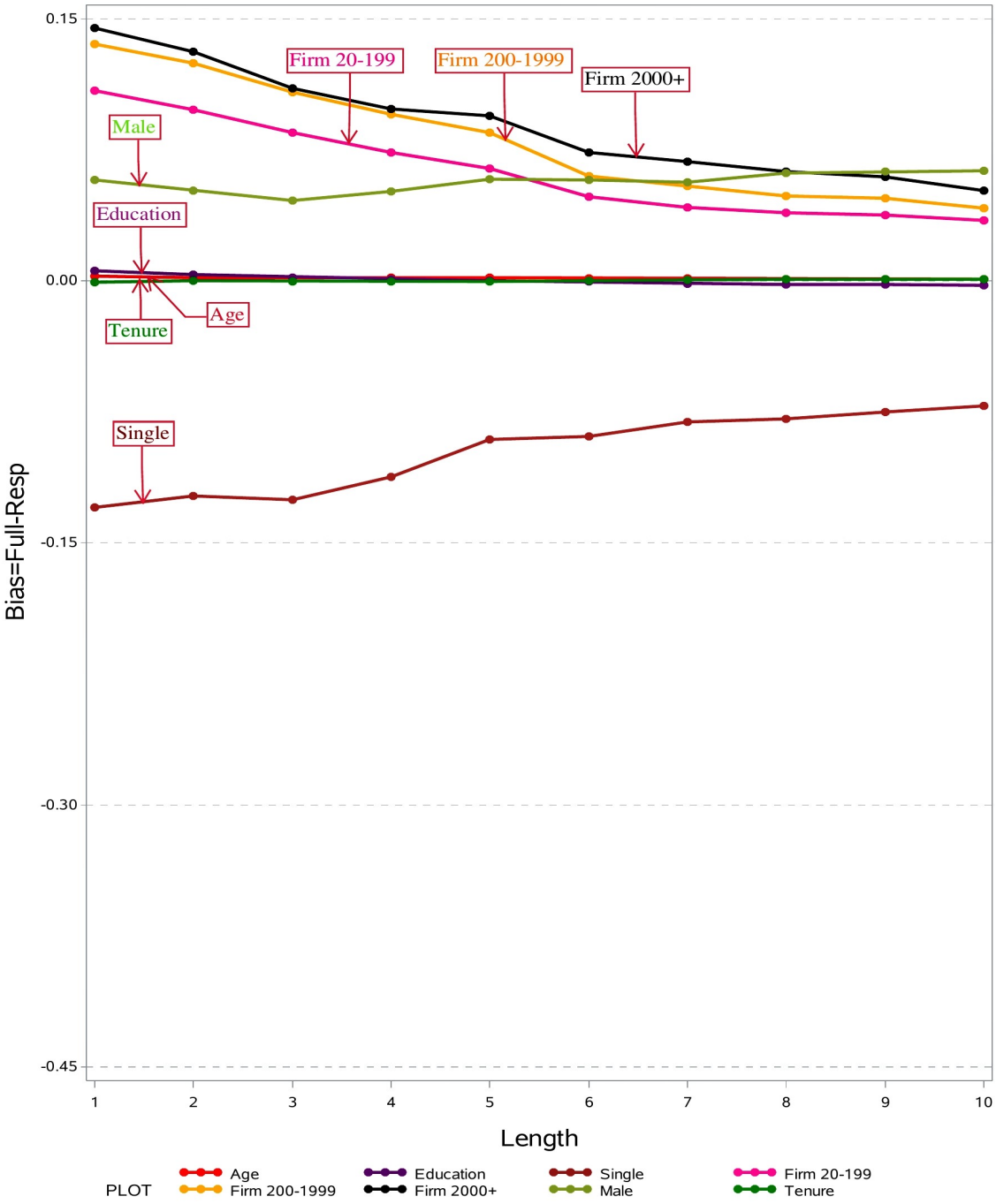


Figure 43: Graphical display of the fade-away of bias of the RE model estimator with auto-correlated errors, with SOEP data and artificial initial non-response.

Note: The vertical axis displays the bias, while the horizontal axis displays the length of the number of panel waves. Number of Monte Carlo replications is $R = 100$. Initial non-response rate is 30%. The points on the graph as highlighted in different colors represent the biases in the estimates in a certain panel length.

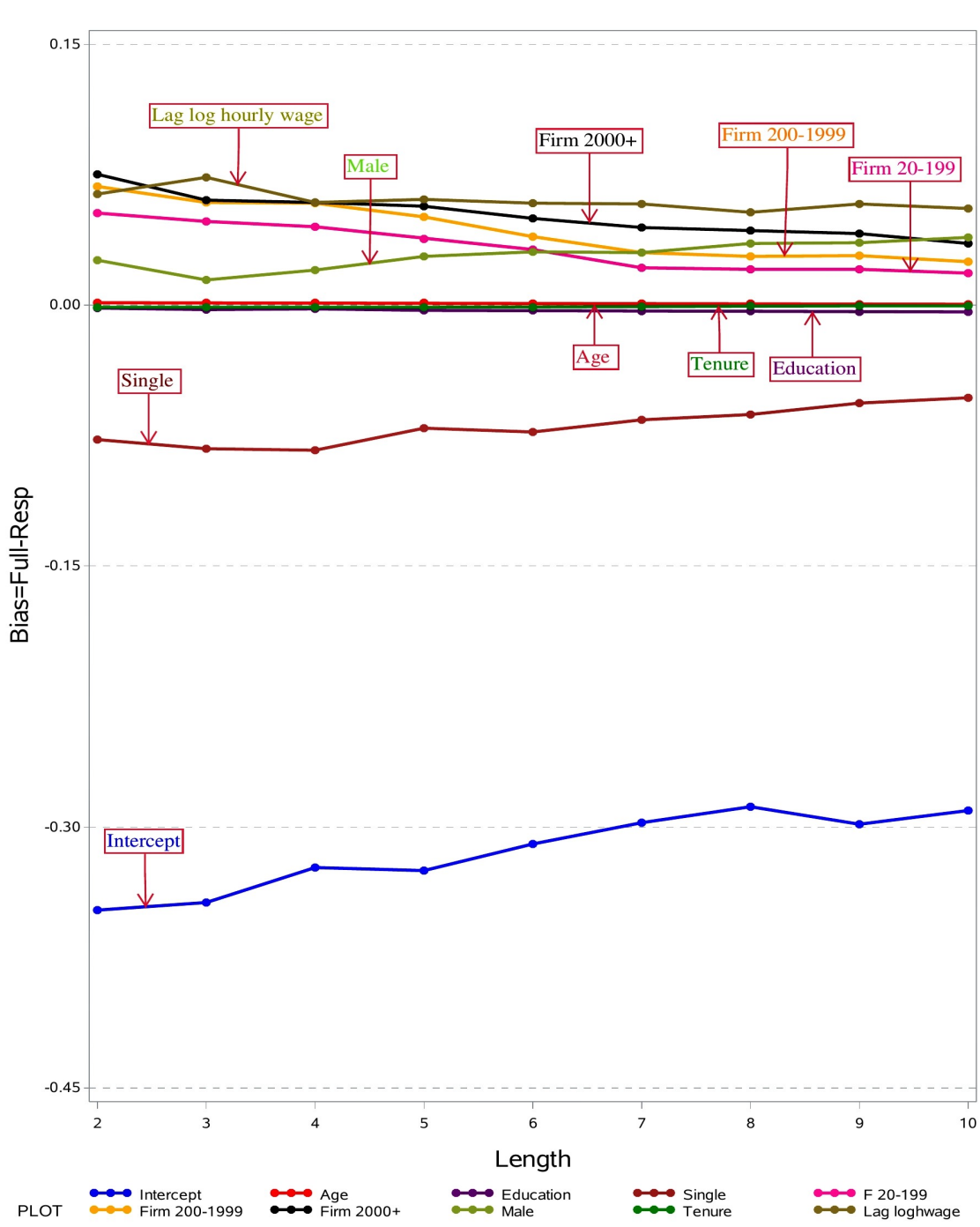


Figure 44: Graphical display of the fade-away of bias of the RE model estimator with lagged $W_{i,t-1}$ using SOEP data and artificial initial non-response.

Note: The vertical axis displays the bias, while the horizontal axis displays the length of the number of panel waves. Number of Monte Carlo replications is $R = 100$. Initial non-response rate is 30%. The points on the graph as highlighted in different colors represent the biases in the estimates in a certain panel length.

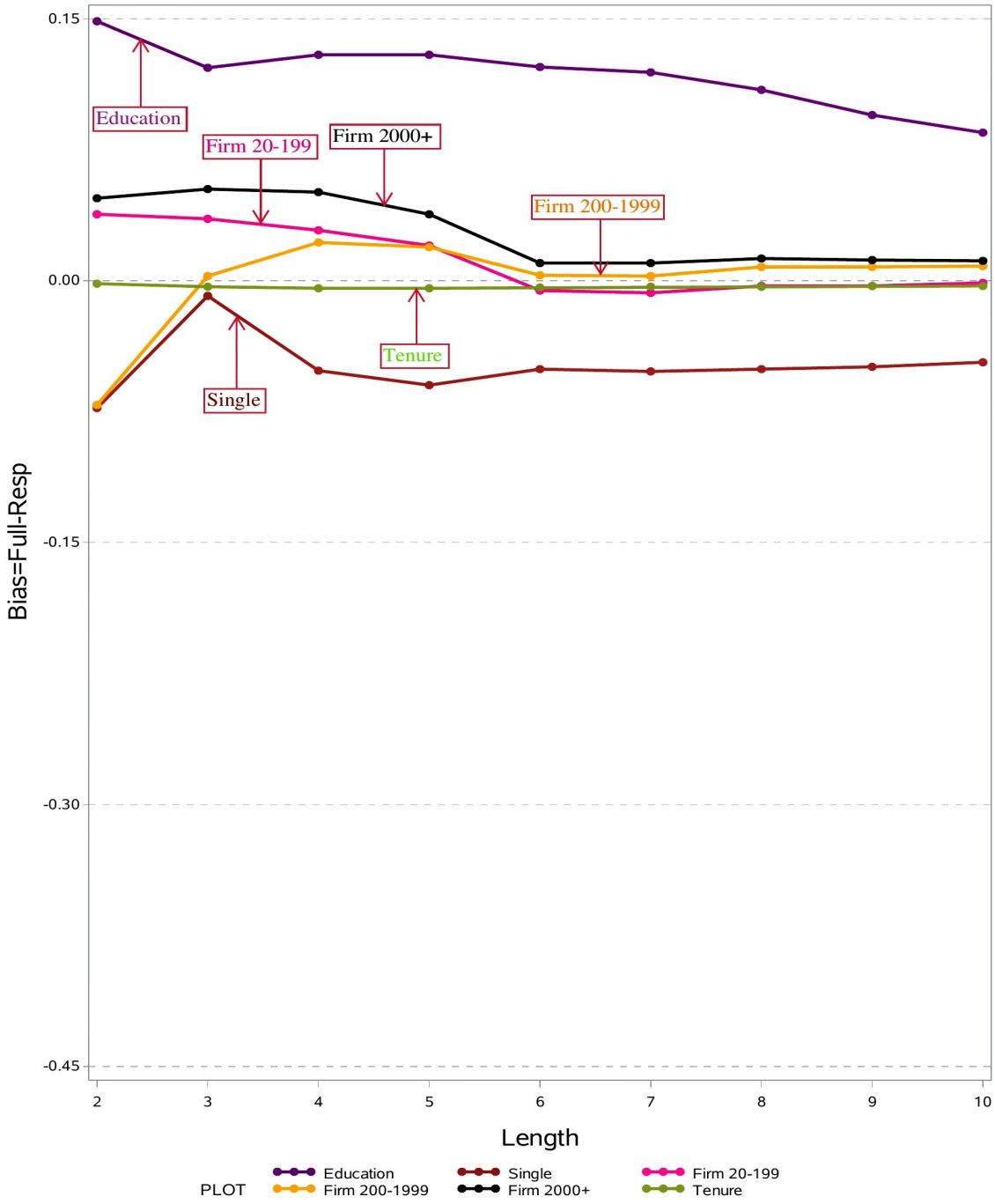


Figure 45: Graphical display of the fade-away of bias of the FE Within model estimator with SOEP data and artificial initial non-response.

Note: The vertical axis displays the bias, while the horizontal axis displays the length of the number of panel waves. Number of Monte Carloreplications is $R = 100$. Initial non-response rate is 30%. The points on the graph as highlighted in different colors represent the biases in the estimates in a certain panel length.

We now turn our discussion to the bias correction method, we use the standard IPW-method to correct for the bias of the model estimates. This method uses a logit model of response. Then the inverse of the logit probability of response is used as a weighting variable. The size of the non-response bias of the IPW estimator is then obtained by subtraction of the Full-Sample vs. Resp-Samples estimates. We start the discussion from the bias of the IPW estimator with RE which is graphically displayed in Figure 46. It can be seen from the figure that in all cases the bias of the estimator is smaller than the bias of the un-weighted RE model estimator in Figure 42. However, the difference is not so large. Thus, in this case, the use of weighting doesn't really help in reducing the bias of the estimates. However, the IPW-method is very beneficial in removing the possible effect of non-response in the RE model with auto-correlated errors. The results are shown in Figure 47. Generally speaking the size of the bias for the coefficients are really very small (except for gender, singles: as was previously discussed these variables are not changing so much over time, so the effect of these variables stays permanent after some panel waves and or length) in the first few panel lengths which disappear or faded-away in later panel lengths. Thus the use of IPW weighting substantially reduces the bias of the estimates against the un-weighted panel estimates if we use the correct weighting model.

Moreover, the inclusion of the lagged dependent income $W_{i,t-1}$ variable in the wage model as a covariate, the estimated coefficients of the IPW estimator with RE are estimated with low biases, as expected we control for the initial non-response in wave 1 of the SOEP. We plot the bias under this estimator in Figure 48. A complete list of analysis tables of the panel model estimators with weighting are presented in Section B.1 of Appendix B, see for example Table 66 to Table 74, respectively.

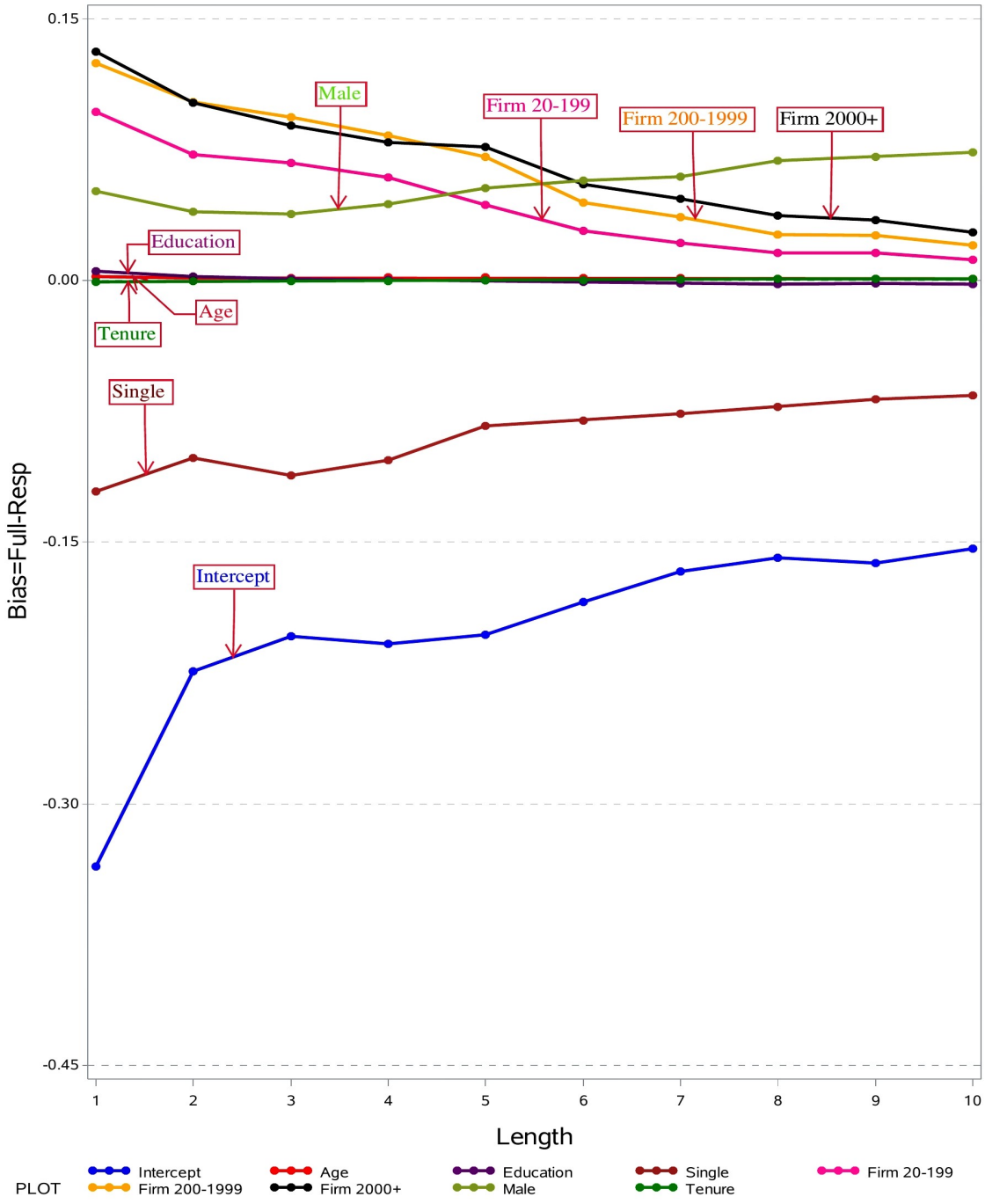


Figure 46: Graphical display of the fade-away of bias of the IPW estimator with RE, using SOEP data and artificial initial non-response.

Note: The vertical axis displays the bias, while the horizontal axis displays the length of the number of panel waves. Number of Monte Carlo replications is $R = 100$. Initial non-response rate is 30%. The points on the graph as highlighted in different colors represent the biases in the estimates in a certain panel length.

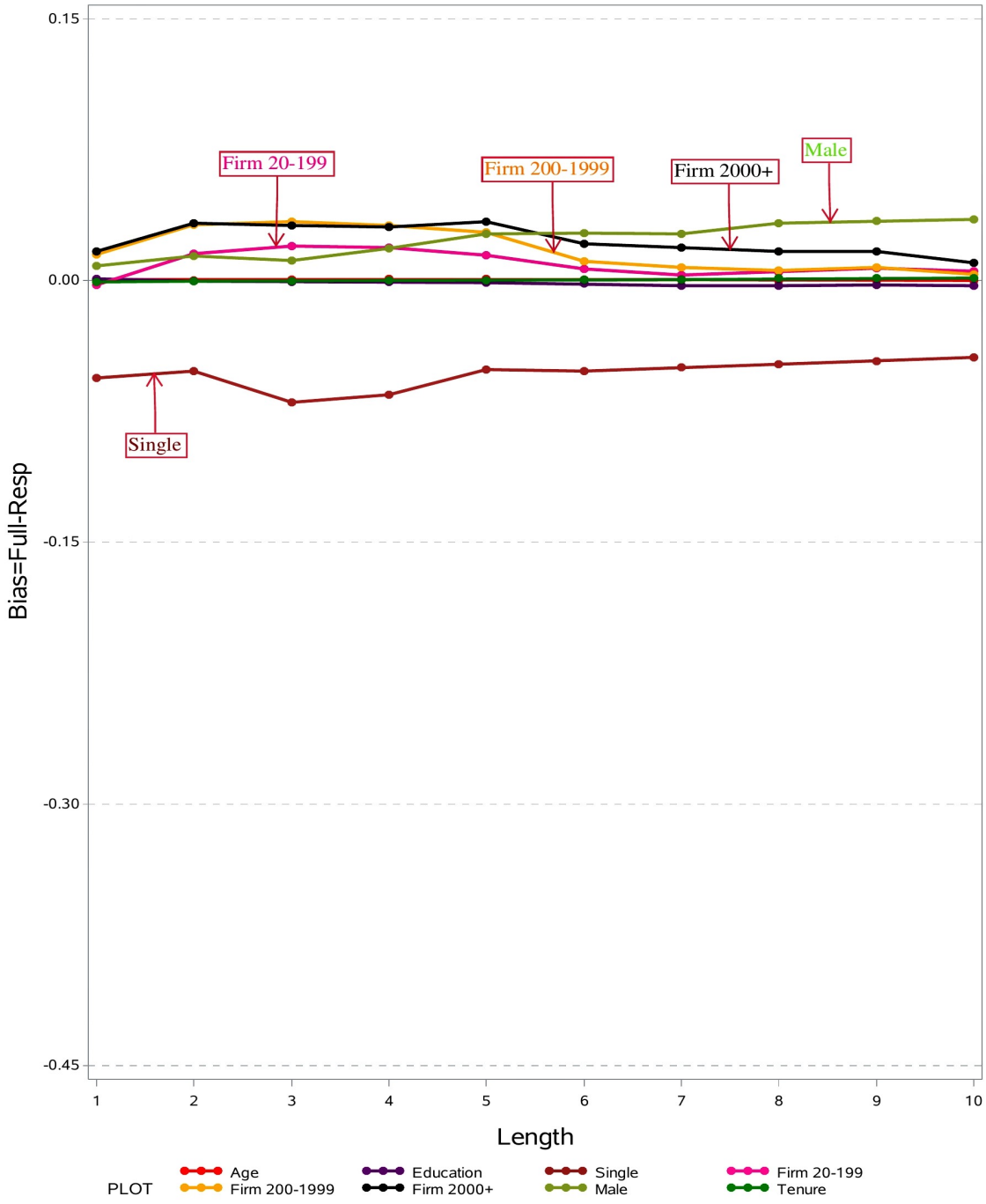


Figure 47: Graphical display of the fade-away of bias of the IPW estimator with RE and auto-correlated errors, using SOEP data and artificial initial non-response.

Note: The vertical axis displays the bias, while the horizontal axis displays the length of the number of panel waves. Number of Monte Carlo replications is $R = 100$. Initial non-response rate is 30%. The points on the graph as highlighted in different colors represent the biases in the estimates in a certain panel length.

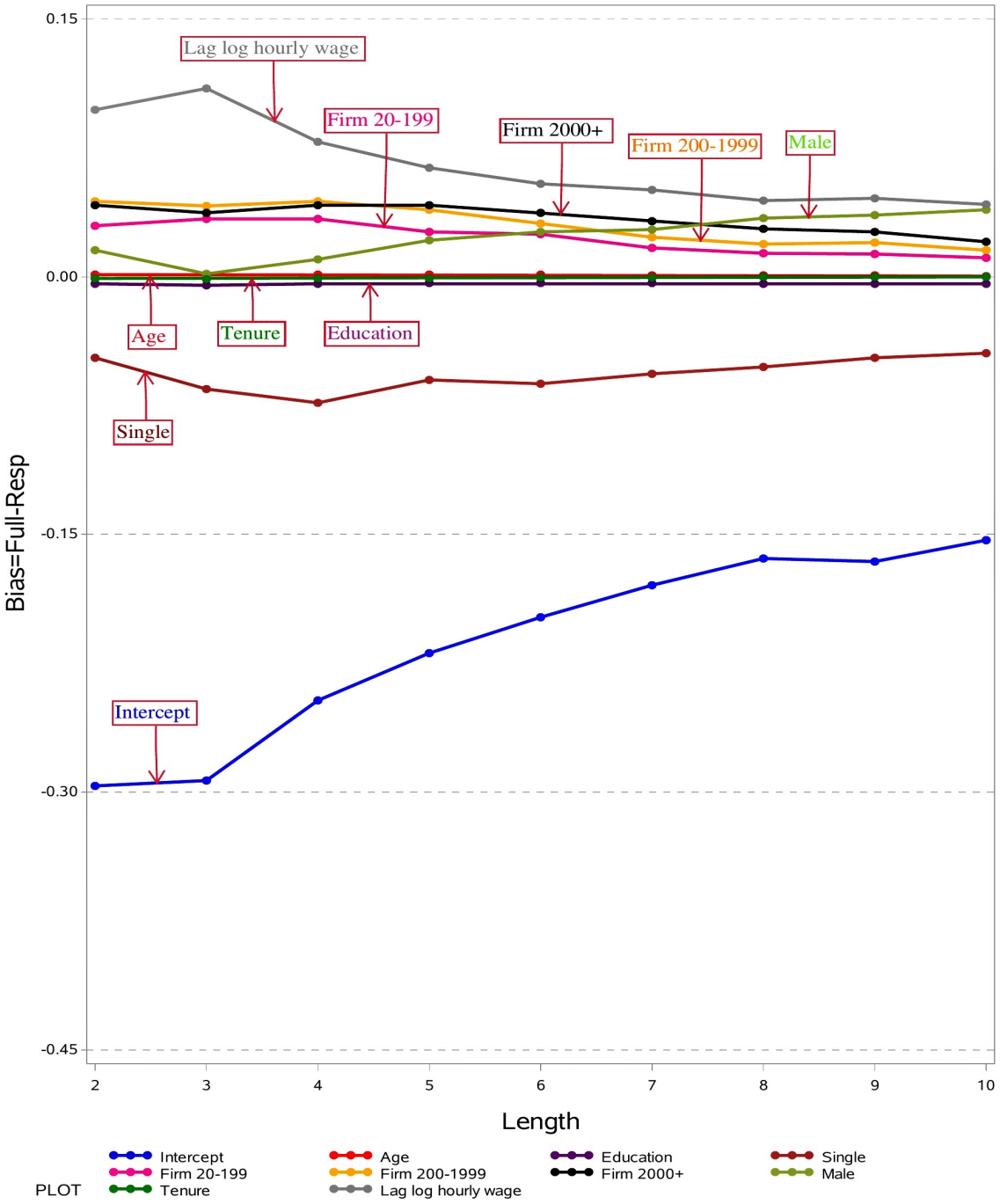


Figure 48: Graphical display of the fade-away of bias of the IPW estimator with RE and lagged income $W_{i,t-1}$, using SOEP data and artificial initial non-response.

Note: The vertical axis displays the bias, while the horizontal axis displays the length of the number of panel waves. Number of Monte Carlo replications is $R = 100$. Initial non-response rate is 30%. The points on the graph as highlighted in different colors represent the biases in the estimates in a certain panel length.

4.3. Application to SOEP life satisfaction data

4.3.1. The models

The German Socio Economic Panel (SOEP) provides rich information on the subjective and objective well-being measures, such as life satisfaction in general, satisfaction with health, number of doctor visits, and number of nights spent in the hospital, etc. In the SOEP life satisfaction comes from the response to the question "How satisfied or dissatisfied are you with your life in general right now ?". It is coded on a scale from 0 (completely dissatisfied) to 10 (completely satisfied). Similar to the life satisfaction in general, the variable "satisfaction with health" has also 11 categories where individuals are asked to report how satisfied or dissatisfied they are with their health on a scale of 0 (totally unhappy) to 10 (totally happy). In order to analyze life satisfaction, we use an ordered logit model for cross-sectional data. As the subjective well-being "life satisfaction" is measured on an ordinal scale, one might use ordered probit or logit models. There exists a vast literature work on the ordinal nature of satisfaction measures by using logit/probit models. One of theses studies Ferrer-i-Carbonell and Frijters (2004) proposed the RE ordered probit model with fixed time effects. The purpose of using theses approaches is to control for the unobserved heterogeneity that may create bias in the estimates. Before writing the logit model for satisfaction with life, first we explain the different variables which are used in model.

The dependent variable $Y_{i,t}$ measures the overall life satisfaction on an 11 point scale reported by individual i in point time t, where t refers to the selected time of interview $t = 1,2,3,\ldots,11$. $X_{i,t}^{'}$ is a vector of observed covariates, α_j are the thresholds for the different categories of life satisfaction and β is a vector of logit coefficients. $Age_{i,t}$, $Age_{i,t}^2$, and $Edu_{i,t}$ are the explanatory variables representing the age, age squared and years of education of the individual, respectively. $Gender_i$ is a qualitative explanatory variable which has two categories male and female, we use a dummy variable $male_i$ for male which is 1 if the person is male and 0 if it is female. Similarly, $Single_{i,t}$, $Widowed_{i,t}$, $Divorced_{i,t}$, and $Separated_{i,t}$ are the dummies for marital statuses: single, widowed, divorced and separated, respectively. The reference category is the married persons. $D_{visits_{i,t}}$ represents the number of annual doctor visits. A dummy variable $H_{stays_{i,t}}$ was created for the event that the person has

spent a night at the hospital last year, which is equal to 1 if the individual spent nights in the hospital last year and 0 otherwise. $HH_{income_{i,t}}$ represents household income.

Finally, $\sum_{k=1}^{6} S_{k(i,t)}$ is a set of dummies for the six categories of "satisfaction with health" starting from category 5 to category 10. For example, S_1 is a dummy variable that is equal to 1 if satisfaction with health is less than or equal to 5 and 0 otherwise, while S_2, S_3, S_4, S_5, and S_6 are the dummy variables for the categories "6-10" of satisfaction with health, respectively. To stabilize the effect of satisfaction with health on life satisfaction in general, we collapsed categories "0-4" with category 5 of satisfaction with health. Consider the following life satisfaction model $Y_{i,t}$, which is a function of, age, age squared, years of education, gender, marital status, doctor visits, hospital stays, household income, and health satisfaction:

$$P(Y_{i,t} \leq j) = \frac{\exp(\alpha_j + X_{i,t}^{'}\beta)}{[1+\exp(\alpha_j + X_{i,t}^{'}\beta)]}, \quad j = 0, 1, 2, \ldots, 10. \tag{4.16}$$

where

$$X_{i,t}^{'}\beta = \beta_1 Age_{i,t} + \beta_2 Age_{i,t}^2 + \beta_3 Edu_{i,t} + \beta_4 Male_i + \beta_5 Single_{i,t} + \beta_6 Widowed_{i,t}$$

$$+ \beta_7 Divorced_{i,t} + \beta_8 Separated_{i,t} + \beta_9 D_{visits_{i,t}} + \beta_{10} H_{stays_{i,t}} + \beta_{11} HH_{income_{i,t}}$$

$$+ \textstyle\sum_{k=1}^{6} \beta_k S_{k(i,t)}.$$

In the context of the longitudinal structure of the data set, the estimation of the model also includes FE with respect to time and individual RE with respect to individuals. The inclusion of the fixed time effects accounts for the yearly changes that are the same for all the individuals. The individual RE account for the unobserved heterogeneity that is constant over time but is different for each individual. So in order to control for the unobserved heterogeneity, we use RE approach.

$$\begin{aligned} Y_{i,t} = \beta_0 + \beta_1 Age_{i,t} + \beta_2 Age_{i,t}^2 + \beta_3 Edu_{i,t} + \beta_4 Male_i + \beta_5 Single_{i,t} + \\ \beta_6 Widowed_{i,t} + \beta_7 Divorced_{i,t} + \beta_8 Separated_{i,t} + \beta_9 D_{visits_{i,t}} + \\ \beta_{10} H_{stays_{i,t}} + \beta_{11} HH_{income_{i,t}} + \sum_{k=1}^{6} \beta_k S_{k(i,t)} + v_i + \eta_{i,t}, \end{aligned} \tag{4.17}$$

where v_i is the time-invariant random intercept associated with each person i, which is assumed to be normally distributed $v_i \overset{iid}{\sim} N(0, \sigma_v^2)$, fixed over time and orthogonal to the explanatory variables. $\eta_{i,t}$ is the error term of the person i over time t. The $\eta_{i,t}$ is assumed to be normally distributed $\eta_{i,t} \overset{iid}{\sim} N(0, \sigma_\eta^2)$ and orthogonal to the covariates included in the model and with v_i.

4.3.2. Data and descriptive statistics

The empirical analysis of this study is based on the 11 waves starting from years 2000 to 2010 of the German Socio Economic Panel (SOEP). For our analysis, we used individual-level data of the 11 waves of the SOEP. Further, we restrict our sample to the Sub-sample F[3] starting in the year 2000. The sample collects information on all individuals aged 17 and above (both men and women). The dependent variable is the general satisfaction with life. The study excluded observations with missing, imputed, negative or less than zero values from the analysis. The final sample after these reductions steps results in an unbalanced panel of 29,628 observations from 3,099 individuals of the 11 panel waves. Based on the above-described data, Table 8 provides summary statistics of the main variables used in the analysis.

[3]Specifically, here we use Sub-sample F of the German Socio Economic Panel (SOEP). In the year 2000, a new refreshment sample "Sub-sample F" was selected independently from all the other samples from the population of private households in Germany. With one exception, the selection scheme was essentially the same as for selecting Sub-sample A (for details see Section 1.4 of Spiess (2000)). The sample covers private households in Germany and greatly increases the sample size of the SOEP. Experience with the previous samples has shown that migrant households display lower response probabilities, that's why households with at least one adult not having the German nationality were over-sampled in the Sub-sample F. The total number of households in the initial wave of Sub-sample F was 6,043.

Table 8: Descriptive statistics of dependent variable and control variables, using SOEP data of individual's aged 17 and above over the sample period 2000-2010.

Variable	Mean	Median	SD	Min	Max
Life satisfaction (5-10)	7.28	8.00	1.40	5.00	10.00
Age of individual	46.66	46.00	12.16	17.00	95.00
Age squared	2324.47	2116.00	1162.09	289.00	9025.00
Years of education	12.34	11.50	2.59	7.00	18.00
Male	0.48	0.00	0.50	0.00	1.00
Single	0.18	0.00	0.38	0.00	1.00
Widowed	0.03	0.00	0.16	0.00	1.00
Divorced	0.08	0.00	0.27	0.00	1.00
Separated	0.02	0.00	0.13	0.00	1.00
Doctor visits	8.97	4.00	15.37	0.00	396.00
Hospital stays	0.10	0.00	0.30	0.00	1.00
Household income (in Euro)	46741.90	42741.00	32738.21	5.00	666832.00
Health satisfaction (5-10)					
5 (less satisfied)	0.12	0.00	0.33	0.00	1.00
6	0.10	0.00	0.30	0.00	1.00
7	0.15	0.00	0.38	0.00	1.00
8	0.26	0.00	0.44	0.00	1.00
9	0.13	0.00	0.34	0.00	1.00
10 (Completely satisfied)	0.09	0.00	0.28	0.00	1.00
Observations			29,628		

Note, that we collapse categories 0-4 with category 5 of life satisfaction and health satisfaction because the case numbers in these categories are not enough.

4.3.3. The design of the simulation study

In order to investigate the fade-away effect for the distributional differences between the distributions of the Full-Sample and the Resp-Samples, we use a simulation approach. We restrict the Full-Sample to all those individuals who continuously participate in the 11 panel waves of the SOEP starting from the year 2000 to 2010. We further exclude all those individuals who leave or temporary dropouts in any panel wave of the selected SOEP data. There are about 2,914 persons belonging to the "Full-Sample" per panel wave. We then artificially introduce an initial non-response

in the first wave of the Full-Sample under the assumption that non-response at the start of the panel is not ignorable for the estimation of population parameters. Consider a life satisfaction data $Y_{i,t}$ of an individual i in time point t in the SOEP, which is decomposed into two parts, the observed part Y_{obs} and the unobserved part Y_{mis}, by a response indicator $R_{i,t}$, such that if $R_{i,t} = 1$ the individual responds (the data is observed) and $R_{i,t} = 0$ if the individual doesn't respond (the data is missing due to initial non-response). For the initial non-response, we assume that it depends on $Y_{i,1}$. Then the logit probability of response for each individual in the initial wave is:

$$P(R_{i,1} = 1|Y_{i,1}) = \frac{\exp(\alpha + \beta Y_{i,1})}{[1 + \exp(\alpha + \beta Y_{i,1})]}, \tag{4.18}$$

where $Y_{i,1}$ is the life satisfaction scores of the individuals in wave 1. α and β are the non-response parameters.

An essential condition to demonstrate the fade-away effect is to choose that non-response at the start of the Resp-Sample should be selective. For this purpose, the non-response parameters are to be selected such that the average response probability from the model (4.18) is something between 60% to 70%. Therefore, for $\alpha = -6.00$ and $\beta = 0.90$ we generate a non-response rate of about 35% (average response probability is 65%) in the initial wave. Note, that if $\beta = 0$ or the probability of participation doesn't depend on $Y_{i,1}$ then there is no fade-away effect present. In order to get more stable results, we replicated the initial non-response 100 times and therefore we have 100 "Resp-Samples". After wave 1 we assumed that no further loss due to attrition occurs. This is motivated to demonstrate the fade-away effect in its elementary form. The Resp-Sample consists of 1,887 persons compared to the 2,914 persons in the Full-Sample. The selected value of β guarantee a substantial NMAR non-response pattern.

For the selected values of $\alpha = -6.00$ and $\beta = 0.90$ we demonstrate the impact of life satisfaction on the response probabilities in the first wave 1 of Resp-Sample in Figure 49. It can be seen that persons with high satisfaction levels have the trend to respond with higher probability than persons with low satisfaction levels. For example, in the case of low satisfaction state, say at state 5 the average of response probability over all replications were estimated at 0.20. While increasing the satisfaction states the corresponding probability increases in a geometric fashion

and is 0.90 in state 10.

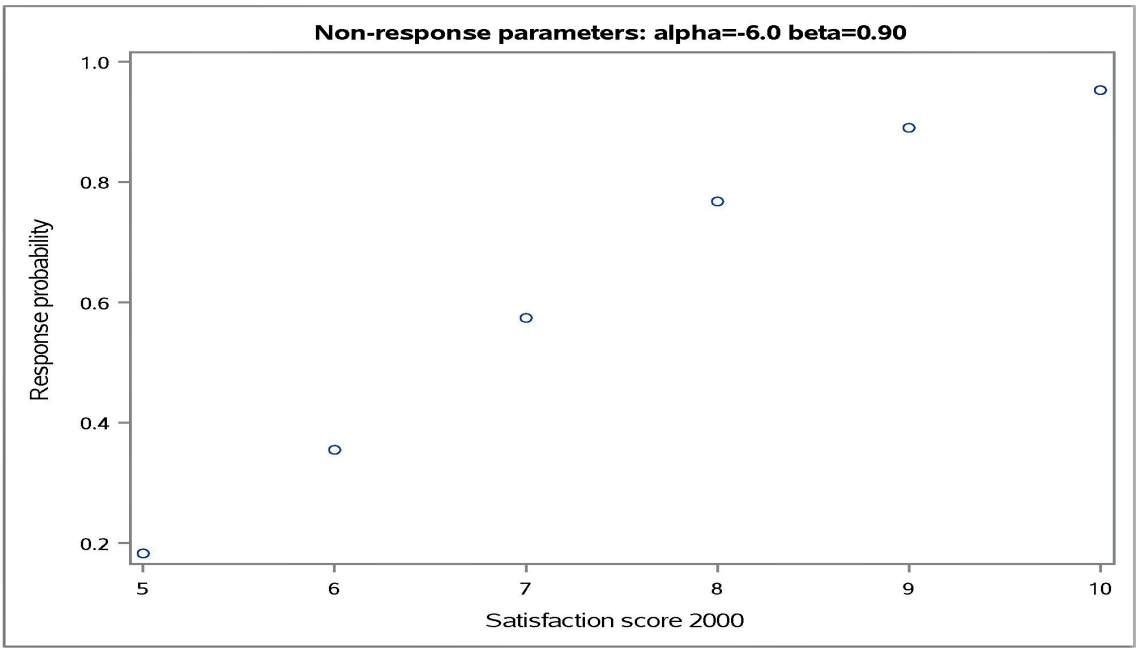


Figure 49: Impact of life satisfaction on the response probabilities.
The vertical axis displays the response probability and the horizontal axis displays life satisfaction scores (5-10) in the year 2000. The average response rate is 65%. Note that we collapsed the original categories 0 to 4 to category 5 because they are not stable.

In the following figure, we compare the distribution on the state space of the Full-Sample and the Resp-Sample in different panel waves of the Sub-sample F of the SOEP. The horizontal axis of the figure displays the satisfaction states of the Full-Sample. While the vertical axis displays the difference of the percent total frequency of the Full-Sample and the Resp-Samples participants who participate in i^{th} satisfaction state in wave t. In the figure, there are 11 different colored lines for each panel wave starting from the year 2000 to 2010. Each line has six points. These points reflect the initial non-response biases which are obtained by the difference of the Full and the Resp samples frequencies in i^{th} satisfaction state in wave t.

It is visible in Figure 50 that there is an under-representation in the Resp-Sample for those who are less satisfied (category 5-7) and over-representation for those who are more satisfied with their life. Regarding the fade-away of the bias, it is obvious that the distributional differences of the Full-Sample and the Resp-Sample are caused by the initial non-response which fade-away in later panel waves up to

some extent but it doesn't vanish completely. According to the Markov chain of the first order, the differences should vanish completely in later panel waves, but here they don't vanish completely. This may happen in a situation where a part of the sample doesn't change their satisfaction scores at all. Such a situation is found in a mover/stayer-model, which is a mixture model where one part, the movers, follow a Markov chain and the second part, the stayers, remain in their position. For a more detailed overview of the mover-stayer models see the discussion in Subsection 2.2.4 of Chapter 2.

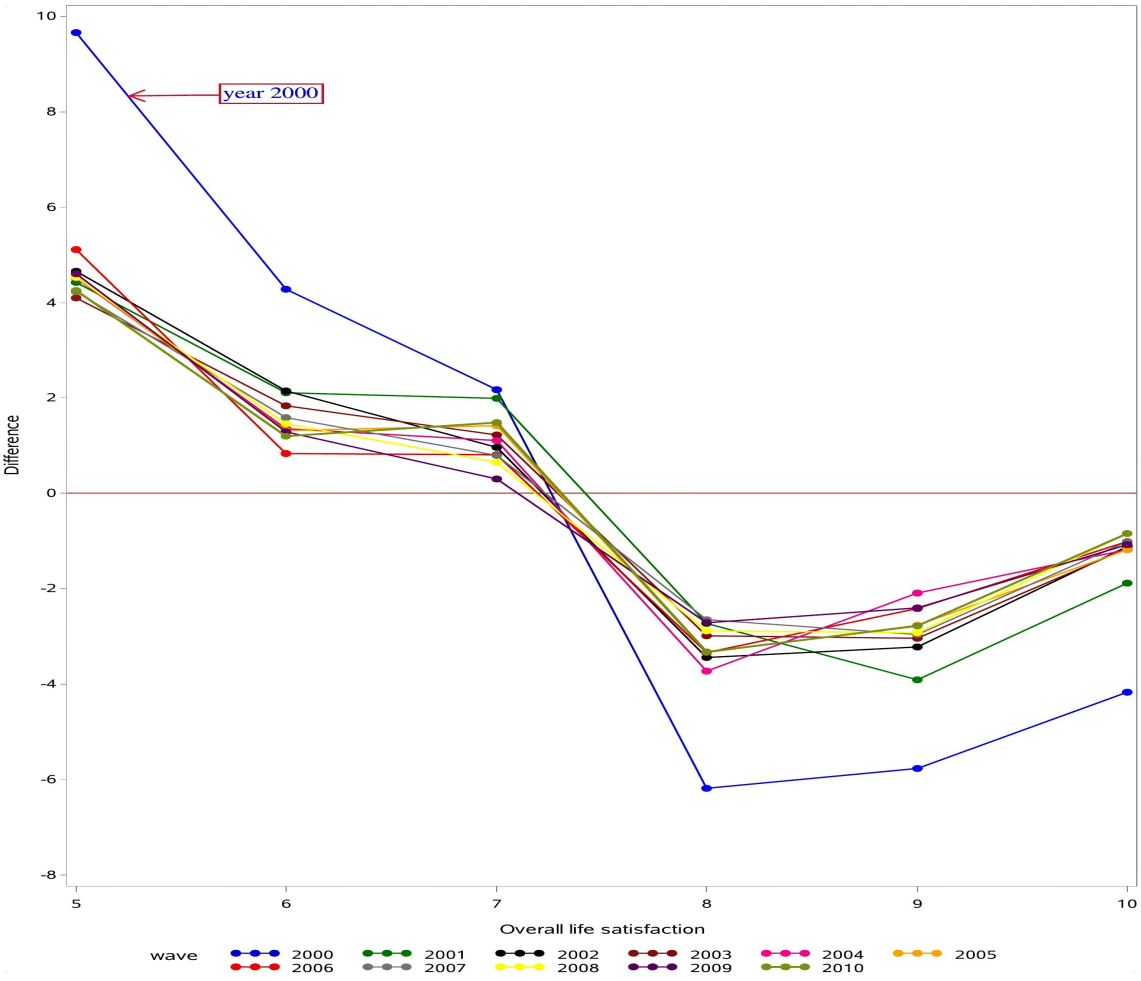


Figure 50: Distribution on satisfaction states, with non-response parameters $\alpha = -6.00$ and $\beta = 0.90$. Number of Monte Carlo replications $R = 100$ times. Initial non-response rate is 35%. Difference=percent total frequency of the Full-Sample minus percent total frequency of the Resp-Sample.

For the assessment of the non-response bias of the slope coefficients, we compared the estimates of the Full-Sample and the Resp-Samples from waves 1 to 11, respectively. Two different estimators were used here, a cross-sectional ordered logit model estimator and a linear panel model estimator with RE, respectively. Here it is important to mention that the logit approach is only used for cross-sectional analysis, whereas, for panel analysis, we used the linear model. We used the linear RE model because of the computational convenience of a longitudinal ordered logit model in the SAS software. Therefore, for panel analysis, we switch to the linear model. In the case of the panel estimator, we increased the length of the included database of the SOEP by adding sequentially further panel waves. The bias estimate in wave t is obtained by using the formula $\text{bias}(\hat{b}_{p,t}) = \hat{b}_{p,t}^{Resp} - \hat{b}_{p,t}^{Full}$, where $t = 1, 2, 3, \ldots, 11$ and the subscript p refers to the covariates: intercept, age, age squared, years of education, male, single, widowed, divorced, separated, doctor visits, hospital stays, household income, and different categories of satisfaction with health, respectively.

4.3.4. Analysis and discussion of results

4.3.4.1. The cross-sectional results

Now, we will discuss the cross-sectional regression results for life satisfaction, using an ordered logit model. Figure 51 to Figure 53 visualize the fade-away effect for different model parameters. Figure 51 displays the fade-away of the bias of the thresholds for the different categories of life satisfaction, while the effect of slope coefficients is compared in Figures 52 and in Figure 53. Figure 51 emphasizes that the effect of initial bias fades-away for the 5 thresholds of the ordered logit model over the life of the panel. For example, an initial bias of the estimate of the intercept 5 (colored blue marked with letter "N5") is 1.70 on the logit scale which reduces very fast in the subsequent waves and is about 0.20 on the logit scale in wave 11. In a similar fashion, the effect of the other categories reduces over time. This bias pattern indicates that the latent thresholds of the Resp-Samples are shifted to lower values. So for equal values of $X^{'}\beta$ we obtain higher probabilities for high values of Y. In fact, the initial response increased the percentage of persons with high Y values substantially. So the shift of thresholds in the Resp-Samples is a direct consequence of the response pattern.

The fade-away effect of the estimated slope coefficients can be seen in Figure 52 and in Figure 53. There are some small and some large initial non-response biases for the slope parameters of the covariates. Estimates having substantial initial biases show a substantial fade-away effect. However, the decline of the bias becomes weaker after about 7 waves. For illustration, consider the effect of singles (with letter "S") in Figure 52. The initial bias fades-away from 0.14 to 0.06 in wave 8 on the logit scale and then it remains stable for the rest of the panel waves. This advocates for a persistent component of the residual terms which affected the distribution by initial non-response. The effect of gender (variable male colored green) is almost stable over the entire panel waves. Also, the effect of a stay in hospital (with letter "Sh") reduces from 0.03 to 0.01 in subsequent waves on a logit scale. There is apparently no bias for the cross-sectional estimation of the effect of age (with letter "A"), age squared (with letter "As"), years of education (with letter "E"), doctor visits (with letter "Dv") and household income (with letter "Hi").

Similarly, for the different categories of health satisfaction the effect of non-response fades-away in the subsequent panel waves. For example, in Figure 53 the effect of health satisfaction 7 (with letter "H7") in the initial wave 1 is -0.39 on the logit scale which reduces to -0.09 in wave 8 and then it becomes very stable reaching to its steady-state distribution. Detailed analysis results in this section are placed in the Appendix B.2. Table 75 and Table 76 in the Appendix B.2, give the results of a regression of life satisfaction on age, age squared, years of education, male, single, widowed, divorced, separated, doctor visits, hospital stays, household income, and different categories of satisfaction with health for the Full and the Resp samples, respectively. The size of the non-response bias of the estimates is presented in Table 77 in Section B.2 of Appendix B.2.

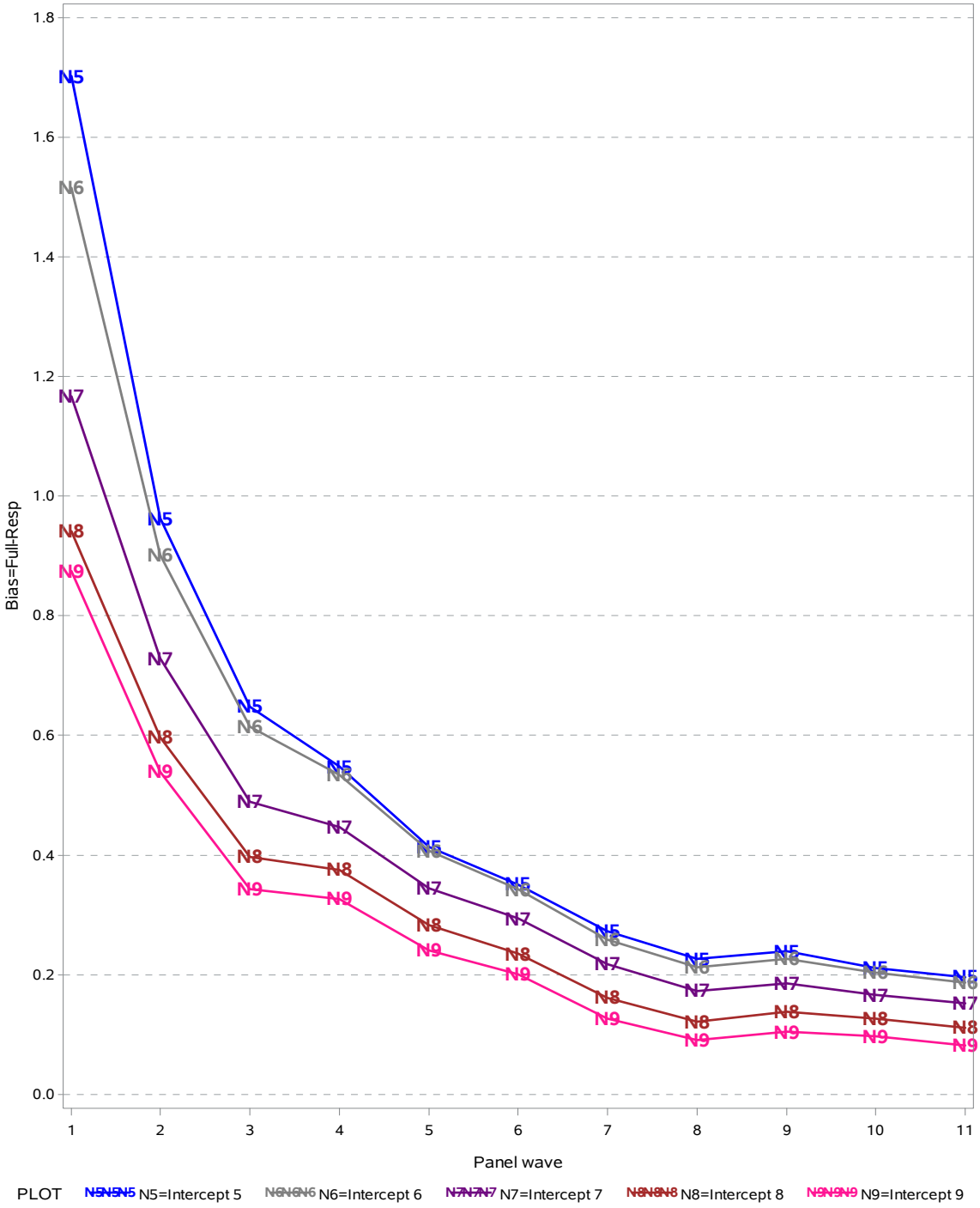


Figure 51: Graphical display of the fade-away of bias of the model thresholds, with SOEP data and artificial initial non-response.

Note: The vertical axis displays the bias of the estimates, whereas the horizontal axis displays the wave of the panel. Number of Monte Carlo replications $R = 100$. Initial non-response rate is 35%. The colored letters on the graph indicate the biases of the thresholds.

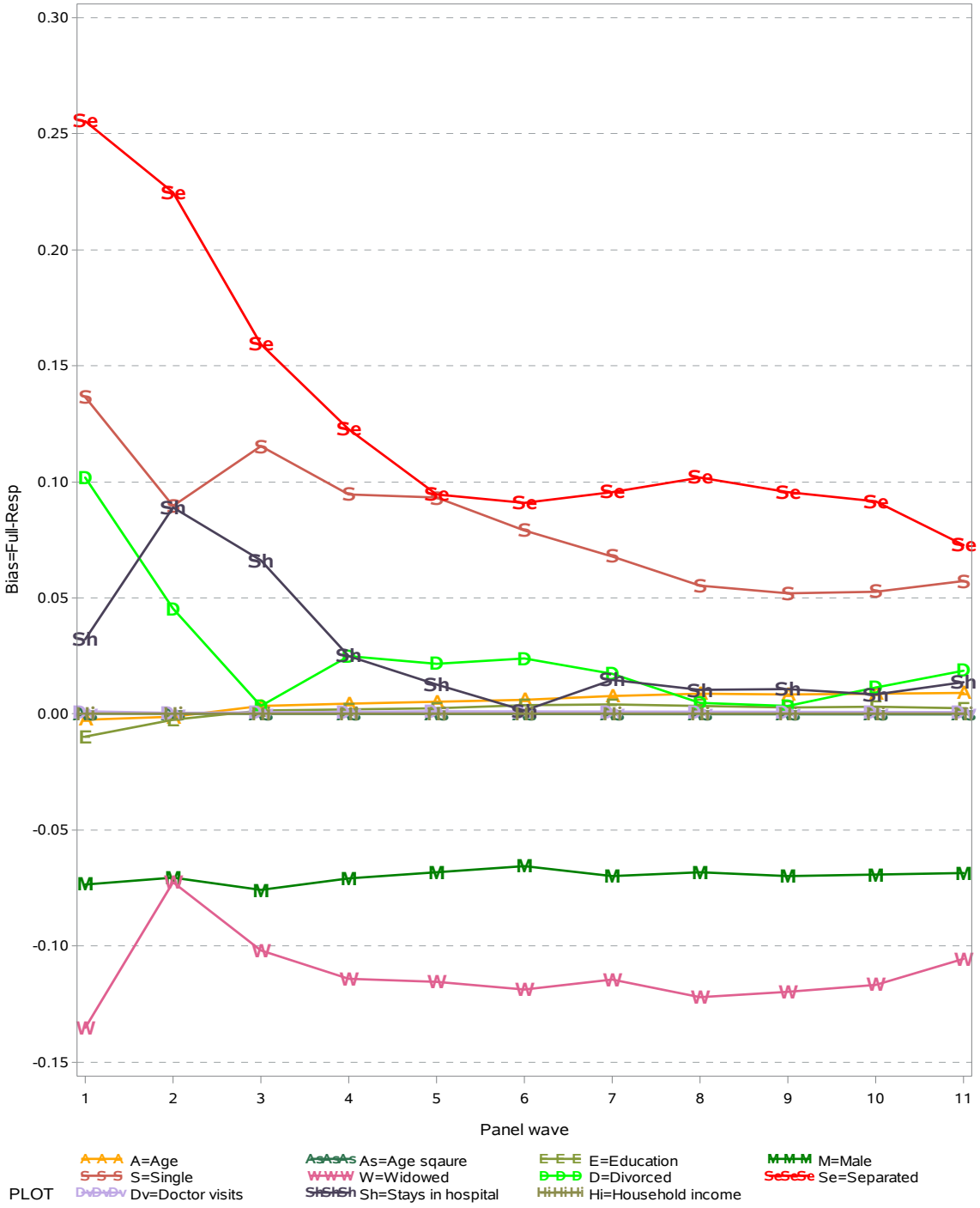


Figure 52: Graphical display of the fade-away of bias of the model estimates, with SOEP data and artificial initial non-response.

Note: The vertical axis displays the bias of the estimates, whereas the horizontal axis displays the wave of the panel. Number of Monte Carlo replications $R = 100$ times. Initial non-response rate is 35%. The letters on the graph as highlighted in different colors shows the biases of the estimates.

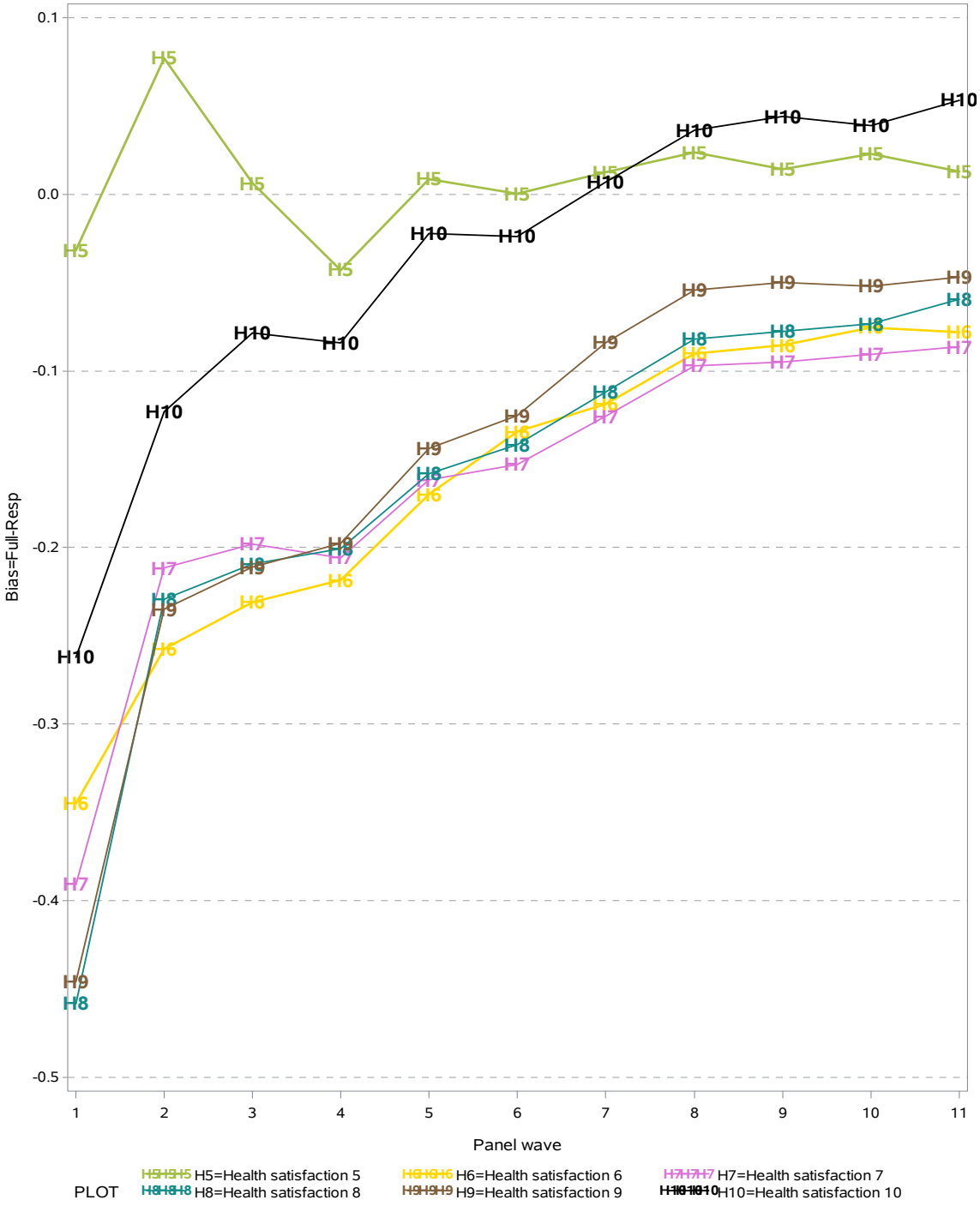
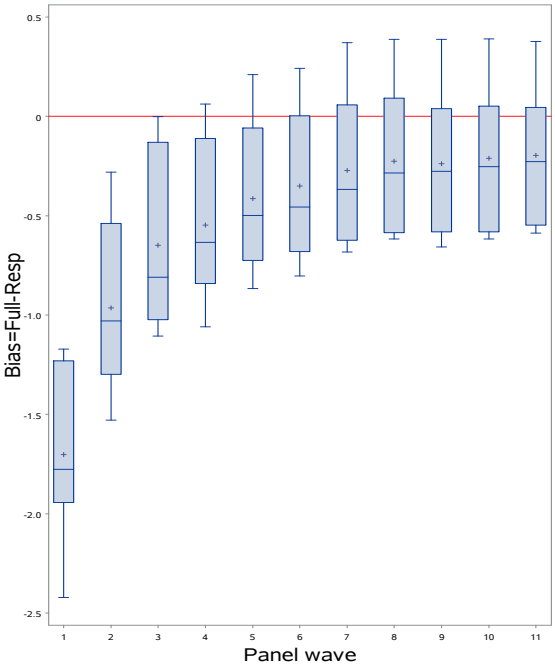


Figure 53: Graphical display of the fade-away of bias of the model estimates, with SOEP data and artificial initial non-response.

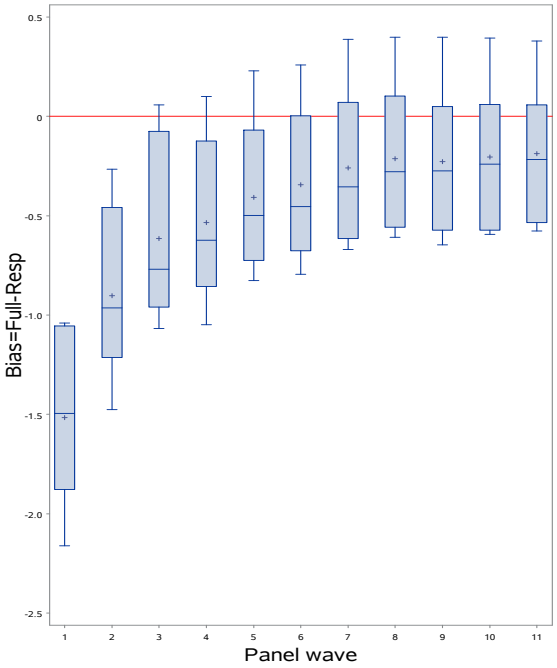
Note: The vertical axis displays the bias of the estimates, while the horizontal axis displays the wave of the panel. Number of Monte Carlo replications $R = 100$ times. Initial non-response rate is 35%. The letter on the graph as highlighted in different colors shows the biases of the estimates.

It is also interesting to visualize the fade-away effect for the distributional differences of the Full and the Resp samples estimates through a box-plot diagram, which reflects the variance of the Resp-Samples results over replications of the initial non-response. Figure 54 to Figure 59 visualize the box-plot representation of the logit estimates. The vertical axis of the figures displays the biases bias($\hat{b}_{p,t}$) of the logit estimates, while the horizontal axis displays the panel waves 1 to 11. The filled circles of the plots show the median. The lower and upper ends of the boxes are the lower and upper quartiles, and the vertical lines are used to indicate the spread and shape of the tails of the distribution. The little white circles outside the boxes indicate outliers in the data. The plots also display a horizontal red line indicating a zero bias. Interestingly, over time the boxes cover the zero bias line and also the centers of the boxes move towards the zero bias line. However, the bias of the slope coefficients for the gender variable in Figure 56 is quite different with respect to size, sign, and behaviour over panel waves. By their nature, they are quite stable over time. This advocates for a permanent error component of life satisfaction which is related to initial non-response. This is in line with the theoretical results of Alho (2015). However, the persistent biases of the estimate of age, age squared, years of education, doctor visits and household income are very small in the absolute numbers of Figure 54 to Figure 59.

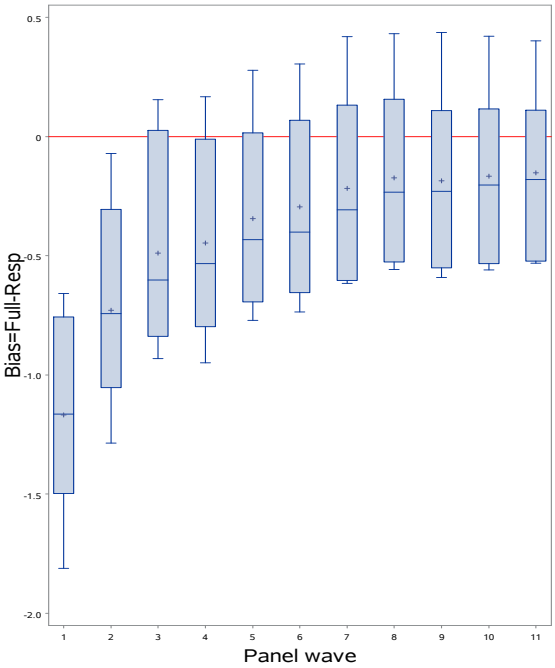
(a) Box plot for intercept 5.



(b) Box plot for intercept 6.



(c) Box plot for intercept 7.



(d) Box plot for intercept 8.

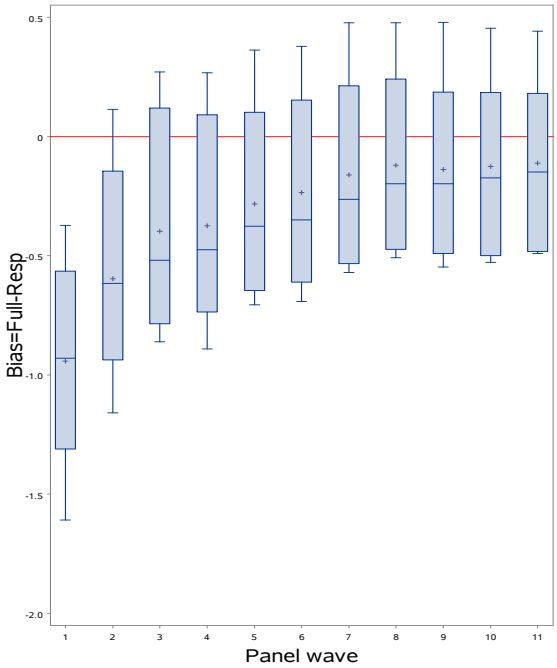
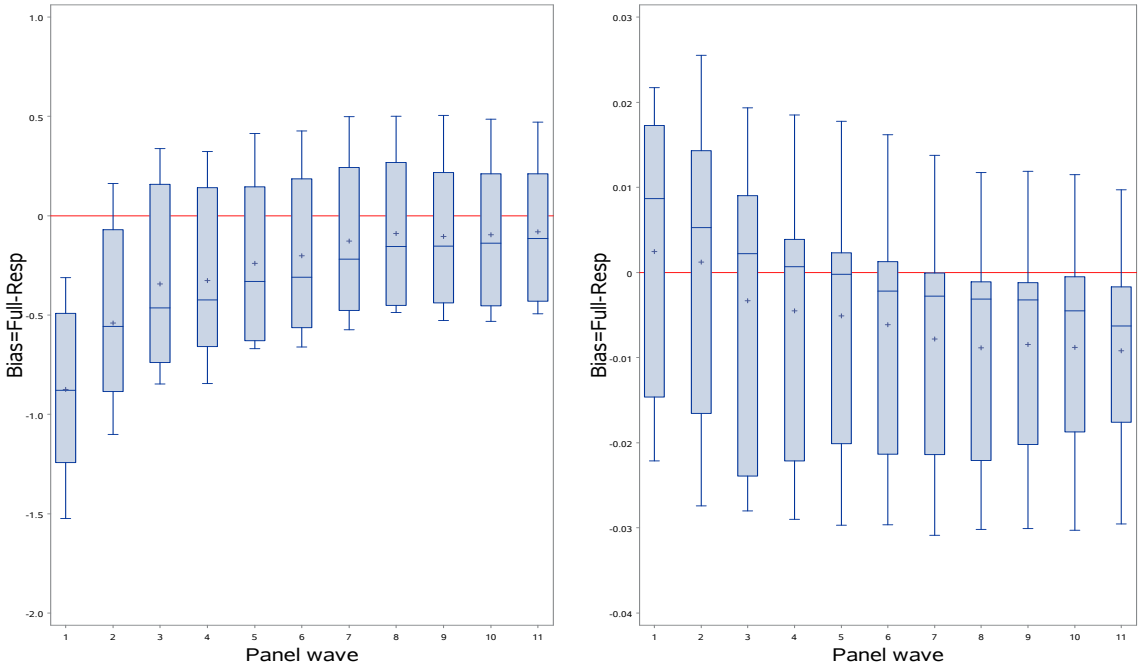


Figure 54: Box plots for the difference of the cross-sectional ordered logit estimates in the Full and the Resp samples, for 100 Monte Carlo replications of the non-response experiment.

(a) Box plot for intercept 9.

(b) Box plot for age.



(c) Box plot for age squared.

(d) Box plot for years of education.

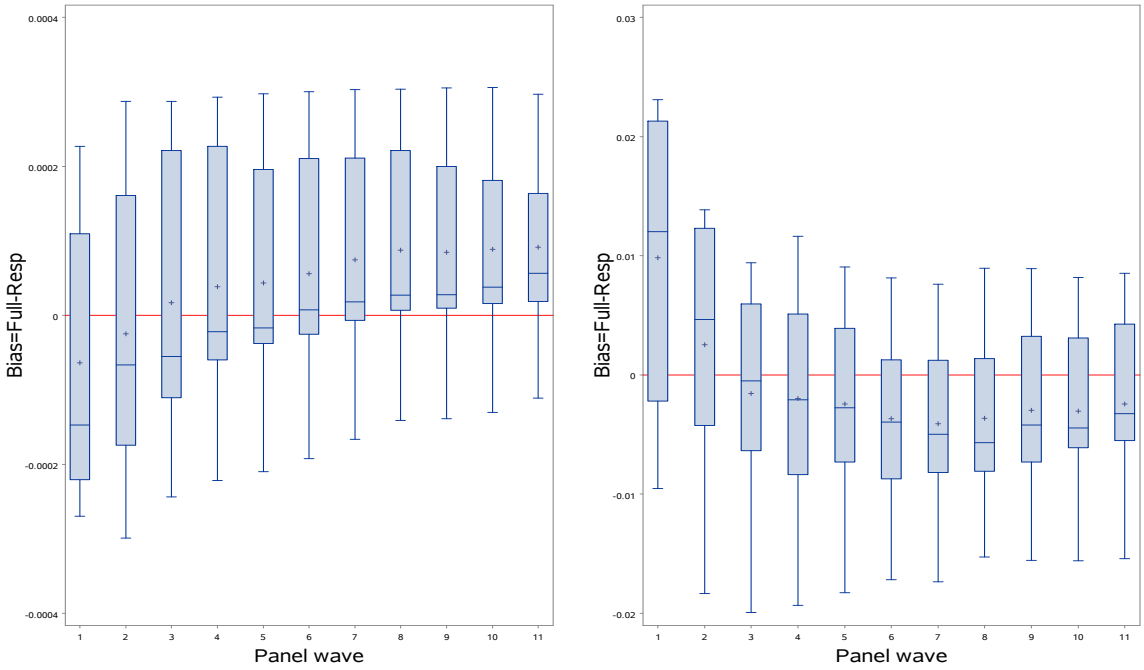
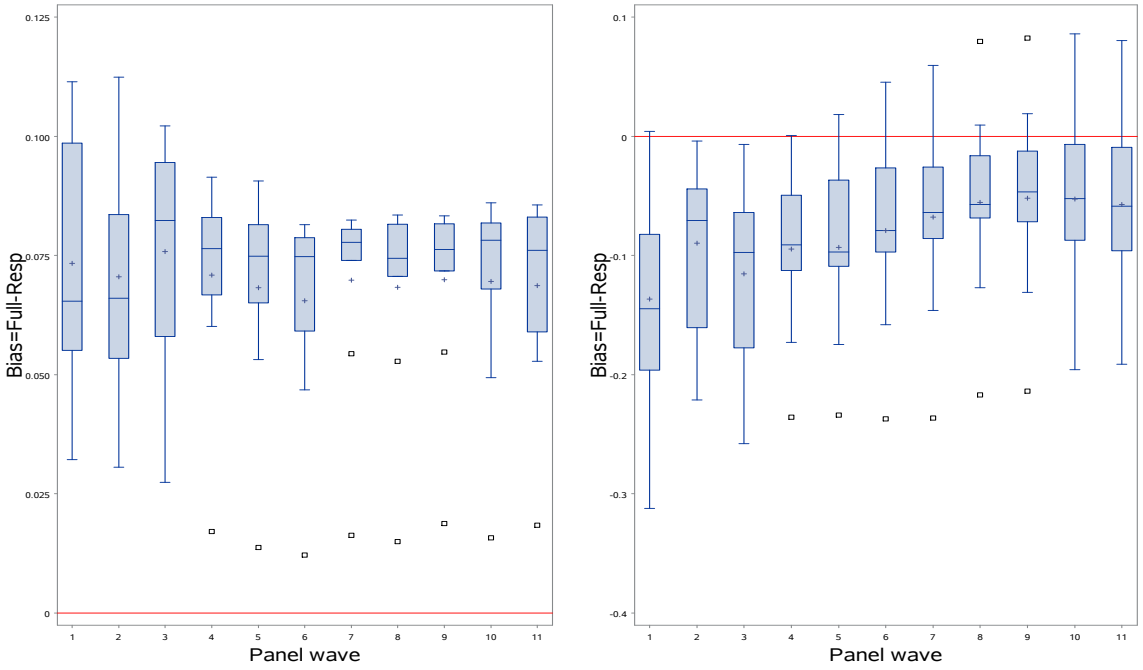


Figure 55: Box plots for the difference of the cross-sectional ordered logit estimates in the Full and the Resp samples, for 100 Monte Carlo replications of the non-response experiment.

(a) Box plot for male.

(b) Box plot for single.



(c) Box plot for widowed.

(d) Box plot for divorced.

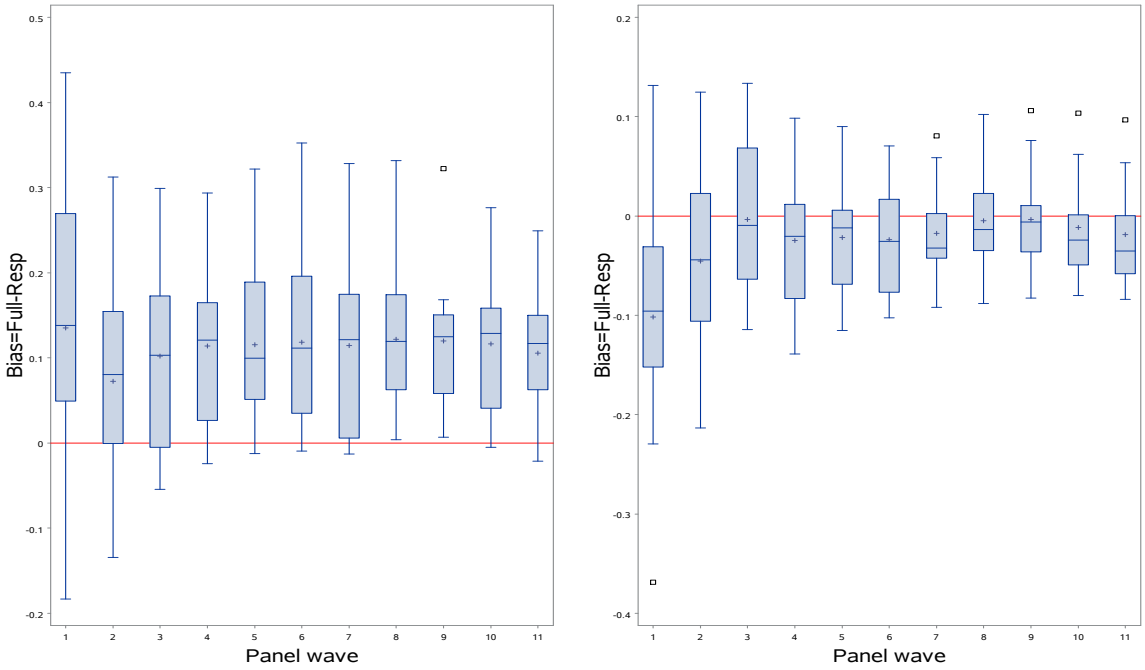
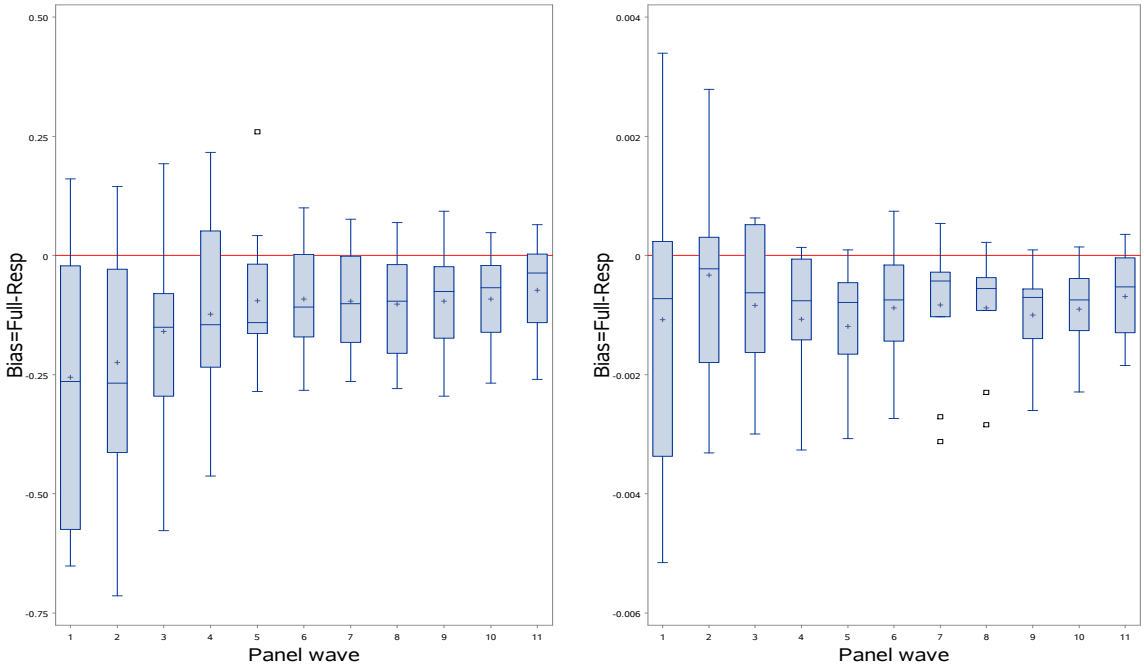


Figure 56: Box plots for the difference of the cross-sectional ordered logit estimates in the Full and the Resp samples, for 100 Monte Carlo replications of the non-response experiment.

(a) Box plot for separated.

(b) Box plot for doctor visits.



(c) Box plot for hospital stays.

(d) Box plot for household income.

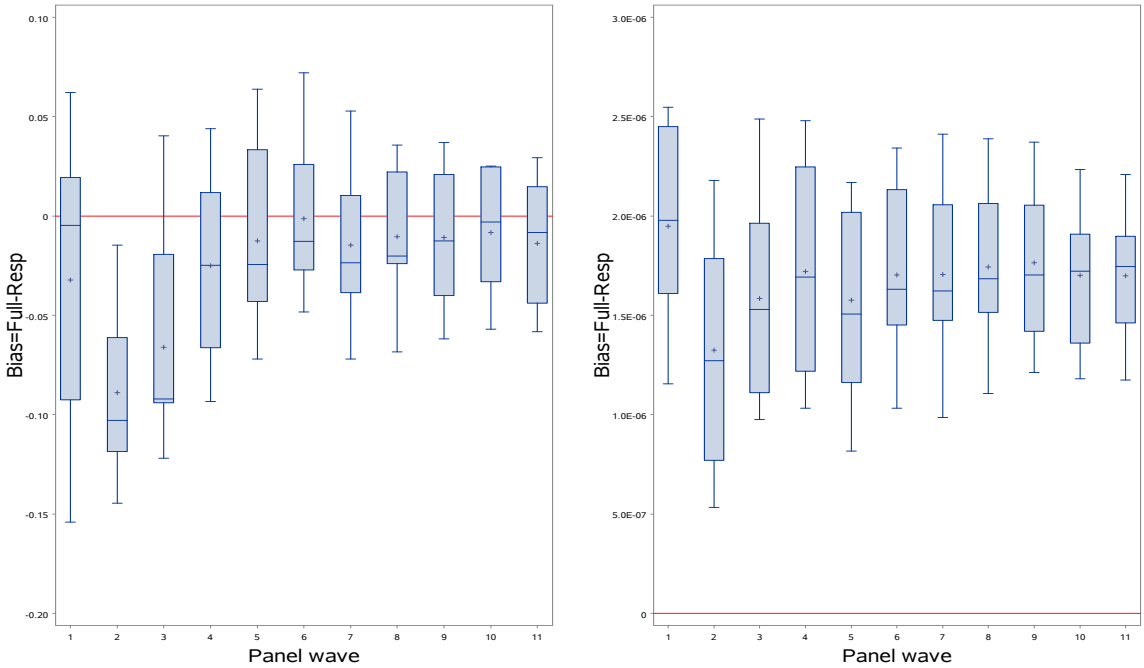
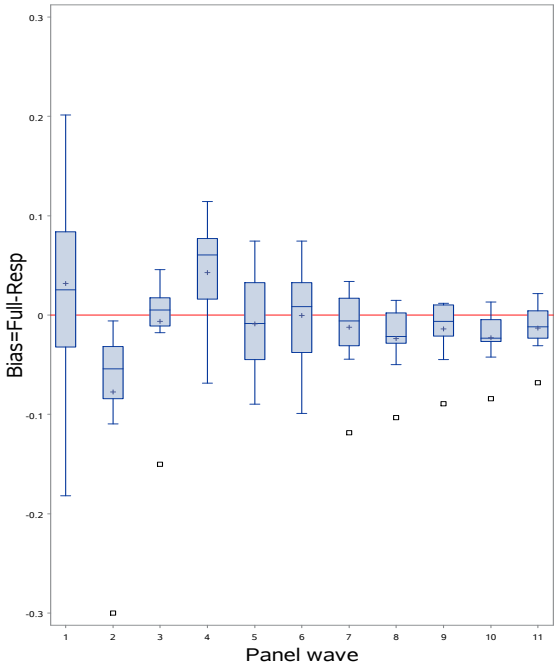
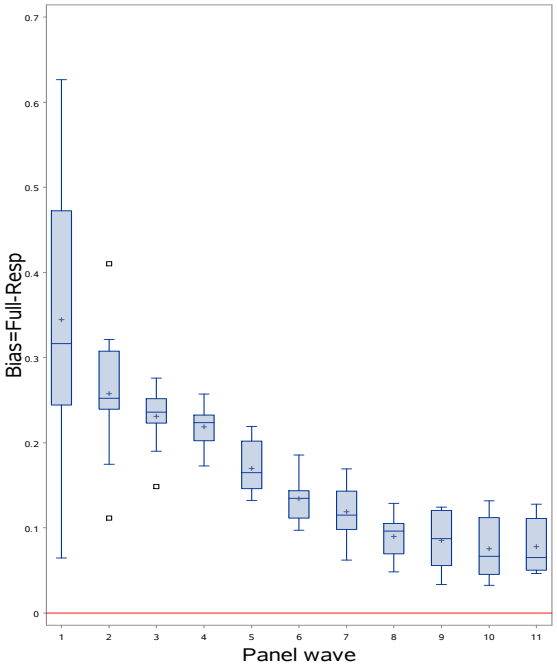


Figure 57: Box plots for the difference of the cross-sectional ordered logit estimates in the Full and the Resp samples, for 100 Monte Carlo replications of the non-response experiment.

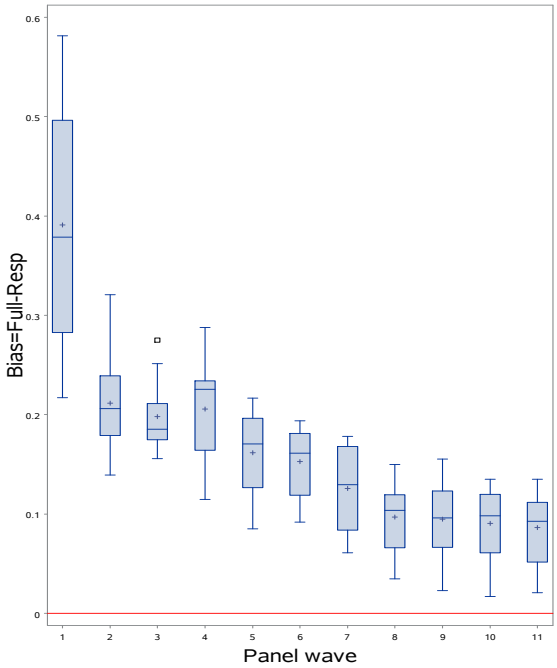
(a) Box plot for satisfaction with health 5.



(b) Box plot for satisfaction with health 6.



(c) Box plot for satisfaction with health 7.



(d) Box plot for health-8.

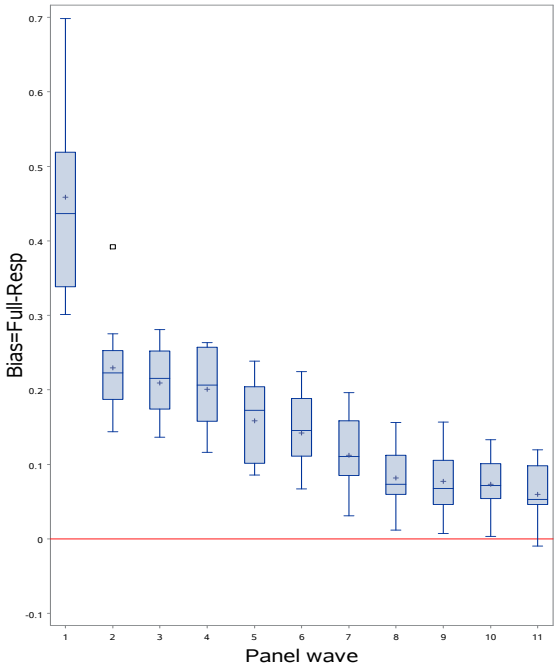
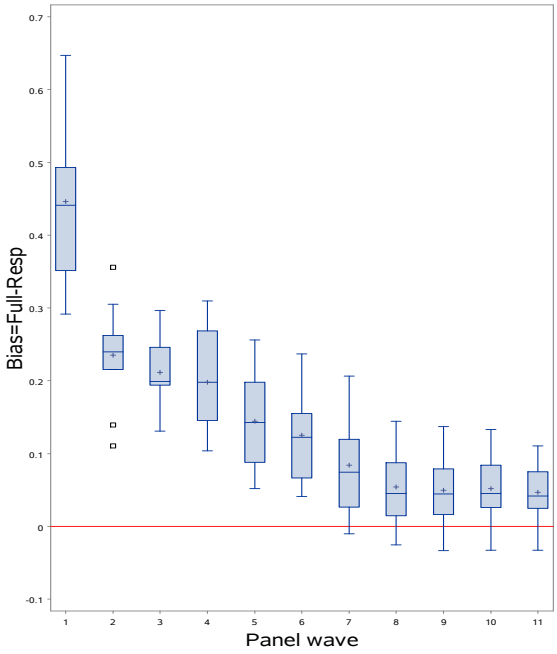


Figure 58: Box plots for the difference of the cross-sectional ordered logit estimates in the Full and the Resp samples, for 100 Monte Carlo replications of the non-response experiment.

(a) Box plot for satisfaction with health 9.



(b) Box plot for satisfaction with health 10.

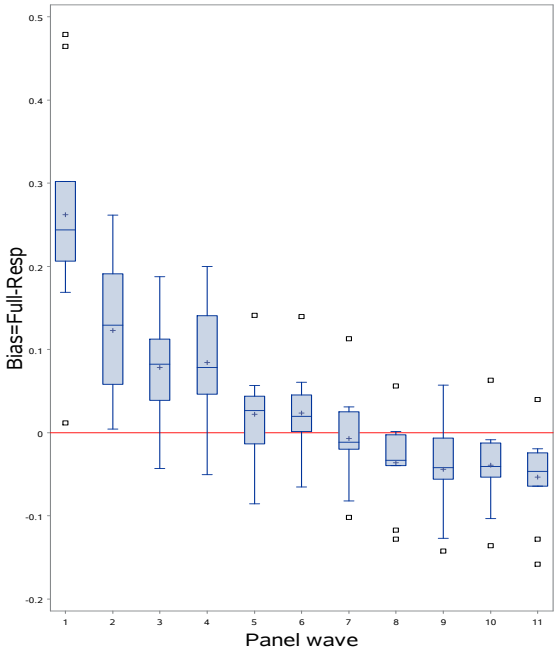


Figure 59: Box plots for the difference of the cross-sectional ordered logit estimates in the Full and the Resp samples, for 100 Monte Carlo replications of the non-response experiment.

4.3.4.2. Longitudinal panel model results

In the previous subsection, we investigated the fade-away effect for the cross-sectional estimates of an ordered logit model using data from wave 1 to 11. In the case of a panel, one might be interested in panel estimates. Therefore, here we use a linear panel model estimator with RE. In order to investigate the fade-away effect for the RE model estimator, we proceed as follows:

There is a different length of the panel waves which enter the estimator. We compute the bias of the panel estimates for the start of the panel based on the first two waves. This consists of the first two panel waves of the SOEP for the year 2000 to 2001. We denote this by length 1. We then add the survey year 2002 (wave 3) to length 1 of the survey year 2000 and 2001 and find the biases for the panel estimates based on the three year longitudinal length. We denote this by length 2. Similarly, we extend the longitudinal lengths up to length 10 when data on all 11 waves of the selected SOEP database from the survey year 2000 to 2010 enter the estimator. In

the cross-sectional regression, the database changes from wave to wave, while in the case of a panel model estimator it changes from length to length.

To see the fade-away effect of the panel model estimator, we plot the biases of the various panel estimates against different panel lengths in Figure 60. For this Table 78 and Table 79 in Section B.2 of Appendix B, provide the estimation results of the regression coefficients for the Full and the Resp-Samples. For the size of the non-response bias of the panel estimates, we compute the bias which is described in Table 80 in Section B.2 of Appendix B. It can be seen from Figure 60, that there are some small and some large initial non-response biases for the different slope parameters of the covariates in longitudinal length 1. However, as the longitudinal lengths increases the corresponding initial biases diminishes in the subsequent lengths. Moreover, it can be seen that the biases don't reduce further after some lengths and remain very stable over the rest of the lengths (length 7 to 10). For example, the effect of intercept (colored blue) decreases from -0.33 to -0.10 at length 7 and then it remains stable for the rest of the lengths. Similarly, the bias of the coefficient health satisfaction fades-away as long as the longitudinal lengths are increased, e.g., the effect of health satisfaction 6 (marked with letter "H6") in the initial length 1 is 0.16 which reduces to 0.01 at length 7 and then it remains permanent in the follow up longer panel lengths.

Similar to cross-sectional estimates there is no fade-away phenomenon present for the panel estimation of the effect of the age (with letter "A"), age squared (with letter "As"), years of education (with letter "E") and for household income (with letter "Hi"). This is because for these coefficients the initial non-response biases are very small. The effect of gender (colored yellow marked with letter "M") is almost stable over all the longitudinal lengths. This is because the variable gender is constant over time and so there is no fade-away effect for gender. Finally, the effect of the different categories of marital status fade-away except for widowed (colored black), which is increasing.

Furthermore, by comparing the speed of the fade-away effect of the panel estimates with the cross-sectional estimates (for detail see Figure 51 to Figure 53 in subsection 4.3.4.1), we notice a much smaller fade-away effect for the longitudinal panel model estimates. This is due to the inclusion of the data from the first panel waves into the estimator. Although, the longitudinal estimators seem to be more efficient because of the use of a larger database they are prone to be affected by biased data from the first

panel waves. To overcome this dilemma, it might be useful to discard observations from the first panel waves. However, the topic of discarding observations from the first panel waves is not discussed in this thesis and is, therefore, the topic for future work.

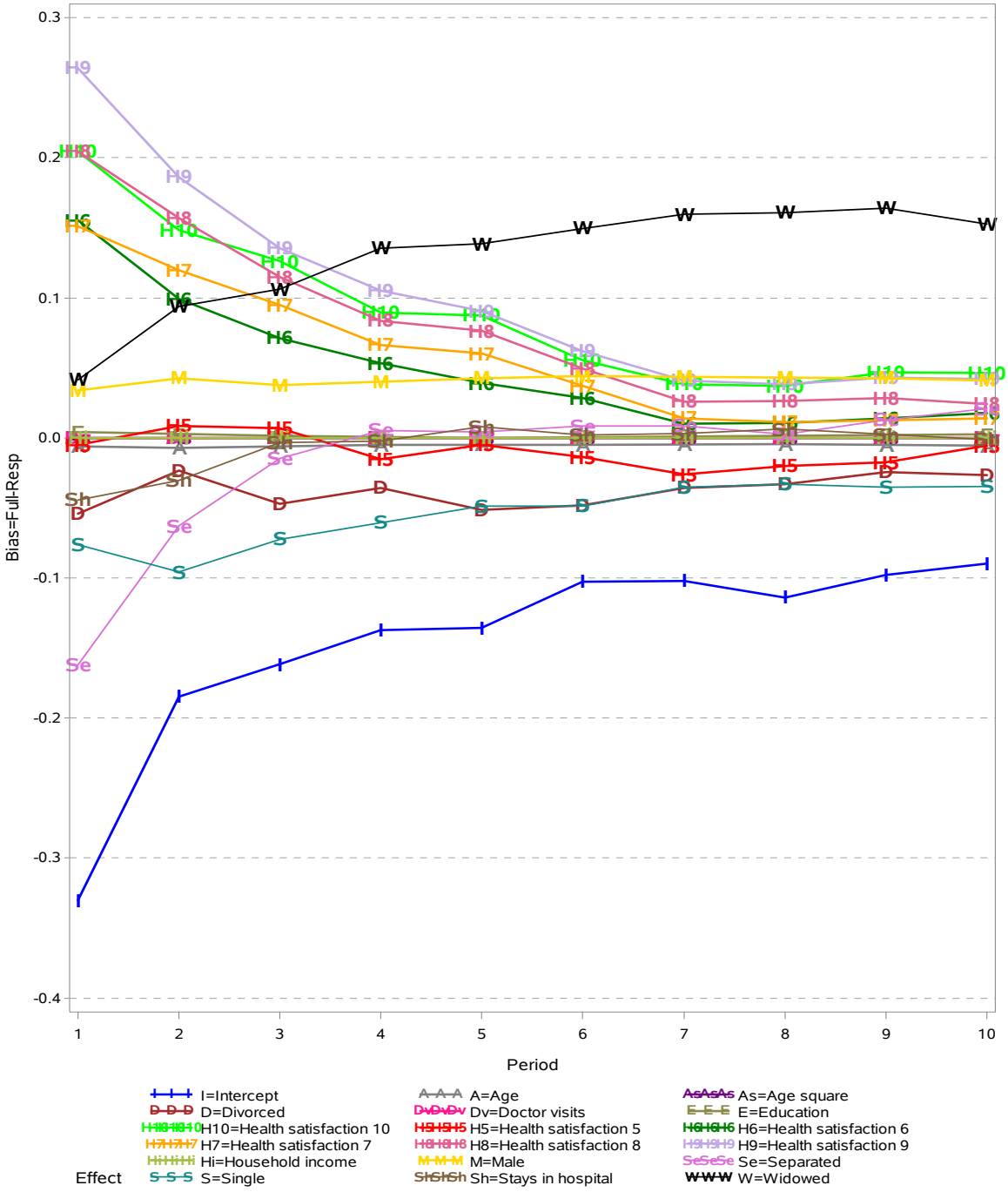


Figure 60: Fade-away of bias of the RE model estimator, with SOEP data and the artificial initial non-response.

Note: The vertical axis displays the bias of the estimates, while the horizontal axis displays the length of the panel. Number of Monte Carlo replications $R = 100$ times. Initial non-response rate is 35%. The alphabets on the graph as highlighted in different colors shows the biases of the estimates.

CHAPTER 5

Conclusion and future research directions

Summary in English:

This chapter provides a short summary of the main findings of this thesis and outlines some research directions for future work. The main findings of this research are summarised in Section 5.1, which consists of the simulation results of Chapter 3 and the empirical results of Chapter 4. Some possible research directions for future work are then addressed in Section 5.2.

5.1. Main findings

This thesis aims to investigate the fade-away effect of the initial non-response bias of the regression model estimators in panel surveys. Specifically, we checked the size of the fade-away effect of the regression estimators: using simulation data and as well as real data from the SOEP. The results of the proposed estimators under simulation setting were set in Part II and the empirical results were presented in Part III.

In simulation Part II, the size of the fade-away effect has been checked for different regression model estimators. These include linear models for cross-sectional data and as well as panel data. For the cross-sectional data, the fade-away effect of the OLS

estimator has been simulated for different stabilities of the covariates and the residual term. Regression weighting procedures are commonly used methods in surveys to reduce the impact of non-response bias on estimates. Therefore, we used different realistic and unrealistic weighting scenarios of the IPW estimator and compared the estimation results with the un-weighted OLS estimator. Concerning the bias correction methods our results indicate that weighting can really help in reducing the impact of non-response/attrition when we use the correct weighting model, while in other cases where we use covariates as a weighting variable the IPW estimator performs very poor with larger bias/variance than the bias/variance of the OLS estimator. However, what is more interesting, is despite using the true response as a weighting variable the bias of the estimator is not completely reduced but still, a bias of 10% is always present in the first wave of the panel. Hence weighting guarantees only consistent parameter estimates.

However, one disadvantage of the cross-sectional OLS/IPW estimators is that it doesn't control the individual unobserved heterogeneity that varies across cross-sections but is constant over time and thus ignores the panel structure of the data. These unobserved time-invariant variables capture all the unobserved time-invariant factors that affect the outcome variable. Ignoring such factors from the model may sometimes result in a heterogeneity bias. Therefore, this gap is covered by using different panel data model estimators. These panel model estimators are: (i) the pooled OLS estimator (Pooled); (ii) the random effects (RE) model estimator (Ranone); (iii) the fixed effects (FE) Within estimator (Fixone). Accordingly, the fade-away effect of the different panel model estimators has been conducted for different stabilities of the covariates and the error term. What is most interesting in comparing the fade-away results of these panel estimators is the fade-away effect of the Within estimator. This is because if non-response is based on the permanent components of the covariates and the error terms then the distorting effect of initial non-response melts down to zero at the second wave when the difference estimator is used. While earlier in Section 3.3 of Chapter 3, we investigated the fade-away effect of the initial non-response bias for the cross-sectional OLS/IPW estimators for different stabilities of the covariates and the error terms. We showed that if the size of a permanent component of the covariates and error terms is large then their distribution stays stable and the distorting effects of initial non-response remain permanent, while this doesn't hold for the stability of transient component which

swings into a steady-state distribution. Since the Within estimator is based on the OLS regression of individual changes, the effect of the persistent components is canceled out by differences at the individual level. Therefore, if non-response is based on the persistent components then under the Within estimator such distorting effect is annihilated by taking differences at the individual level. While the fade-away effect of the RE model estimator is always between the fade-away effect of the Within estimator and the pooled OLS estimator. This holds for all the stabilities of the covariates and the error term. Therefore, the use of the FE Within estimator plays an important role in reducing the effect of non-response based on the permanent components, and is therefore robust against non-response based on the permanent components. This also holds for large transient components of the covariates and the error terms which swing into the steady-state distribution of the Markov chain, thus showing a fade-away behaviour.

The last section (Section 3.7) of Chapter 3 is devoted to the estimation of the non-linear ordered logit model. Here we compared the un-weighted estimates of ordered logit model with several weighting approaches. This consists of realistic and unrealistic scenarios. The realistic scenario consists of two cases: in the first case (WOL 1), we correct for both the initial non-response and attrition through weighting, while in the second case (WOL 2) we control only for the initial non-response through weighting but we don't control for attrition. Under this aspect, the initial weights are constructed by using the information on the covariate $X_{i,t=1}$ which is supposed to be known for the respondents and the non-respondents. Here we take $X_{i,t=1}$ as a predictor for the unknown value of $D_{i,t=1}$. The weights in later panel waves are updated by using the information of the lagged dependent variable $D_{i,t-1}$ as an explanatory variable for attrition. In the unrealistic case (WOL 3), we used the true information on $D_{i,1}$ for the construction of non-response weighting. Similarly, for the estimation of attrition weights, we used the information on lagged dependent variable $D_{i,t-1}$ in the attrition model (similar to attrition weights in the realistic case (WOL 1)).

By comparing the fade-away effect of the various estimators we see a considerable variation in the speed of the fade-away effect. However, the strength of the fade-away effect of the estimators is fast in later panel waves without any corrections by weighting procedures. This is because under the realistic cases we used a wrong weighting model for the probability of response which makes the estimates even more

biased compared to doing nothing. In the unrealistic case with full information on the observed response variable $D_{i,1}$ performs best in wave 1. However, such gain from weighting is completely disappeared quite soon after wave 1 depending on the size of permanent components.

In the empirical part (Part III) of this thesis, the performance of the fade-away effect of the various regression estimators was assessed by using data from the SOEP. The empirical part is further divided into two sections which include the estimation of wage model in Section 4.2 and a life satisfaction score model in Section 4.3.

In Section 4.2, the study aims to examine the effect of initial non-response on the estimation of a wage equation by using data from the first ten panel waves of the SOEP starting from the year 1984 to 1993. To check the fade-away effect we have used three different model settings: (i) the random effects (RE) model for wages with and without lagged dependent variable; (ii) the random effects (RE) model with auto-correlated errors for wages; (iii) the fixed effects (FE) wage model. In order to demonstrate the fade-away effect we gradually extend the database from one to ten panel waves. This covers the period 1984 to 1993 of the SOEP. The analysis is performed for the Full-Sample (containing respondents and non-respondents), and for the Resp-Sample (only respondents). Hence, the possible effect of initial non-response bias is displayed by comparing the Full-Sample with the Resp-Samples estimates.

Our analysis revealed that the estimates of panel models over longer time intervals show only a moderate fade-away effect. This is due to the fact that the corresponding estimators also use the database from the first panel waves. Under this aspect, it might be attractive not to use the data from the first panel waves. The cross-sectional estimators are under this aspect are the extreme solution. They show a statistical fade-away effect. However, the decline of the bias becomes weaker after about six waves. This advocates for a permanent component of the residual which is affected by initial non-response. Then the fade-away effect is only linked to the distribution of the transient component of the error term. The bias of the slope coefficients for the single covariate is quite different with respect to size, sign, and behaviour over panel waves. By their nature, the effect of single is quite stable over time. So in our example, there is no obvious candidate for a fast changing covariate with a high fade-away effect. There seems to be no clear relationship between the stability of the covariate and the size of the bias. Model misspecification may interact with a

non-response bias, as was shown in the case of the lagged hourly wages. Such a bias declined fast in a cross-sectional model but quite slow in an estimator of a panel model.

In addition to the application of the simulation study to SOEP income data in Section 4.2, the SOEP simulation study was replicated with the analysis of satisfaction scores in Section 4.3. Therefore, the next section analyzed the effect of initial-response on the estimation of a model that explains life satisfaction using SOEP data. To examine the fade-away effect, we used two models: (i) the cross-sectional ordered logit model; (ii) the RE model for life satisfaction by using data from the first eleven years of the SOEP Sub-sample F (starting from the year 2000 to 2010). To demonstrate the fade-away effect, we gradually extended the database from one to eleven panel waves. The size of non-response bias is estimated by the difference between the Resp-Samples and the Full-Sample.

Based on our empirical results, we find that age, years of education, household income and doctor visits have only a minor effect on life satisfaction, while marital status, stays in the hospital and health satisfaction have a significant effect on life satisfaction. In comparing the distributional differences of the Full-Sample and the Resp-Sample our results show that the distributional differences between the distribution of the Full-Sample and the Resp-Samples fade-away as the panel goes on. However, the speed of the fade-away effect varies considerably between different slope parameters of the covariates. As the panel estimator uses the information on the first panel waves the fade-away effect is much smaller than the cross-sectional case. This is due to the fact that the corresponding estimators also utilize information from the previous panel waves of the database. Under this aspect, it might be attractive not to use the data from the first panel waves. The cross-sectional estimators under this aspect are the extreme solution. They show a fade-away effect. However, after some waves, the fade-away effect becomes weaker and weaker. This advocates for the existence of a permanent component of the residual which is affected by initial non-response. Then the fade-away effect is only linked to the distribution of the transient component of the error term. The bias of the slope coefficients for the gender (male) covariate is quite different with respect to size, sign and behaviour over panel waves. By their nature, these covariates are quite stable over time. So in our example, there are estimates of the thresholds and estimates of the different categories of health satisfaction for the fast changing covariates with a high fade-away

effect.

5.2. Future research

The research on the fade-away effect is a very interesting topic especially in the context of regression analysis, and therefore deserves further attention for future research work. Besides, our main findings in Chapter 3 and Chapter 4 indicated that the area is still open for further research and needs much attention for further theoretical and practical research. Returning to the main research issues/topics that could be the topics for future research, we highlight some theoretical and empirical topics as follows:

Starting from the simulation study in Chapter 3, where the aim of the simulation study to verify the approximate results of Alho (2015) and demonstrate the fade-away effect in longer time periods. This covers the performance of the cross-sectional OLS estimators. We then correct for the bias in the cross-sectional OLS estimates through IPW, and compare both weighted with un-weighted OLS results. We also discussed the behaviour different linear model estimators over different longitudinal lengths: the pooled OLS estimator, the RE model estimator, and the FE Within estimator. However, we observed that the speed of the fade-away effect for panel estimates was smaller than the cross-section OLS/IPW estimators. This is due to the inclusion of the data from the first panel waves into the panel estimates. Therefore, for future work, it could be also worthwhile to use different longitudinal lengths which exlude data from first panel waves. Also, the use of IPW in the case of panel estimators is not investigated in this thesis and therefore will be the topic for future research, see Rendtel and Harms (2009).

In the first part of Chapter 4, we checked the size of the fade-away effect of the initial non-response bias of the linear regression estimators using SOEP income data. Further, there is no attrition scenario used in the SOEP simulation approach. Therefore, it would also be worthwhile to extend the SOEP simulation study to different attrition scenarios and check the size of the fade-away effect of the estimators. This includes the behaviour of cross-sectional OLS estimator and panel model estimators. In addition to the fade-away effect of the un-weighted estimators under different attrition scenarios, the use of alternative estimators, on the other

hand, can also be very useful in reducing the distorting effect of initial non-response bias. For example, the use of IPW or one can use information from the steady-state distribution of the Markov chain. Therefore, it would be interesting to check the fade-away effect of the IPW estimators, and the comparison of the weighted with un-weighted estimators in the presence of attrition. Further, extension of the work is to repeat the SOEP simulation for the analysis of income data (transition between income quintiles). Finally, it would be also very interesting to analyze the fade-away effect of the estimators in a design-based setting. This consists of the correction of design-based estimates by using calibration estimations, see e.g., Estevao and Särndal (2006) for general use of sample information for calibration, and Rendtel and Harms (2009) for calibration in panel surveys.

Similarly, in the second part of Chapter 4, we examined the fade-away effect of the cross-sectional ordered logit model and a linear panel model with RE using SOEP life satisfaction data. In this SOEP based simulation approach attrition was ignored. However, panel surveys are also affected by panel attrition, which occurs after the first panel wave. Therefore, it would be interesting to visualize the fade-away effect of the above estimators in the presence of panel attrition. Bias correction methods, such as IPW and imputation can be used as well to cope with non-response and attrition biases.

Zusammenfassung in deutscher Sprache:

Dieses Kapitel bietet eine kurze Zusammenfassung der wichtigsten Ergebnisse dieser Arbeit und skizziert einige Möglichkeiten für die Ausrichtung zukünftiger Forschung. Die wichtigsten Ergebnisse dieser Arbeit sind in Abschnitt 5.3 zusammengefasst. Diese bestehen aus den Simulationsergebnissen aus Kapitel 3 und den Ergebnissen der empirischen Untersuchung aus Kapitel 4. Mögliche Ausrichtungen zukünftiger Forschung werden dann in Abschnitt 5.4 behandelt.

5.3. Wichtigste Ergebnisse

Diese Arbeit zielt darauf ab, den Abschwächungseffekt der Verzerrung von Regressionsschätzern durch anfängliche Nichtbeantwortung in Panelbefragungen zu untersuchen. Konkret haben wir die Größe des Fade-Away Effekts der Regressionsschätzer anhand von Simulationsdaten und realen Daten aus dem SOEP überprüft. Die Ergebnisse der vorgeschlagenen Schätzer unter Simulationsbedingungen wurden in Teil II besprochen und die empirischen Ergebnisse wurden in Teil III vorgestellt.

In der Simulation Part II wurde die Größe des Fade-Away Effekts für verschiedene Regressionsmodellschätzer überprüft. Dazu gehören lineare Modelle für Querschnittsdaten und Paneldaten. Für die Querschnittsdaten wurde der Abschwächungseffekt des Kleinste-Quadrate-Schätzers für verschiedene Stabilitäten der Kovariaten und des Residualterms simuliert. Gewichtete Regressionsmethoden sind gängige Methoden für Umfragedaten, um die Auswirkungen von Nonresponse Bias auf Schätzungen zu reduzieren. Daher haben wir verschiedene realistische und unrealistische Gewichtungsszenarien des IPW-Schätzers verwendet und die Schätzergebnisse mit dem ungewichteten OLS-Schätzer verglichen. In Bezug auf die Korrekturmethoden für Verzerrungen, deuten unsere Ergebnisse darauf hin, dass Gewichtung wirklich dazu beitragen kann, die Auswirkungen von Antwortausfallverzerrung/Panelmortalität zu reduzieren, wenn wir das richtige Gewichtungsmodell verwenden. In anderen Fällen, in denen wir Kovariate als Gewichtungsvariable verwenden, liefert der IPW schätzer sehr schlechte Resultate mit einer größeren Verzerrung/Varianz als die Verzerrung/Varianz des OLS-schätzers. Interessant ist jedoch, dass trotz der Verwendung der wahren Antwort als Gewichtungsvariable die Verzerrung des Schätzers nicht vollständig verschwunden ist. In der ersten Welle des Panels tritt stets eine verzerrung von 10% auf. Die Gewichtung garantiert daher nur konsistente Parameterschätzungen, nicht aber unverzerrte. Die Querschnitts-OLS/IPW-Schätzer kontrollieren jedoch nicht für die individuelle unbeobachtete Heterogenität, die über die Querschnitte variiert, aber im laufe der zeit konstant ist und ignorieren somit die Panelstruktur der Daten. Diese unbeobachteten zeitinvarianten variablen erfassen alle unbeobachteten zeitinvarianten Faktoren, die die Ergebnisvariable beeinflussen. Werden solche Faktoren durch das Modell ignoriert, kann das zu einer Heterogenitätsverzerrung führen. Um mit diesem Problem umzugehen, werden verschiedene Schätzmethoden für Paneldaten eingesetzt. Diese Panelschätzer sind: (i) der gepoolte OLS-Schätzer

(Gepoolt); (ii) das Paneldatenmodell (Ranone) mit zufälligen Effekten (RE); (iii) der Within-Schätzer (Fixone) mit festen Effekten (FE). Dementsprechend wurde der Abschwächungseffekt der verschiedenen Panelmodellschätzer für unterschiedliche Stabilitäten der Kovariaten und des Fehlerterms durchgeführt. Am interessantesten beim Vergleich der Ergebnisse dieser Panelschätzer ist der Fade-Away Effekt des Within-Schätzers. Dies liegt daran, dass bei Anwendung des Within-Schätzers der verzerrende Effekt des anfänglichen Antwortausfalls bei der zweiten Welle auf null fällt, wenn der Differenzschätzer verwendet wird.

Die Geschwindigkeit des Abschwächungseffekts hängt von der Größe der permanenten Komponente und der vorübergehenden Komponente ab. Wenn der Wert einer permanenten Komponente groß ist, bleibt ihre Verteilung stabil und die verzerrenden Auswirkungen des anfänglichen Antwortausfalls bleiben dauerhaft. Dies gilt nicht für die Stabilität der vorübergehenden Komponente, die in eine steady-state Verteilung übergeht. Da der Within-Schätzer auf der OLS-Regression einzelner Änderungen basiert, wird die Wirkung der persistenten Komponenten durch unterscheide auf der individueller Ebene aufgehoben. Wenn also die Nicht-Antwort auf den persistenten Komponenten basiert, wird ein solcher verzerrender Effekt unter Verwendung des Within-Schätzers durch die Differenzenbildung auf individueller Ebene zunichte gemacht. Der Fade-Away Effekt des RE-Schätzers liegt immer zwischen dem Fade-Away Effekt des Within-Schätzers und des gepoolten OLS-Schätzers. Dies gilt für alle Stabilitäten der Kovariaten und des Fehlerterms. Daher spielt die Verwendung des FE Within-Schätzers eine wichtige Rolle bei der Verringerung der Auswirkung von Antwortausfall auf Grundlage permanenter Komponenten. Damit ist der Schätzer robust gegen Ausfälle auf Basis persistenter Komponenten.

Der letzte Abschnitt (Abschnitt 3.7)von Kapitel 3 beschäftigt sich mit der Schätzung eines nichtlinearen geordneten logistischen Modells. Wir vergleichen die ungewichteten Schätzergebnisse des geordneten logistischen Modells mit mehreren gewichteten Ansätzen. Dieser Vergleich basiert sowohl auf realistischen, als auch auf unrealistischen Szenarien. Das realistische Szenario besteht aus zwei Fällen: Im ersten Fall (WOL 1), korrigieren wir für anfängliche Nicht-Antwort und Attrition mittels Gewichtung, während im zweiten Fall (WOL 2) durch Gewichtung nur für anfängliche Nicht-Antwort, aber nicht für Attrition kontrolliert wird. Unter diesem Gesichtspunkt, werden die anfänglichen Gewichte mit Hilfe der Kovariate $X_{i,t=1}$ konstruiert, welche für Befragte und Nichtbefragte bekannt sein soll. Wir betrachten

hierbei $X_{i,t=1}$ als Prädiktor für das unbekannte $D_{i,t=1}$. Die Gewichte in den späteren Panelwellen werden durch die Verwendung von verzögerten Werten der abhängigen Variable $D_{i,t-1}$ als erklärende Variable für Attrition aktualisiert. Im unrealistischen Fall (WOL 3) wurde die wahre Information über $D_{i,1}$ zur Konstruktion von Antwortausfallgewichten herangezogen. Ganz ähnlich verhält es sich mit der Schätzung von Attritionsgewichten. Es werden die Werte der verzögerten, abhängigen Variable $D_{i,t-1}$ im Attritionsmodell verwendet (vergleichbar zu den Attritionsgewichten im realistischen Fall (WOL 1)).

Durch den Vergleich des Fade-Away Effekts der verschiedenen Schätzer können wir beträchtliche Unterschiede in der Geschwindigkeit des Fade-Away Effekts beobachten. Allerdings ist die Stärke des des Effekts der Schätzer in späteren Panelwellen ohne Korrekturen durch Gewichtung groß. Der Grund dafür liegt in den realistischen Fällen in der Verwendung eines falschen Gewichtungsmodells für die Wahrscheinlichkeit zu antworten, was die Schätzwerte noch mehr verzerrt, als nichts zu tun. Der unrealistische Fall mit vollständiger Information über die beobachtete abhängige Variable $D_{i,1}$ schneidet in Welle eins am besten ab. Allerdings verschwindet ein solcher Vorteil durch Gewichtung vollständig relativ schnell nach der ersten Welle in Abhängigkeit von der Größenordnung der permanenten Komponente.

Im empirischen teil (Teil III) dieser Arbeit wurde die Leistung des Abschwächungseffekts der verschiedenen Regressionsschätzer anhand von Daten aus dem SOEP bewertet. Der empirische teil ist in zwei Abschnitte unterteilt, die die Schätzung eines Lohnmodells in Abschnitt 4.2 und eines Score-Modells der Lebenszufriedenheit in Abschnitt 4.3 beinhalten. In Abschnitt 4.2 zielt die Studie darauf ab, die Auswirkungen des anfänglichen Antwortausfalls auf die Schätzung einer Lohngleichung zu untersuchen, indem Daten aus den ersten zehn Panelwellen des SOEP aus den Jahren 1984 bis 1993 verwendet werden. Um den Abschwächungseffekt zu überprüfen, haben wir drei verschiedene Modelle verwendet: (i) das Panelmodell mit zufälligen Effekten (RE) für Löhne mit und ohne verzögerte abhängige Variable; (ii) das Panelmodell mit zufälligen Effekten (RE) mit autokorrelierten Fehlern für Löhne; (iii) das Panelmodell mit festen Effekten (FE). Um den Fade-Away Effekt zu demonstrieren, erweitern wir die Datenbasis schrittweise von einer auf zehn Panelwellen. Dies betrifft den Zeitraum 1984 bis 1993 des SOEP. Die Analyse wird für die Gesamtstichprobe (mit Befragten und Nichtbefragten) und für die Resp-Stichprobe (nur Befragte) durchgeführt. Daher wird der mögliche Effekt der anfänglichen Antwor-

tausfallsverzerrung durch den Vergleich der Gesamtstichprobe mit den Schätzungen der Resp-Stichproben angezeigt.

Unsere Analyse ergab, dass die Schätzungen von Panelmodellen über längere Zeiträume nur einen moderaten Abschwächungseffekt zeigen. Dies liegt daran, dass die entsprechenden Schätzer auch die Datenbasis aus den ersten Panelwellen nutzen. Unter diesem Aspekt könnte es sinnvoll sein, die Daten der ersten Panelwellen nicht zu verwenden. Die Querschnittsschätzer sind unter dieser Betrachtung sind die extreme Lösung. Sie zeigen einen statistischen Abschwächungseffekt. Der Rückgang der Verzerrung wird jedoch nach etwa sechs Wellen schwächer. Dieser spricht für eine dauerhafte Komponente der Residuen, die durch anfänglichen Antwortausfall beeinflusst wurde. Dann ist der Fade-Away Effekt nur noch mit der Verteilung der vorübergehenden Komponente des Fehlerterms verknüpft. Die verzerrung der Steigungskoeffizienten für einzelne Kovariate ist in Bezug auf Größe, Vorzeichen und Verhalten über Panelwellen sehr unterschiedlich. Naturgemäß sind sie im Laufe der Zeit recht stabil. In unserem Beispiel gibt es keinen offensichtlichen Kandidaten für eine sich schnell verändernde Kovariate mit einem hohen Abschwächungseffekt. Es scheint keinen klaren Zusammenhang zwischen der Stabilität der Kovariate und der Größe der Verzerrung zu geben. Modellfehlspezifikationen können mit einer durch Antwortausfall verursachten Verzerrung interagieren, wie sich bei den verzögerten Stundenlöhnen zeigte. Eine solche Verzerrung nahm in einem Querschnittsmodell schnell ab, in einem Panelmodell jedoch recht langsam.

Zusätzlich zur Anwendung der Simulationsstudie auf die SOEP-Einkommensdaten in Abschnitt 4.2 wurde die SOEP Simulationsstudie mit der Analyse von Zufriedenheitswerten in Abschnitt 4.3 repliziert. Aus diesem Grund wurde im nächsten Abschnitt der Einfluss von anfänglichem Antwortausfall auf die Schätzung eines Modells analysiert, das die Lebenszufriedenheit anhand von SOEP-Daten erklärt. Um den Fade-Away Effekt zu untersuchen, haben wir zwei Modelle verwendet: (i) das Querschnittsmodell für geordnete Logitregression; (ii) das RE-Modell für Lebenszufriedenheit durch die Verwendung von Daten aus den ersten elf Jahren der SOEP-Unterstichprobe F (aus den Jahren 2000 bis 2010). Um den Abschwächungseffekt zu demonstrieren, haben wir die Datenbasis von einer bis zu elf Panelwellen erweitert. Die Größe der Verzerrung durch Antwortausfall wird durch die Differenz zwischen den Resp-Stichproben und der Gesamtstichprobe geschätzt.

Basierend auf unseren empirischen Ergebnissen stellen wir fest, dass Alter, Bil-

dungsjahre, Haushaltseinkommen und Arztbesuche nur einen geringen Einfluss auf unterschiedliche Lebenszufriedenheit haben, während hingegen Familienstand, Krankenhausaufenthalte und Gesundheitszufriedenheit einen signifikanten Einfluss auf die Lebenszufriedenheit aufweisen. Beim Vergleich der Verteilungsunterschiede zwischen Gesamtstichprobe und Resp-Stichprobe zeigen unsere Ergebnisse, dass die Verteilungsunterschiede zwischen der im Verlauf des Panels verschwinden. Die Geschwindigkeit des Fade-Away Effekts variiert jedoch stark zwischen den verschiedenen Steigungsparametern der Kovariaten. Da der Panelschätzer die Informationen über der ersten Panelwellen verwendet, ist der Abschwächungseffekt viel kleiner als im Querschnittsfall. Dies liegt daran, dass die entsprechenden Schätzer auch Informationen aus den vorherigen Panelwellen der Datenbasis nutzen. Unter diesem Aspekt könnte es sinnvoll sein, die Daten der ersten Panelwellen nicht zu verwenden. Die Querschnittsschätzer sind in dieser Hinsicht die Extremlösung. Sie zeigen einen Abschwächungseffekt. Nach einigen Wellen wird der Fade-Away Effekt jedoch immer schwächer. Dies deutet auf das Vorhandensein einer dauerhaften Komponente der Residuen hin, in welcher die Verteilung von einem anfänglichen Antwortausfall betroffen ist. Dann ist der Fade-Away Effekt nur noch mit der Verteilung der vorübergehenden Komponente des Fehlerterms verknüpft. Die Verzerrung der Steigungskoeffizienten für die Geschlechterkovariate (männlich) ist in Bezug auf Größe, Vorzeichen und Verhalten über Panelwellen sehr unterschiedlich. Von Natur aus sind diese Kovariaten im Laufe der Zeit recht stabil. In unserem Beispiel gibt es also Schätzungen der Schwellenwerte und Schätzungen verschiedener Kategorien der Gesundheitszufriedenheit für die sich schnell verändernden Kovariaten mit hohem Abschwächungseffekt.

5.4. Zukünftige Forschung

Die Erforschung des Fade-Away Effekts ist ein sehr interessantes Thema, insbesondere im Rahmen von Regressionsanalysen und verdient daher weitere Aufmerksamkeit über zukünftige Forschungsarbeit. Außerdem, haben unsere Hauptergebnisse in Kapitel 3 und Kapitel 4 gezeigt, dass das Gebiet noch für weitere Untersuchungen offen ist und braucht viel Aufmerksamkeit für weitere theoretische und praktische Forschung. Im folgenden werden einige theoretische und empirische Themen, welche

sich als Hauptforschungsgegenstände eignen vorgestellt:

Ausgehend von der Simulationsstudie in Kapitel 3, die das Ziel hatte die approximierten Ergebnisse von Alho (2015) zu belegen und den Fade-Away Effekt in längeren Zeiträumen zu belegen. Dies umfasst das Abschneiden des Querschnitts OLS-Schätzers und der Panelschätzer. Wir korrigieren anschließend für die Verzerrung in den Querschnitts OLS-Schätzungen durch IPW und vergleichen die beiden gewichteten und ungewichteten OLS-Ergebnisse. Wir haben auch das Verhalten verschiedener linearer Modellschätzer über verschiedene Zeithorizonte verglichen: der gepoolte OLS Schätzer, der RE Modellschätzer und der FE-Within Schätzer. Wir haben dabei beobachtet, dass die Geschwindigkeit des Fade-Away Effekts für Panelschätzungen geringer war, als die des Querschnitts OLS/ IPW Schätzers. Dies liegt in der Verwendung der Daten der ersten Panelwellen in den Panelschätzungen begründet. Daher wäre es für zukünftige Forschung interessant, verschiedene Zeithorizonte zu betrachten, aber die Daten erster Panelwellen auszuschließen. Außerdem wurde die Verwendung des IPW im Fall eines Panelschätzers nicht in dieser Arbeit untersucht und bleibt deshalb ein Thema für zukünftige Forschung (siehe Rendtel and Harms (2009)).

Im ersten Teil des Kapitels 4, überprüften wir die Größe des Fade-Away Effekts der Verzerrung durch anfänglichen Antwortausfall der linearen Regressionsschätzer unter Verwendung von SOEP-Einkommensdaten. Darüber hinaus gibt es kein Attritionsszenario, das im SOEP simulationsansatz verwendet wird. Daher wäre es sinnvoll, die SOEP-Simulationsstudie auf verschiedene Mortalitätsszenarien auszudehnen und die Größe des Abschwächungseffekts der Schätzer zu überprüfen. Dazu gehört auch das Verhalten von Querschnitts OLS-Schätzern und Panelmodellschätzern. Neben dem Fade-Away Effekt der ungewichteten Schätzer unter verschiedenen Attritionsszenarien kann der Einsatz alternativer Schätzer andererseits auch sehr nützlich sein, um den verzerrenden Effekt des anfänglichen Antwortausfalls zu reduzieren. Beispielsweise sind die Verwendung von IPW oder von Informationen aus der stationären Verteilung der Markovkette denkbar. Daher wäre es interessant, den Abschwächungseffekt der IPW-Schätzer zu untersuchen und den Vergleich zwischen gewichteten und ungewichteten Schätzern bei Panelattrition zu ziehen. Darüber hinaus könnte eine Erweiterung der Arbeit darin bestehen, die SOEP-Simulation für die Analyse von Einkommensdaten (Übergang zwischen Einkommensquintilen) wiederholen. Schließlich wäre es auch sehr interessant, den Abschwächungseffekt

der Schätzer in einem designbasierten Umfeld zu analysieren. Dieses besteht aus der Korrektur designbasierter Schätzungen unter der Verwendung von Kalibrationsschätzungen, siehe zum Beispiel Estevao and Särndal (2006) für allgemeine Verwendung von Stichprobeninformation zur Kalibration und Rendtel and Harms (2009) zu Kalibration in Panelbefragungen.

Ebenso untersuchten wir im zweiten Teil von Kapitel 4 den Abschwächungseffekt des geordneten Logitmodells für Querschnittsdaten und eines linearen Panelmodells mit zufälligen Effekten unter Verwendung von SOEP-Lebenszufriedenheitsdaten. In diesem SOEP-basierten Simulationsansatz wurde die Panelmortalität ignoriert. Die Panelbefragungen werden aber auch von der Panelattrition beeinflusst, die nach der ersten Panelwelle auftritt. Deshalb ist wäre interessant, den Abschwächungseffekt der obigen Schätzer bei Vorhandensein von Panelattrition abzubilden. Verzerrungskorrekturmethoden wie IPW und Imputation können ebenfalls eingesetzt werden, um mit Verzerrungen bei Antwortausfall und Mortalität umzugehen.

Bibliography

Abowd, J., Crepon, B., and Kramarz, F. (2001). Moment estimation with attrition: An application to economic models. *Journal of the American Statistical Association*, 96(456):1223–1231.

Agresti, A. (2010). *Analysis of ordinal categorical data*. New York: Wiley, 2nd edition.

Aitchison, J. and Silvey, S. D. (1957). The generalization of probit analysis to the case of multiple repsonses. *Biometrika*, 44:131–140.

Aldrich, J. H. and Nelson, F. D. (1984). *Linear probability, logit, and probit models*. London: Sage Publications.

Alho, J. (2015). On the fade away phenomenon in follow-up studies. Unpublished manuscript.

Alho, U., Müller, G., Pflieger, V., and Rendtel, U. (2017). The fade-away of an initial bias in longitudinal surveys, Discussion paper Economics FB Wirtschaftswissenschaft FUB 2017/25 . Online available at http://hdl.handle.net/10419/168347.

Allison, P. D. (2000). Multiple imputation for missing data: A Cautionary Tale. *Sociological Methods and Research*, 28(3):301–309.

Allison, P. D. (2001). *Missing data*. Thousand Oaks, CA: Sage.

Altman, E. I. and Kao, D. L. (1991). Examining and modeling corporate bond rating drift, New York University Salomon Center working paper series, S-91-39.

Amemiya, T. (1981). Qualitative response models: A survey. *Journal of Economic Literature*, 19:1483–1536.

Antonakis, J., Bendahan, S., Jacquart, P., and Lalive, R. (2014). Causality and endogeneity: Problems and solutions. In D.V. Day (ed.), The Oxford Handbook of Leadership and Organizations (pp. 93-117). New York: Oxford University Press.

Arellano, M. (2003). *Panel data econometrics*. Oxford University Press: Advanced texts in econometrics.

Armstrong, J. S. and Overton, T. S. (1977). Estimating non-response bias in mail surveys. *Journal of Marketing Research*, 14:396–402.

Bandeen-Roche, K., Miglioretti, D., Zeger, S., and Rathouz, P. (1997). Latent variable regression for multiple discrete outcomes. *Journal of the American Statistical Association*, 92:1375–1386.

Bartels, L. (1993). Messages received: The political impact of media exposure. *American Political Science Review*, 88:267–285.

Behr, A., Bellgardt, E., and Rendtel, U. (2005). Extent and determinants of panel attrition in the European Community Household Panel. *European Sociological Review*, 21:489–512.

Blumen, I., Kogan, M., and Mccarthy, P. (1955). *The industrial mobility of labor as a probability process*. Ithaca, New York: Cornell University Press.

Breen, R., Karlson, K. B., and Holm, A. (2018). Interpreting and understanding logits, probits, and other non-linear probability models. *Annual Review of Sociology*, 44:39–54.

Brick, M. J. (2013). Unit non-response and weighting adjustments: A critical review. *Journal of Official Statistics*, 29(3):329–353.

Briggs, A., Clark, T., Wolstenholme, J., and Clarke, P. (2003). Missing presumed at random: cost-analysis of incomplete data. *Health Economics*, 12:377–392.

Campanelli, P. and O'Muircheartaigh, C. (1999). Interviewers, interviewer continuity, and panel survey non-response. *Quality and Quantity*, 33:59–76.

Carpenter, J. R. and Kenward, M. G. (2013). *Multiple imputation and its application*. Hoboken, NJ: Wiley.

Chambers, R. L. (1996). Robust case-weighting for multipurpose establishment surveys. *Journal of Official Statistics*, 12:3–32.

Daniels, M. J. and Hogan, J. W. (2000). Reparameterizing the pattern mixture model for sensitivity analysis under informative dropout. *Biometrics*, 56:1241–1248.

Das, M. (2004). Instrumental variables estimation of non-parametric models with discrete endogenous regressors. *Journal of Econometrics*, 124:335–361.

Demirtas, H. and Schafer, J. L. (2003). On the performance of random coefficient pattern mixture models for non-ignorable dropout. *Statistics in Medicine*, 22:2553–2575.

Deville, J. C. and Särndal, C. E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87:376–382.

Diggle, P. J., Heagerty, P., Liang, K. Y., and Zeger, S. L. (2002). *Analysis of longitudinal data*. Oxford University Press, 2nd edition.

DiNardo, J., McCrary, H., and Sanbonmatsu, L. (2006). Constructive proposals for dealing with attrition: An empirical example. Ann Arbor: University of Michigan, working paper. Online available at https://pdfs.semanticscholar.org/b431/c3e7112e443eeab310b90016030ed53830f6.pdf.

Durrant, G. B. (2009). Imputation methods for handling item non-response in practice: Methodological issues and recent debates. *International Journal of Social Research Methodology*, 12(4):293–304.

Enders, C. K. (2010). *Applied missing data analysis*. Guilford Press, New York, US.

Estevao, V. M. and Särndal, C. E. (2000). A functional approach to calibration. *Journal of Official Statistics*, 16:379–399.

Estevao, V. M. and Särndal, C. E. (2006). Survey estimates by calibration on complex auxiliary information. *International Statistical Review*, 74(2):127–147.

Ferrer-i-Carbonell , A. and Frijters, P. (2004). The effect of methodology on the determinants of happiness. *The Economic Journal*, 114(497):641–659.

Firth, D. and Bennett, K. E. (1998). Robust models in probability sampling. *Journal of the Royal Statistical Society: Series B*, 60:3–21.

Fitzgerald, J., Gottschalk, P., and Moffitt, R. (1998). An analysis of sample attrition in panel data, The Michigan Panel Study of Income Dynamics. *The Journal of Human Resources*, 33(2):251–299.

Frydman, H. (1984). Maximum likelihood estimation in the mover-stayer model. *Journal of the American Statistical Association*, 79:632–637.

Frydman, H. and Kadam, A. (2002). Estimation in the continuous time mover-stayer model with an application to bond ratings migration, Statistics working papers series. Online available at SSRN: https://ssrn.com/abstract=1293601.

Frydman, H., Kallberg, J. G., and Kao, D. L. (1985). Testing the adequacy of Markov chains and mover-stayer models as representations of credit behavior. *Operations Research*, 33:1203–1214.

Fuller, W. A. (2009). *Sampling statistics*. Wiley, Hoboken.

Fullerton, A. S. (2009). A conceptual framework for ordered logistic regression models. *Sociological Methods and Research*, 38:306–347.

Goodman, L. A. (1961). Statistical methods for the mover-stayer model. *Journal of the American Statistical Association*, 56:841–868.

Grabka, M. M. (2012). Codebook for the $PEQUIV File 1984-2011, CNEF variables with extended income information for the SOEP. Data documentation No. 65. German Institute for Economic Research (DIW), Berlin. Online available at https://www.econstor.eu/bitstream/10419/64634/1/722934114.pdf.

Greene, W. H. (2008). *Econometric analysis*. Englewood Cliffs, Prentice Hall, 6th edition.

Groves, R. M. and Couper, M. P. (1998). *Non-response in household interview surveys*. New York: John Wiley & Sons.

Groves, R. M., Dillman, D. A., Eltinge, J. L., and Little, R. J. A. (2002). *Survey non-response*. New York: John Wiley & Sons.

Grubb, D. and Magee, L. (1988). A variance comparison of OLS and feasible GLS estimators. *Econometric Theory*, 4(2):329–335.

Häggström, O. (2002). *Finite Markov chains and algorithmic applications*. Cambridge University Press.

Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement*, 5:475–492.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, 47:153–161.

Hirano, K., Imbens, G. W., and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189.

Hogan, J. W. and Daniels, M. J. (2008). *Missing data in longitudinal studies*. Boca Raton: Chapman and Hall.

Horowitz, J. L. and Manski, C. F. (1998). Censoring of outcomes and regressors due to survey non-response: Identification and estimation using weights and imputations. *Journal of Econometrics*, 84(1):37–58.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47:663–685.

Hosmer, D. W. and Lemeshow, S. (2000). *Applied logistic regression*. NY: John Wiley & Sons, 2nd edition.

Hox, J. J. and deLeeuw, E. D. (1994). A comparison of non-response in mail, telephone, and face-to-face surveys-applying multilevel modeling to meta-analysis. *Quality and Quantity*, 28:329–344.

Hsiao, C. (1986). *Analysis of panel data*. Cambridge University Press, 2nd edition.

Ibrahim, J. G., Chen, M. H., Lipsitz, S. R., and Herring, A. H. (2005). Missing data methods for generalized linear models. *Journal of the American Statistical Association*, 100(469):332–346.

Isaki, C. and Fuller, W. A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77:89–96.

Junes, T. (2012). Initial wave non-response and panel attrition in the Finnish sub-sample of EU-SILC. Master's thesis, Department of Social Statistics, University of Helsinki, Helsinki. Available at https://helda.helsinki.fi/handle/10138/34131?show=full.

Kalton, G. and Flores-Cervantes, I. (2003). Weighting methods. *Journal of Official Statistics*, 19:81–97.

Kim, J. K. and Park, M. (2010). Calibration estimation in survey sampling. *International Statistical Review*, 78(1):21–39.

Kott, P. S. (2006). Using calibration weighting to adjust for non-response and coverage errors. *Survey Methodology*, 32:133–142.

Kutner, M. H., Nachtsheim, C. J., Neter, J., and Li, W. (2005). *Applied linear statistical models*. McGraw-Hill/Irwin, New York, 5th edition.

Lessler, J. T. and Kalsbeek, W. D. (1992). *Non-sampling error in surveys*. New York: John Wiley.

Li, X., Basu, S., Miller, M. B., Iacono, W. G., and McGue, M. (2011). A rapid generalized least squares model for a genome-wide quantitative trait association analysis in families. *Human Heredity*, 71(1):67–82.

Liao, T. M. (1994). *Interpreting probability models logit, probit, and other generalized linear models*. London: Sage Publications.

Little, R. J. and Rubin, D. B. (2002). *Statistical analysis with missing data*. New Jersey: Wiley, 2nd edition.

Little, R. J. A. (1988). Missing data adjustments in large surveys. *Journal of Business and Economic Statistics*, 6(3):287–296.

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

Little, R. J. A. and Rubin, D. B. (1987). *Statistical analysis with missing data*. John Wiley & Sons, New York.

Little, R. J. A. and Rubin, D. B. (1989). The analysis of social science data with missing values. *Sociological Methods Research*, 18(2-3):292–326.

Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. California: Sage Publications.

Luca, G. D. and Peracchi, F. (2007). A sample selection model for unit and item non-response in cross-sectional surveys (March 2007), CEIS working paper No. 99. Available at SSRN: https://ssrn.com/abstract=967391.

Maddala, G. (1983). *Limited dependent and qualitative variables in econometrics*. Cambridge University Press, Cambridge, UK.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society Series B*, 42:109–142.

McKelvey, R. D. and Zavoina, W. (1975). A statistical model for the analysis of ordinal level dependent variables. *Journal of Mathematical Sociology*, 4:103–120.

Mincer, J. A. (1974). *Schooling, experience and earnings*. Columbia University Press: New York.

Moffitt, R., Fitzgerald, J., and Gottschalk, P. (1999). Sample attrition in panel data: the role of selection on observables. *Annales D'Economie et de Statistique*, 55-56:129–152.

Molenberghs, G., Michiels, B., Kenward, M. G., and Diggle, P. J. (1988). Monotone missing data and pattern mixture models. *Statistica Neerlandica*, 52:153–161.

Montanari, G. E. and Ranalli, M. G. (2005). Non-parametric model calibration estimation in survey sampling. *Journal of the American Statistical Association*, 100:1429–1442.

Nakai, M. and Weiming, K. (2011). Review of methods for handling missing data in longitudinal data analysis. *International Journal of Mathematical Analysis*, 5(1):1–13.

Norris, J. R. (1997). *Markov chains*. Cambridge University Press.

O'Connell, A. A. (2006). *Logistic regression models for ordinal response variables*. Thousand Oaks, CA: Sage.

Olsen, R, J. (2005). The problem of respondent attrition: Survey methodology is key. *Monthly Labor Review*, 128:63–71.

Park, M. and Fuller, W. A. (2005). Towards non-negative regression weights for survey samples. *Survey Methodology*, 31:85–93.

Poulsen, C. S. (1983). *Latent structure analysis with choice modelling applications*. PhD thesis, Aarhus School of Business Administration and Economics, University of Pennsylvania. Dissertation available at https://repository.upenn.edu/dissertations/AAI8316074/.

Rao, J. N. K. (1994). Estimating totals and distribution functions using auxiliary information at the estimation stage. *Journal of Official Statistics*, 10:153–165.

Reiter, J. P. Raghunathan, T. E. (2007). The multiple adaptations of multiple imputation. *Journal of the American Statistical Association*, 102(480):1462–1471.

Rendtel, U. (2002). Attrition in Household Panels: A Survey, CHINTEX working paper No. 4. Available at http://www.destatis.de/chintex/download/paper4.pdf.

Rendtel, U. (2003). Attrition effects in the European Community Household Panel, Bulletin of the ISI 54th Session, Contributed papers, Volume LX, Book 2, 316-317.

Rendtel, U. (2013). The fade-away effect of initial non-response in panel surveys, Empirical results for EU-SILC. Euro-stat methodologies and working papers (ISSN 1977-0375) edition 2013, doi:10.2785/21863.

Rendtel, U., Behr, A., Bellgardt, E., Neukirch, T., Pyy-Martikainen, M., Sisto, J., Lehtonen, R., Harms, T., Basic, E., and Marek, I. (2004). Report on panel effects. Results of work package 6 of the CHINTEX project, CHINTEX. Available at https://www.destatis.de/DE/Methoden/Methodenpapiere/Chintex/ Projekt/-Workpackage6.html.

Rendtel, U. and Harms, T. (2009). *Weighting and calibration for household surveys*, chapter 15, pages 265–286. Methodology of Longitudinal Surveys, Chichester: John Wiley & Sons.

Riphahn, R. T. and Serfling, O. (2005). Item non-response on income and wealth questions. *Empirical Economics*, 30(2):521–538.

Robins, J. M. and Rotnitzky, A. (1995). Semi-parametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90:122–129.

Robins, J. M., Rotnitzky, A., and Zhao, L. (1995). Analysis of semi-parametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90:106–121.

Rotnitzky, A. and Robins, J. M. (1995). Semi-parametric regression estimation in the presence of dependent censoring. *Biometrika*, 82:805–820.

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

Rubin, D. B. (1987). *Multiple imputation for non-response in surveys*. John Wiley, New York.

Särndal, C. E. (2007). The calibration approach in survey theory and practice. *Survey Methodology*, 33:99–119.

Särndal, C. E., Swenson, B., and Wretman, J. H. (1989). The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika*, 76:527–537.

Särndal, C. E., Swensson, B., and Wretman, J. (1992). *Model assisted survey sampling*. Springer-Verlag, New York.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. Chapman Hall Press, New York, US.

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999). Adjusting for non-ignorable drop-out using semi-parametric non-response models. *Journal of the American Statistical Association*, 94:1096–1120.

Singer, B. and Spilerman, S. (1976). The representation of social processes by Markov models. *American Journal of Sociology*, 82:1–54.

Sisto, J. (2003). Attrition effects on the design-based estimates of disposable household income. CHINTEX working paper No. 9 03/2003, URL: www.destatis.de/chintex.

Snell, E. J. (1964). A scaling procedure for ordered categorical data. *Biometrics*, 20:592–607.

Spiess, M. (2000). Derivation of design weights: The case of the German Socio Economic Panel (GSOEP). Discussion paper No. 197, DIW Berlin. Available at https://econpapers.repec.org/paper/diwdiwwpp/dp197.htm.

Spilerman, N. (1972). Extensions of the mover-stayer model. *American Journal of Sociology*, 78:599–626.

Thijs, H., Molenberghs, G., Michiels, B., Verbeke, G., and Curran, D. (2000). Strategies to fit pattern mixture models. *Biostatistics*, 3:245–265.

Van de Pol, F. and Langeheine, R. (1989). Mixed Markov models, mover-stayer models and the EM algorithm. With an application to labor market data from the Netherlands Socio Economic Panel. In R. Coppi and S. Bolasco (Eds.). *Multiway Data Analysis. Amsterdam: North-Holland*, pages 485–495.

Wagner, G. G., Frick, J. R., and Schupp, j. (2007). The German Socio-Economic Panel Study (SOEP) - scope, evolution and enhancements. *Schmollers Jahrbuch*, 127(1):139–169.

Watson, D. (2003). Sample attrition between waves 1 and 5 in the European Community Household Panel. *European Sociological Review*, 19(4):361–378.

Wawro, G. (2002). Estimating dynamic panel data models in political science. *Political Analysis*, 10:25–48.

Williams, R. (2006). Generalized ordered logit/partial proportional odds models for ordinal dependent variables. *The Stata Journal*, 6(1):58–82.

Winship, C. and Mare, R. D. (1984). Regression models with ordinal variables. *American Sociological Review*, 49:512–525.

Wooldridge, J. (2002). Inverse probability weighted M-estimators for sample selection, attrition and stratification. *Portuguese Economic Journal*, 1:141–162.

Wooldridge, J. (2007). Inverse probability weighted estimation for general missing data problems. *Journal of Econometrics*, 141:1281–1301.

Wooldridge, J. M. (2009). *Econometric analysis of cross section and panel data*. The MIT Press Cambridge, Massachusetts London, England, 2nd edition.

Wu, C. and Sitter, R. R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96:185–193.

CHAPTER 6

Appendix

A. Appendix of Chapter 3

In Chapter 3 of this thesis, we graphically displayed the fade-away effect for the different model estimators using simulation data. These consist of linear models for cross-sectional data and longitudinal data. The numerical results of these estimators are summarized in Appendix A. For the sake of convenience we display the estimation results of different estimators in different sections of this appendix. In the first section (Section A.1) we show the fade-away effect of the cross-sectional OLS estimator in a four wave panel data. We then compared the results of the weighted with un-weighted cross-sectional OLS estimator in Section A.2. The results of the different panel model estimators are given in Section A.3. Finally, the results of the non-linear model estimator are presented in Section A.4: Particularly, in this, we compared the fade-away results of the weighted and un-weighted estimates of ordered logit model under Scenario A-D.

A.1. Fade-away effect for the cross-sectional OLS estimator in a four wave panel data under Scenario A-G

Table 9: Response probabilities P_t, percent relative biases B_{tsim} and the speed of the fade-away effect λ_{tsim}, for fix non-response parameters $\alpha = 0.80, \beta = 0.05$ and attrition parameters $\alpha^\star = 0.80, \beta^\star = 0.05$.

σ^2	κ, γ	Response probability				Relative bias (RB)*100				Relative factor (λ)		
	ρ, ϕ	P_1	P_2	P_3	P_4	B_{1sim}	B_{2sim}	B_{3sim}	B_{4sim}	λ_{1sim}	λ_{2sim}	λ_{3sim}
0.2	0.2	0.90	0.90	0.90	0.90	0.14	0.15	0.15	0.18	1.07	1.00	1.20
0.4	0.2	0.90	0.90	0.90	0.90	0.30	0.33	0.39	0.42	1.10	1.18	1.08
0.6	0.2	0.90	0.90	0.90	0.90	0.49	0.51	0.60	0.60	1.04	1.18	1.00
0.8	0.2	0.90	0.90	0.90	0.90	0.63	0.77	0.86	0.90	1.22	1.12	1.05
1.0	0.2	0.90	0.90	0.90	0.90	0.87	1.02	1.03	1.07	1.17	1.01	1.04
0.2	0.4	0.90	0.90	0.90	0.90	0.14	0.23	0.23	0.27	1.64	1.00	1.17
0.4	0.4	0.90	0.90	0.90	0.90	0.30	0.48	0.51	0.59	1.60	1.06	1.16
0.6	0.4	0.90	0.90	0.90	0.91	0.47	0.71	0.81	0.92	1.51	1.14	1.14
0.8	0.4	0.90	0.90	0.90	0.91	0.65	1.01	1.15	1.29	1.55	1.14	1.12
1.0	0.4	0.90	0.90	0.90	0.91	0.83	1.27	1.47	1.66	1.53	1.16	1.13
0.2	0.6	0.90	0.90	0.90	0.91	0.15	0.25	0.32	0.39	1.67	1.28	1.22
0.4	0.6	0.90	0.90	0.91	0.91	0.27	0.50	0.69	0.87	1.85	1.38	1.26
0.6	0.6	0.90	0.90	0.91	0.91	0.50	0.84	1.14	1.45	1.68	1.36	1.27
0.8	0.6	0.90	0.90	0.91	0.91	0.71	1.19	1.63	1.95	1.68	1.37	1.20
1.0	0.6	0.90	0.90	0.91	0.91	0.84	1.53	2.06	2.51	1.82	1.35	1.22

Table 10: Speed of the fade-away phenomenon of initial non-response bias in Scenario A ($\kappa = \gamma = \rho = \phi = 0.10$) without any attrition pattern.

σ^2	Relative bias*100						Relative factor (λ)			
	B_{1com}	B_{1sim}	B_{2com}	B_{2sim}	B_{3sim}	B_{4sim}	λ_{1com}	λ_{1sim}	λ_{2sim}	λ_{3sim}
0.10	2.84	3.98	0.08	0.00	0.00	0.00	0.03	0.00	0.00	0.00
0.20	5.69	6.88	0.16	0.13	0.10	0.30	0.03	0.02	0.77	3.00
0.30	8.53	9.88	0.24	0.04	0.04	0.05	0.03	0.01	0.01	1.25
0.40	11.31	12.68	0.32	0.20	0.50	0.57	0.03	0.02	2.50	1.14
0.50	14.22	14.82	0.41	0.90	0.16	0.06	0.03	0.06	0.17	0.38
0.60	17.07	17.14	0.49	0.30	0.35	0.08	0.03	0.02	1.17	0.23
0.70	19.91	18.99	0.57	0.16	0.17	0.37	0.03	0.01	1.06	2.18
0.80	22.76	20.67	0.65	0.81	0.18	0.79	0.03	0.04	0.22	4.39
0.90	25.60	22.80	0.73	1.35	0.51	0.14	0.03	0.06	0.38	0.28
1.00	28.44	24.19	0.81	1.42	0.51	0.02	0.03	0.06	0.36	0.04

Note: The bias of the OLS estimates through approximation formula is represented by B_{tcom}, while the relative factors of the speed of the fade-away effect are denoted by λ_{tcom}. Similarly, the bias of the OLS estimates through simulation study are represented by B_{tsim}, while the relative factors of the fade-away effect are denoted by λ_{tsim}, where $t = 1, 2, 3, 4$.

Table 11: Speed of the fade-away phenomenon of initial non-response bias in Scenario B ($\kappa = \gamma = \rho = \phi = 0.50$) without any attrition pattern.

σ^2	Relative bias (RB)*100						Relative factor (λ)			
	B_{1com}	B_{1sim}	B_{2com}	B_{2sim}	B_{3sim}	B_{4sim}	λ_{1com}	λ_{1sim}	λ_{2sim}	λ_{3sim}
0.10	2.84	3.74	1.42	1.72	0.97	0.78	0.50	0.46	0.56	0.80
0.20	5.69	7.05	2.85	3.45	2.18	2.07	0.50	0.49	0.63	0.95
0.30	8.53	10.25	4.27	4.82	3.21	2.61	0.50	0.47	0.67	0.81
0.40	11.31	12.61	5.69	6.05	4.26	2.98	0.50	0.48	0.70	0.70
0.50	14.22	14.97	7.12	7.92	4.59	3.72	0.50	0.53	0.58	0.81
0.60	17.07	17.57	8.54	8.46	5.87	4.63	0.50	0.48	0.69	0.79
0.70	19.91	19.10	9.96	9.53	6.19	5.37	0.50	0.50	0.65	0.87
0.80	22.76	20.53	11.38	10.44	6.72	5.65	0.50	0.51	0.64	0.84
0.90	25.60	22.85	12.81	11.75	7.99	6.48	0.50	0.51	0.68	0.81
1.00	28.44	24.36	14.23	12.63	8.31	6.78	0.50	0.52	0.66	0.82

Table 12: Speed of the fade-away phenomenon of initial non-response bias in Scenario C ($\kappa = \gamma = \rho = \phi = 0.70$) without any attrition pattern.

σ^2	Relative bias (RB)*100						Relative factor (λ)			
	B_{1com}	B_{1sim}	B_{2com}	B_{2sim}	B_{3sim}	B_{4sim}	λ_{1com}	λ_{1sim}	λ_{2sim}	λ_{3sim}
0.10	2.84	3.72	2.25	2.85	2.36	2.10	0.79	0.77	0.83	0.89
0.20	5.69	7.17	4.49	5.34	4.65	4.18	0.79	0.75	0.87	0.90
0.30	8.53	10.33	6.74	8.17	6.72	5.87	0.79	0.79	0.82	0.87
0.40	11.31	12.83	8.98	10.10	8.39	7.53	0.79	0.79	0.83	0.90
0.50	14.22	14.93	11.23	11.71	9.82	8.59	0.79	0.78	0.84	0.88
0.60	17.07	17.45	13.48	13.63	11.34	10.27	0.79	0.78	0.83	0.91
0.70	19.91	19.55	15.72	15.46	13.24	11.68	0.79	0.79	0.86	0.88
0.80	22.76	21.57	17.97	17.26	14.53	12.79	0.79	0.80	0.84	0.88
0.90	25.60	23.19	20.21	18.33	15.06	13.27	0.79	0.79	0.82	0.88
1.00	28.44	24.54	22.46	19.23	16.24	14.15	0.79	0.78	0.85	0.87

Table 13: Speed of the fade-away phenomenon of initial non-response bias in Scenario D ($\kappa = \gamma = \rho = \phi = 0.90$) without any attrition pattern.

σ^2	Relative bias (RB)*100						Relative factor (λ)			
	B_{1com}	B_{1sim}	B_{2com}	B_{2sim}	B_{3sim}	B_{4sim}	λ_{1com}	λ_{1sim}	λ_{2sim}	λ_{3sim}
0.10	2.84	3.88	2.77	3.80	3.75	3.67	0.89	0.98	0.99	0.98
0.20	5.69	6.87	5.54	6.70	6.62	6.45	0.97	0.98	0.99	0.97
0.30	8.53	10.02	8.32	9.77	9.61	9.35	0.98	0.98	0.98	0.97
0.40	11.31	12.60	11.09	12.32	12.13	11.88	0.98	0.98	0.99	0.98
0.50	14.22	15.48	13.86	15.12	14.79	14.45	0.98	0.98	0.98	0.98
0.60	17.07	17.02	16.63	16.66	16.24	15.83	0.97	0.98	0.98	0.98
0.70	19.91	19.28	19.40	18.84	18.42	18.03	0.97	0.98	0.98	0.98
0.80	22.76	21.05	22.17	20.53	20.05	19.53	0.97	0.98	0.98	0.97
0.90	25.60	22.95	24.95	22.32	21.87	21.34	0.97	0.97	0.98	0.98
1.00	28.44	23.65	27.72	23.14	22.60	22.20	0.98	0.98	0.98	0.98

Table 14: Speed of the fade-away phenomenon of initial non-response bias in Scenario E ($\kappa = \rho = 0.10, \gamma = 0.01$ and $\phi = 0.70$) without any attrition pattern.

σ^2	Relative bias (RB)*100						Relative factor (λ)			
	B_{1com}	B_{1sim}	B_{2com}	B_{2sim}	B_{3sim}	B_{4sim}	λ_{1com}	λ_{1sim}	λ_{2sim}	λ_{3sim}
0.10	2.84	3.83	0.30	0.32	0.22	0.05	0.11	0.08	0.69	0.23
0.20	5.69	7.16	0.60	0.71	0.28	0.08	0.11	0.10	0.39	0.29
0.30	8.53	10.11	0.90	1.10	0.40	0.19	0.11	0.11	0.36	0.47
0.40	11.31	12.73	1.20	1.38	0.61	0.35	0.11	0.11	0.44	0.57
0.50	14.22	15.60	1.50	1.63	0.66	0.56	0.11	011	0.41	0.85
0.60	17.07	17.06	1.79	1.86	0.74	0.30	0.11	0.11	0.40	0.41
0.70	19.91	18.70	2.09	1.78	0.51	0.48	0.11	0.10	0.29	0.94
0.80	22.76	20.90	2.39	2.49	0.96	0.61	0.11	0.12	0.39	0.64
0.90	25.60	22.85	2.69	1.88	1.02	0.43	0.11	0.08	0.54	0.42
1.00	28.44	23.64	2.99	2.67	0.74	0.33	0.11	0.11	0.28	0.45

Table 15: Speed of the fade-away phenomenon of initial non-response bias in Scenario F ($\kappa = \rho = 0.50, \gamma = 0.01$ and $\phi = 0.70$) without any attrition pattern.

σ^2	Relative bias (RB)*100						Relative factor (λ)			
	B_{1com}	B_{1sim}	B_{2com}	B_{2sim}	B_{3sim}	B_{4sim}	λ_{1com}	λ_{1sim}	λ_{2sim}	λ_{3sim}
0.10	2.84	3.85	1.33	1.77	0.95	0.45	0.47	0.46	0.54	0.47
0.20	5.69	7.01	2.67	3.20	1.86	1.06	0.47	0.46	0.58	0.57
0.30	8.53	10.10	04.00	4.73	2.60	1.45	0.47	0.47	0.55	0.56
0.40	11.31	12.94	5.34	5.79	3.18	1.93	0.47	0.45	0.55	0.61
0.50	14.22	14.79	6.67	7.34	4.10	2.74	0.47	0.45	0.56	0.67
0.60	17.07	17.23	8.01	8.45	4.76	3.04	0.47	0.49	0.56	0.64
0.70	19.91	19.23	9.34	8.91	4.93	2.89	0.47	0.46	0.55	0.59
0.80	22.76	21.30	10.68	9.99	6.03	3.76	0.47	0.47	0.60	0.62
0.90	25.60	23.08	12.01	10.96	6.30	3.75	0.47	0.48	0.58	0.60
1.00	28.44	23.42	13.34	11.25	6.49	3.76	0.47	0.48	0.58	0.58

Table 16: Speed of the fade-away phenomenon of initial non-response bias in Scenario G ($\kappa = \rho = 0.70, \gamma = 0.01$ and $\phi = 0.70$) without any attrition pattern.

σ^2	Relative bias (RB)*100						Relative factor (λ)			
	B_{1com}	B_{1sim}	B_{2com}	B_{2sim}	B_{3sim}	B_{4sim}	λ_{1com}	λ_{1sim}	λ_{2sim}	λ_{3sim}
0.10	2.84	3.76	1.74	2.33	1.40	0.84	0.61	0.62	0.60	0.60
0.20	5.69	7.32	3.47	4.45	2.91	2.09	0.61	0.61	0.65	0.72
0.30	8.53	10.23	5.21	6.23	3.59	2.43	0.61	0.61	0.58	0.68
0.40	11.31	12.95	6.94	8.09	5.03	3.08	0.61	0.63	0.62	0.61
0.50	14.22	14.83	8.68	9.51	6.05	3.97	0.61	0.64	0.64	0.66
0.60	17.07	17.12	10.41	10.30	6.64	4.34	0.61	0.60	0.65	0.65
0.70	19.91	18.85	12.15	11.99	7.33	4.67	0.61	0.64	0.61	0.64
0.80	22.76	20.79	13.89	12.60	8.25	4.97	0.61	0.61	0.66	0.60
0.90	25.60	22.33	15.62	13.43	8.69	5.55	0.61	0.60	0.65	0.64
1.00	28.44	24.02	17.36	14.72	9.21	6.14	0.61	0.61	0.63	0.67

Table 17: Speed of the fade-away phenomenon of initial non-response bias in Scenario H ($\kappa = \rho = 0.90, \gamma = 0.01$ and $\phi = 0.70$) without any attrition pattern.

σ^2	Relative bias (RB)*100						Relative factor (λ)			
	B_{1com}	B_{1sim}	B_{2com}	B_{2sim}	B_{3sim}	B_{4sim}	λ_{1com}	λ_{1sim}	λ_{2sim}	λ_{3sim}
0.10	2.84	3.69	1.97	2.50	1.72	1.19	0.69	0.68	0.69	0.69
0.20	5.69	6.98	3.94	4.82	3.42	2.51	0.69	0.69	0.71	0.73
0.30	8.53	10.09	5.91	7.07	4.72	3.69	0.69	0.70	0.67	0.78
0.40	11.31	13.13	7.88	8.96	6.24	4.60	0.70	0.68	0.70	0.74
0.50	14.22	14.97	9.85	11.15	8.06	5.97	0.69	0.75	0.72	0.74
0.60	17.07	17.41	11.82	12.07	8.37	6.05	0.69	0.69	0.69	0.72
0.70	19.91	19.09	13.79	13.25	8.97	6.17	0.69	0.69	0.68	0.69
0.80	22.76	21.09	15.82	14.75	10.40	7.42	0.70	0.70	0.71	0.71
0.90	25.60	22.68	17.72	15.53	11.02	7.61	0.69	0.69	0.71	0.69
1.00	28.44	23.85	19.69	16.47	11.22	7.59	0.69	0.69	0.68	0.68

Table 18: Speed of the fade-away phenomenon of initial non-response bias in Scenario A ($\kappa = \gamma = \rho = \phi = 0.10$), for fix non-response parameters $\alpha = 0.05, \beta = 0.40$ and attrition parameters $\alpha^\star = 0.80, \beta^\star = 0.05$.

σ^2	Relative bias (RB)*100				Relative factor (λ)		
	B_{1sim}	B_{2sim}	B_{3sim}	B_{4sim}	λ_{1sim}	λ_{2sim}	λ_{3sim}
0.10	3.84	0.29	0.22	0.26	0.08	0.76	1.18
0.20	7.06	0.52	0.40	0.23	0.07	0.77	0.58
0.30	9.65	0.72	0.27	0.19	0.08	0.38	0.70
0.40	12.68	0.69	0.40	0.57	0.05	0.58	1.43
0.50	15.50	0.71	0.98	0.53	0.05	1.38	0.54
0.60	16.98	0.32	0.37	1.25	0.02	1.16	3.38
0.70	19.15	1.01	0.89	0.64	0.05	0.88	0.72
0.80	21.08	1.29	0.97	0.20	0.06	0.75	0.21
0.90	22.71	1.08	1.24	0.91	0.05	1.15	0.73
1.00	24.02	1.09	1.33	0.88	0.05	1.22	0.66

Note: The bias of the cross-sectional OLS estimates in each panel wave is represented by B_{tsim}, where $t = 1, 2, 3, 4$, while the relative factors of the fade-away effect are denoted by λ_{tsim}.

Table 19: Speed of the fade-away phenomenon of initial non-response bias in Scenario B ($\kappa = \gamma = \rho = \phi = 0.50$), for fix non-response parameters $\alpha = 0.05, \beta = 0.40$ and attrition parameters $\alpha^\star = 0.80, \beta^\star = 0.05$.

σ^2	Relative bias (RB)*100				Relative factor (λ)		
	B_{1sim}	B_{2sim}	B_{3sim}	B_{4sim}	λ_{1sim}	λ_{2sim}	λ_{3sim}
0.10	3.82	1.68	1.18	1.04	0.44	0.70	0.88
0.20	7.02	3.59	2.66	1.94	0.51	0.74	0.73
0.30	10.34	4.98	3.82	2.96	0.48	0.77	0.78
0.40	12.38	6.52	4.81	3.86	0.53	0.74	0.80
0.50	15.04	7.88	5.14	3.75	0.52	0.65	0.73
0.60	17.25	8.84	6.49	5.21	0.51	0.73	0.80
0.70	19.39	10.27	7.62	6.89	0.53	0.74	0.90
0.80	21.29	10.78	7.87	6.89	0.51	0.73	0.88
0.90	23.45	12.92	9.10	7.76	0.55	0.70	0.85
1.00	24.68	13.14	9.71	8.10	0.53	0.74	0.83

Note: The bias of the cross-sectional OLS estimates in each panel wave is represented by B_{tsim}, where $t = 1, 2, 3, 4$, while the relative factors of the fade-away effect are denoted by λ_{tsim}.

Table 20: Speed of the fade-away phenomenon of initial non-response bias in Scenario C ($\kappa = \gamma = \rho = \phi = 0.70$), for fix non-response parameters $\alpha = 0.05, \beta = 0.40$ and attrition parameters $\alpha^\star = 0.80, \beta^\star = 0.05$.

σ^2	Relative bias (RB)*100				Relative factor (λ)		
	B_{1sim}	B_{2sim}	B_{3sim}	B_{4sim}	λ_{1sim}	λ_{2sim}	λ_{3sim}
0.10	3.73	2.82	2.48	2.31	0.76	0.00	0.93
0.20	7.10	5.69	4.73	4.16	0.80	0.83	0.88
0.30	9.74	7.58	6.62	6.01	0.78	0.87	0.91
0.40	12.68	9.99	8.39	7.47	0.79	0.84	0.89
0.50	15.36	12.10	10.27	9.30	0.79	0.85	0.91
0.60	17.50	14.05	12.01	10.80	0.80	0.86	0.90
0.70	19.15	15.56	13.56	12.29	0.81	0.87	0.91
0.80	21.20	17.21	14.55	13.12	0.81	0.85	0.90
0.90	22.75	18.27	15.32	14.60	0.80	0.84	0.95
1.00	24.16	19.34	16.61	15.12	0.80	0.86	0.91

Note: The bias of the cross-sectional OLS estimates in each panel wave is represented by B_{tsim}, where $t = 1, 2, 3, 4$, while the relative factors of the fade-away effect are denoted by λ_{tsim}.

Table 21: Speed of the fade-away phenomenon of initial non-response bias in Scenario D ($\kappa = \gamma = \rho = \phi = 0.90$), for fix non-response parameters $\alpha = 0.05, \beta = 0.40$ and attrition parameters $\alpha^\star = 0.80, \beta^\star = 0.05$.

σ^2	Relative bias (RB)*100				Relative factor (λ)		
	B_{1sim}	B_{2sim}	B_{3sim}	B_{4sim}	λ_{1sim}	λ_{2sim}	λ_{3sim}
0.10	3.72	3.59	3.44	3.29	0.97	0.96	0.96
0.20	6.87	6.62	6.47	6.33	0.96	0.98	0.98
0.30	10.10	9.71	9.48	9.42	0.96	0.98	0.99
0.40	12.66	12.38	11.93	11.60	0.98	0.96	0.97
0.50	14.73	14.20	13.64	13.31	0.96	0.96	0.98
0.60	17.63	17.01	16.50	16.07	0.97	0.97	0.97
0.70	18.87	18.24	17.89	17.82	1.00	1.00	1.00
0.80	20.35	19.72	19.18	19.03	0.97	0.97	0.99
0.90	22.75	21.85	21.45	21.17	0.96	0.98	0.99
1.00	24.80	24.21	23.84	23.39	0.98	0.99	0.98

Note: The bias of the cross-sectional OLS estimates in each panel wave is represented by B_{tsim}, where $t = 1, 2, 3, 4$, while the relative factors of the fade-away effect are denoted by λ_{tsim}.

Table 22: Speed of the fade-away phenomenon of initial non-response bias in Scenario A ($\kappa = \gamma = \rho = \phi = 0.10$), for fix non-response parameters $\alpha = 0.05, \beta = 0.40$ and attrition parameters $\alpha^{\star} = 0.70, \beta^{\star} = 0.10$.

σ^2	Relative bias (RB)*100				Relative factor (λ)		
	B_{1sim}	B_{2sim}	B_{3sim}	B_{4sim}	λ_{1sim}	λ_{2sim}	λ_{3sim}
0.10	3.76	0.78	0.39	0.22	0.21	0.50	0.56
0.20	6.96	1.19	0.68	1.02	0.17	0.57	1.50
0.30	10.06	1.71	1.50	0.91	0.17	0.88	0.61
0.40	12.53	1.66	1.91	1.49	0.13	1.15	0.78
0.50	14.95	1.77	1.69	2.68	0.12	0.96	1.59
0.60	17.37	2.54	3.20	2.83	0.15	1.26	0.88
0.70	19.44	3.48	3.30	2.48	0.18	0.95	0.75
0.80	21.03	4.10	3.93	2.87	0.20	0.96	0.73
0.90	22.84	3.25	2.62	2.83	0.14	0.81	1.08
1.00	24.52	4.15	3.34	3.41	0.17	0.81	1.02

Note: The bias of the cross-sectional OLS estimates in each panel wave is represented by B_{tsim}, where $t = 1, 2, 3, 4$, while the relative factors of the fade-away effect are denoted by λ_{tsim}.

Table 23: Speed of the fade-away phenomenon of initial non-response bias in Scenario B ($\kappa = \gamma = \rho = \phi = 0.50$), for fix non-response parameters $\alpha = 0.05, \beta = 0.40$ and attrition parameters $\alpha^{\star} = 0.70, \beta^{\star} = 0.10$.

σ^2	Relative bias (RB)*100				Relative factor (λ)		
	B_{1sim}	B_{2sim}	B_{3sim}	B_{4sim}	λ_{1sim}	λ_{2sim}	λ_{3sim}
0.10	3.48	1.88	1.49	1.57	0.54	0.79	1.05
0.20	7.10	4.14	3.47	3.48	0.58	0.84	1.00
0.30	10.07	6.28	4.58	4.48	0.62	0.73	0.98
0.40	12.75	7.61	6.03	5.92	0.60	0.79	0.98
0.50	14.89	8.97	7.19	6.41	0.60	0.80	0.89
0.60	17.10	10.23	8.68	7.73	0.60	0.85	0.89
0.70	18.79	11.35	9.41	9.02	0.60	0.83	0.96
0.80	21.03	13.35	10.96	10.38	0.64	0.82	0.95
0.90	22.45	13.46	10.87	10.54	0.60	0.81	0.97
1.00	23.84	14.04	12.26	10.58	0.59	0.87	0.86

Note: The bias of the cross-sectional OLS estimates in each panel wave is represented by B_{tsim}, where $t = 1, 2, 3, 4$, while the relative factors of the fade-away effect are denoted by λ_{tsim}.

Table 24: Speed of the fade-away phenomenon of initial non-response bias in Scenario C ($\kappa = \gamma = \rho = \phi = 0.70$), for fix non-response parameters $\alpha = 0.05, \beta = 0.40$ and attrition parameters $\alpha^\star = 0.70, \beta^\star = 0.10$.

σ^2	Relative bias (RB)*100				Relative factor (λ)		
	B_{1sim}	B_{2sim}	B_{3sim}	B_{4sim}	λ_{1sim}	λ_{2sim}	λ_{3sim}
0.10	3.74	3.16	3.04	3.08	0.85	0.96	1.01
0.20	7.48	6.30	6.07	5.73	0.84	0.96	0.94
0.30	9.88	8.45	7.78	7.78	0.86	0.92	1.00
0.40	12.66	10.99	10.11	10.18	0.87	0.92	1.01
0.50	14.93	13.07	12.22	11.85	0.88	0.94	0.97
0.60	17.59	15.23	14.07	13.80	0.87	0.92	0.98
0.70	19.40	16.63	15.42	15.21	0.86	0.93	0.99
0.80	21.10	18.24	17.37	16.74	0.87	0.95	0.96
0.90	21.98	18.90	17.73	17.56	0.86	0.94	0.99
1.00	23.45	20.18	19.06	18.25	0.86	0.95	0.96

Note: The bias of the cross-sectional OLS estimates in each panel wave is represented by B_{tsim}, where $t = 1, 2, 3, 4$, while the relative factors of the fade-away effect are denoted by λ_{tsim}.

Table 25: Speed of the fade-away phenomenon of initial non-response bias in Scenario D ($\kappa = \gamma = \rho = \phi = 0.90$), for fix non-response parameters $\alpha = 0.05, \beta = 0.40$ and attrition parameters $\alpha^\star = 0.70, \beta^\star = 0.10$.

σ^2	Relative bias (RB)*100				Relative factor (λ)		
	B_{1sim}	B_{2sim}	B_{3sim}	B_{4sim}	λ_{1sim}	λ_{2sim}	λ_{3sim}
0.10	3.81	3.94	4.09	4.33	1.03	1.04	1.06
0.20	7.05	7.33	7.69	8.07	1.04	1.05	1.05
0.30	10.21	10.47	10.86	11.39	1.03	1.04	1.05
0.40	12.79	13.29	13.93	14.61	1.04	1.05	1.05
0.50	15.02	15.48	16.05	16.91	1.03	1.04	1.05
0.60	17.15	17.80	18.47	19.33	1.04	1.04	1.05
0.70	19.49	19.78	20.41	21.27	1.02	1.03	1.04
0.80	21.34	21.96	22.80	23.94	1.03	1.04	1.05
0.90	23.09	23.72	24.66	25.46	1.03	1.04	1.03
1.00	24.32	24.98	25.95	26.89	1.03	1.04	1.04

Note: The bias of the cross-sectional OLS estimates in each panel wave is represented by B_{tsim}, where $t = 1, 2, 3, 4$, while the relative factors of the fade-away effect are denoted by λ_{tsim}.

Table 26: Speed of the fade-away phenomenon of initial non-response bias in Scenario A ($\kappa = \gamma = \rho = \phi = 0.10$), for fix non-response parameters $\alpha = 0.05, \beta = 0.40$ and attrition parameters $\alpha^{\star} = 0.50, \beta^{\star} = 0.20$.

σ^2	Relative bias (RB)*100				Relative factor (λ)		
	B_{1sim}	B_{2sim}	B_{3sim}	B_{4sim}	λ_{1sim}	λ_{2sim}	λ_{3sim}
0.10	3.66	1.38	1.14	1.33	0.38	0.83	1.17
0.20	7.20	2.14	2.55	2.27	0.30	1.19	0.89
0.30	10.03	3.67	3.15	3.14	0.37	0.86	1.00
0.40	12.99	4.70	4.48	4.36	0.36	0.95	0.97
0.50	15.19	5.73	5.66	5.43	0.38	0.99	0.96
0.60	17.26	5.96	5.77	6.16	0.35	0.97	1.07
0.70	19.30	6.74	7.16	7.53	0.35	1.06	1.05
0.80	20.91	8.02	8.01	8.67	0.38	1.00	1.08
0.90	23.00	9.96	9.59	8.13	0.43	0.96	0.85
1.00	24.17	10.27	9.73	8.94	0.45	0.95	0.92

Note: The bias of the cross-sectional OLS estimates in each panel wave is represented by B_{tsim}, where $t = 1, 2, 3, 4$, while the relative factors of the fade-away effect are denoted by λ_{tsim}.

Table 27: Speed of the fade-away phenomenon of initial non-response bias in Scenario B ($\kappa = \gamma = \rho = \phi = 0.50$), for fix non-response parameters $\alpha = 0.05, \beta = 0.40$ and attrition parameters $\alpha^{\star} = 0.50, \beta^{\star} = 0.20$.

σ^2	Relative bias (RB)*100				Relative factor (λ)		
	B_{1sim}	B_{2sim}	B_{3sim}	B_{4sim}	λ_{1sim}	λ_{2sim}	λ_{3sim}
0.10	3.89	2.88	2.78	2.89	0.74	0.97	1.04
0.20	7.04	5.31	5.33	5.61	0.75	1.00	1.05
0.30	10.02	7.72	7.42	7.54	0.77	0.96	1.02
0.40	12.66	9.51	9.13	9.31	0.75	0.96	1.02
0.50	15.16	11.03	10.84	11.17	0.73	0.98	1.03
0.60	17.31	12.65	11.66	12.54	0.73	0.92	1.08
0.70	19.02	14.38	13.64	13.71	0.76	0.95	1.01
0.80	20.48	15.84	14.90	15.38	0.77	0.94	1.03
0.90	22.56	17.18	16.63	17.11	0.76	0.97	1.03
1.00	24.65	18.47	17.74	17.59	0.75	0.96	0.99

Note: The bias of the cross-sectional OLS estimates in each panel wave is represented by B_{tsim}, where $t = 1, 2, 3, 4$, while the relative factors of the fade-away effect are denoted by λ_{tsim}.

Table 28: Speed of the fade-away phenomenon of initial non-response bias in Scenario C ($\kappa = \gamma = \rho = \phi = 0.70$), for fix non-response parameters $\alpha = 0.05, \beta = 0.40$ and attrition parameters $\alpha^{\star} = 0.50, \beta^{\star} = 0.20$.

σ^2	Relative bias (RB)*100				Relative factor (λ)		
	B_{1sim}	B_{2sim}	B_{3sim}	B_{4sim}	λ_{1sim}	λ_{2sim}	λ_{3sim}
0.10	3.80	3.96	4.18	4.56	1.04	1.06	1.09
0.20	7.24	7.23	7.29	8.09	1.00	1.01	1.11
0.30	10.21	10.26	10.93	11.44	1.01	1.07	1.05
0.40	13.26	13.38	13.82	14.68	1.01	1.03	1.06
0.50	15.29	15.31	15.97	16.32	1.00	1.04	1.02
0.60	17.46	17.16	17.99	18.96	0.98	1.05	1.05
0.70	19.98	19.46	19.77	20.38	0.97	1.02	1.03
0.80	20.90	20.26	20.99	21.64	0.97	1.04	1.03
0.90	22.53	21.59	22.07	23.33	0.96	1.02	1.06
1.00	24.53	23.92	23.78	24.69	0.98	0.99	1.04

Note: The bias of the cross-sectional OLS estimates in each panel wave is represented by B_{tsim}, where $t = 1, 2, 3, 4$, while the relative factors of the fade-away effect are denoted by λ_{tsim}.

Table 29: Speed of the fade-away phenomenon of initial non-response bias in Scenario D ($\kappa = \gamma = \rho = \phi = 0.90$), for fix non-response parameters $\alpha = 0.05, \beta = 0.40$ and attrition parameters $\alpha^{\star} = 0.50, \beta^{\star} = 0.20$.

σ^2	Relative bias (RB)*100				Relative factor (λ)		
	B_{1sim}	B_{2sim}	B_{3sim}	B_{4sim}	λ_{1sim}	λ_{2sim}	λ_{3sim}
0.10	3.71	4.37	5.17	5.82	1.18	1.18	1.13
0.20	6.83	8.06	9.22	10.47	1.18	1.14	1.14
0.30	9.98	11.55	13.14	14.58	1.16	1.14	1.11
0.40	12.95	14.77	16.64	18.50	1.14	1.13	1.11
0.50	15.49	17.59	19.77	21.57	1.14	1.12	1.09
0.60	17.18	19.36	21.57	23.68	1.13	1.11	1.10
0.70	18.76	20.91	23.22	25.70	1.12	1.11	1.11
0.80	21.11	23.51	25.95	28.24	1.11	1.10	1.09
0.90	22.61	25.09	27.58	29.90	1.11	1.10	1.08
1.00	24.47	26.91	29.50	31.56	1.10	1.10	1.07

Note: The bias of the cross-sectional OLS estimates in each panel wave is represented by B_{tsim}, where $t = 1, 2, 3, 4$, while the relative factors of the fade-away effect are denoted by λ_{tsim}.

A.2. Comparison of the weighted and un-weighted cross-sectional OLS estimators in Scenario A-D

Table 30: Fade-away effect for the weighted and un-weighted cross-sectional OLS estimators in Scenario A ($\kappa = \gamma = \rho = \phi = 0.10$) for fix non-response parameters $\alpha = -4.50, \beta = 3.20$ and attrition parameters $\alpha^\star = 0.90, \beta^\star = 0.90$.

Panel wave	Case numbers	Relative bias (RB)*100				
		Bias($\hat{b}^{OLS}_{Full}$)	Bias($\hat{b}^{OLS}_{Resp}$)	Bias($\hat{b}^{IPW_1}_{Resp}$)	Bias($\hat{b}^{IPW_2}_{Resp}$)	Bias($\hat{b}^{IPW_3}_{Resp}$)
1	65128	0.71 (0.10)	-29.93 (9.09)	-35.57 (12.86)	-35.57 (12.86)	-9.82 (2.83)
2	59128	-0.17 (0.09)	-0.88 (0.15)	-0.84 (0.22)	-6.18 (0.61)	-5.59 (1.22)
3	53318	-0.18 (0.09)	-0.29 (0.18)	-0.11 (0.29)	-5.98 (0.63)	-5.27 (1.39)
4	48094	-0.18 (0.09)	-0.69 (0.16)	-0.58 (0.29)	-5.94 (0.60)	-5.93 (1.45)
5	43382	-0.18 (0.09)	-0.86 (0.19)	-0.60 (0.31)	-6.49 (0.72)	-5.44 (1.50)
6	39298	-0.18 (0.09)	-0.73 (0.19)	-0.38 (0.40)	-5.87 (0.63)	-5.48 (1.72)
7	35547	-0.18 (0.09)	-0.80 (0.26)	-1.31 (0.58)	-5.79 (0.69)	-5.33 (1.70)
8	32190	-0.18 (0.09)	-0.72 (0.31)	-1.09 (0.58)	-6.17 (0.69)	-5.81 (2.07)
9	29234	-0.18 (0.09)	-1.15 (0.34)	-1.31 (0.59)	-5.82 (0.76)	-4.83 (2.36)
10	26534	-0.18 (0.09)	-0.80 (0.27)	-0.80 (0.63)	-5.75 (0.75)	-5.68 (2.10)

Note: The first column represents the wave of the panel, the second column refers to the number of respondents. Column 3 to 5 report the bias of the least-squares estimates: the results in columns 3 and 4 are obtained from un-weighted OLS regression, while the results in column 5 are obtained from weighted OLS regression, where the regression is weighted by the inverse of the estimated response probability. The MSE of the estimates is given in parenthesis, which is multiplied by 100.

Table 31: Fade-away effect for the weighted and un-weighted cross-sectional OLS estimators in Scenario B ($\kappa = \gamma = \rho = \phi = 0.50$) for fix non-response parameters $\alpha = -4.50, \beta = 3.20$ and attrition parameters $\alpha^\star = 0.90, \beta^\star = 0.90$.

Panel wave	Case numbers	Relative bias (RB)*100				
		Bias($\hat{b}^{OLS}_{Full}$)	Bias($\hat{b}^{OLS}_{Resp}$)	Bias($\hat{b}^{IPW_1}_{Resp}$)	Bias($\hat{b}^{IPW_2}_{Resp}$)	Bias($\hat{b}^{IPW_3}_{Resp}$)
1	65219	0.49 (0.11)	-30.05 (9.18)	-35.37 (12.71)	-35.37 (12.71)	-10.06 (2.88)
2	61226	-0.16 (0.09)	-15.28 (2.50)	-19.86 (4.16)	-22.19 (5.13)	-9.37 (2.40)
3	57189	-0.30 (0.09)	-11.05 (1.41)	-13.76 (2.17)	-17.05 (3.13)	-9.47 (2.20)
4	53433	-0.32 (0.09)	-9.82 (1.19)	-11.52 (1.60)	-15.45 (2.72)	-10.14 (2.59)
5	49860	-0.32 (0.09)	-10.25 (1.24)	-11.40 (1.65)	-15.43 (2.63)	-11.33 (2.76)
6	46666	-0.32 (0.09)	-10.08 (1.21)	-10.10 (1.56)	-15.60 (2.76)	-11.47 (3.05)
7	43713	-0.32 (0.09)	-9.77 (1.16)	-8.86 (1.47)	-15.89 (2.80)	-11.69 (3.51)
8	40985	-0.32 (0.09)	-10.11 (1.30)	-9.39 (1.46)	-16.49 (3.09)	-12.96 (3.94)
9	38569	-0.32 (0.09)	-10.25 (1.28)	-9.94 (1.47)	-16.26 (3.11)	-12.76 (4.16)
10	36340	-0.32 (0.09)	-11.08 (1.46)	-8.75 (1.81)	-15.88 (3.02)	-13.33 (4.21)

Table 32: Fade-away effect for the weighted and un-weighted cross-sectional OLS estimators in Scenario C ($\kappa = \gamma = \rho = \phi = 0.70$) for fix non-response parameters $\alpha = -4.50, \beta = 3.20$ and attrition parameters $\alpha^{\star} = 0.90, \beta^{\star} = 0.90$.

Panel wave	Case numbers	Relative bias (RB)*100				
		Bias($\hat{b}^{OLS}_{Full}$)	Bias($\hat{b}^{OLS}_{Resp}$)	Bias($\hat{b}^{IPW_1}_{Resp}$)	Bias($\hat{b}^{IPW_2}_{Resp}$)	Bias($\hat{b}^{IPW_3}_{Resp}$)
1	65249	0.29 (0.11)	-30.42 (9.40)	-35.81 (13.01)	-35.81 (13.01)	-10.30 (3.04)
2	61920	-0.09 (0.09)	-23.72 (5.80)	-29.44 (8.91)	-30.33 (9.41)	-10.53 (2.92)
3	58601	-0.27 (0.09)	-20.59 (4.42)	-25.11 (6.55)	-27.40 (7.71)	-12.31 (3.04)
4	55504	-0.33 (0.09)	-19.00 (3.76)	-22.51 (5.31)	-26.29 (7.14)	-13.42 (3.58)
5	52432	-0.35 (0.09)	-18.69 (3.68)	-21.67 (4.96)	-25.70 (6.92)	-13.32 (3.73)
6	49682	-0.35 (0.09)	-18.45 (3.61)	-20.70 (4.72)	-24.93 (6.50)	-13.65 (4.37)
7	47108	-0.35 (0.09)	-18.47 (3.60)	-20.44 (4.63)	-24.85 (6.51)	-15.04 (4.50)
8	44707	-0.35 (0.09)	-18.59 (3.66)	-19.95 (4.45)	-24.90 (6.48)	-15.84 (4.85)
9	42542	-0.35 (0.09)	-18.80 (3.75)	-19.06 (4.31)	-25.55 (6.90)	-16.50 (5.50)
10	40522	-0.35 (0.09)	-18.63 (3.71)	-19.23 (4.28)	-25.69 (7.01)	-17.10 (5.78)

Table 33: Fade-away effect for the weighted and un-weighted cross-sectional OLS estimators in Scenario D ($\kappa = \gamma = \rho = \phi = 0.90$) for fix non-response parameters $\alpha = -4.50, \beta = 3.20$ and attrition parameters $\alpha^{\star} = 0.90, \beta^{\star} = 0.90$.

Panel wave	Case numbers	Relative bias (RB)*100				
		Bias($\hat{b}^{OLS}_{Full}$)	Bias($\hat{b}^{OLS}_{Resp}$)	Bias($\hat{b}^{IPW_1}_{Resp}$)	Bias($\hat{b}^{IPW_2}_{Resp}$)	Bias($\hat{b}^{IPW_3}_{Resp}$)
1	65264	0.02 (0.11)	-30.79 (9.63)	-36.56 (13.55)	-36.56 (13.55)	-9.69 (2.87)
2	62290	-0.10 (0.10)	-30.37 (9.38)	-36.00 (13.18)	-36.27 (13.36)	-10.76 (2.91)
3	59400	-0.19 (0.10)	-30.10 (9.23)	-35.66 (12.92)	-36.17 (13.27)	-12.13 (3.41)
4	56736	-0.26 (0.09)	-29.62 (8.94)	-34.97 (12.46)	-35.75 (12.96)	-13.80 (3.76)
5	54160	-0.31 (0.09)	-29.49 (8.88)	-34.63 (12.22)	-35.29 (12.64)	-14.96 (4.02)
6	51872	-0.34 (0.09)	-29.19 (8.70)	-34.16 (11.95)	-35.04 (12.47)	-15.55 (4.43)
7	49680	-0.36 (0.09)	-29.18 (8.70)	-33.69 (11.61)	-35.24 (12.61)	-16.09 (4.96)
8	47629	-0.38 (0.09)	-29.05 (8.64)	-33.32 (11.39)	-35.23 (12.63)	-16.88 (5.39)
9	45743	-0.39 (0.09)	-29.20 (8.73)	-32.91 (11.12)	-35.09 (12.56)	-18.09 (5.79)
10	43937	-0.39 (0.09)	-29.17 (8.71)	-32.66 (10.99)	-35.30 (12.70)	-19.49 (6.49)

A.3. Fade-away effect for the panel model estimators in Scenario A-D

Table 34: Bias and MSE of the panel model estimators in Scenario A ($\kappa = \gamma = \rho = \phi = 0.10$) for fix non-response parameters $\alpha = -4.50, \beta = 3.20$ and attrition parameters $\alpha^\star = 0.90, \beta^\star = 0.90$.

Length	Case	Relative bias (RB)*100					
	numbers	Bias($\hat{b}^{Pooled}_{Full}$)	Bias($\hat{b}^{Pooled}_{Resp}$)	Bias($\hat{b}^{Fixone}_{Full}$)	Bias($\hat{b}^{Fixone}_{Resp}$)	Bias($\hat{b}^{Ranone}_{Full}$)	Bias($\hat{b}^{Ranone}_{Resp}$)
2	62385	0.27 (0.06)	-10.76 (1.25)	0.28 (0.10)	-3.27 (0.29)	0.26 (0.06)	-10.01 (1.09)
3	56538	0.10 (0.06)	-6.64 (0.50)	0.28 (0.10)	-2.11 (0.13)	0.18 (0.06)	-5.84 (0.40)
4	50931	0.02 (0.06)	-4.94 (0.29)	0.28 (0.10)	-1.51 (0.08)	0.18 (0.07)	-4.17 (0.22)
5	45968	-0.02 (0.07)	-4.06 (0.20)	0.28 (0.10)	-1.07 (0.05)	0.20 (0.08)	-3.28 (0.14)
6	41704	-0.05 (0.07)	-3.55 (0.16)	0.28 (0.10)	-0.76 (0.04)	0.21 (0.08)	-2.75 (0.10)
7	37698	-0.07 (0.07)	-3.24 (0.13)	0.28 (0.10)	-0.68 (0.04)	0.22 (0.09)	-2.45 (0.09)
8	34071	-0.08 (0.08)	-2.99 (0.12)	0.28 (0.10)	-0.54 (0.03)	0.22 (0.09)	-2.19 (0.07)
9	30906	-0.09 (0.08)	-2.81 (0.10)	0.28 (0.10)	-0.44 (0.03)	0.23 (0.09)	-2.00 (0.06)
10	28084	-0.11 (0.08)	-2.67 (0.10)	0.28 (0.10)	-0.39 (0.03)	0.23 (0.09)	-1.86 (0.06)

Note: The first column represents the length of the database. The second column refers to the number of individuals that participate in the survey at a given length. The third and fourth column contains the bias of the pooled OLS estimator under the Full and the Resp samples, respectively. Similarly, the next two columns contain the bias of the FE Within estimator under the Full and the Resp samples, while the last two columns display the bias of the RE model estimator for the Full and the Resp samples, respectively. The MSE of the estimates is given in parenthesis, which is multiplied by 100.

Table 35: Bias and MSE of the panel model estimators in Scenario B ($\kappa = \gamma = \rho = \phi = 0.50$) for fix non-response parameters $\alpha = -4.50, \beta = 3.20$ and attrition parameters $\alpha^\star = 0.90, \beta^\star = 0.90$.

Length	Case	Relative bias (RB)*100					
	numbers	Bias($\hat{b}^{Pooled}_{Full}$)	Bias($\hat{b}^{Pooled}_{Resp}$)	Bias($\hat{b}^{Fixone}_{Full}$)	Bias($\hat{b}^{Fixone}_{Resp}$)	Bias($\hat{b}^{Ranone}_{Full}$)	Bias($\hat{b}^{Ranone}_{Resp}$)
2	62249	0.17 (0.08)	-22.08 (5.00)	0.16 (0.10)	-1.00 (0.16)	0.15 (0.05)	-13.80 (1.99)
3	58490	-0.03 (0.07)	-18.17 (3.41)	0.16 (0.10)	-0.67 (0.09)	0.07 (0.06)	-8.65 (0.81)
4	54697	-0.12 (0.07)	-16.06 (2.68)	0.16 (0.10)	-0.58 (0.08)	0.07 (0.07)	-6.39 (0.47)
5	50979	-0.18 (0.07)	-14.82 (2.28)	0.16 (0.10)	-0.50 (0.06)	0.07 (0.08)	-5.16 (0.31)
6	47695	-0.21 (0.07)	-14.11 (2.07)	0.16 (0.10)	-0.55 (0.04)	0.08 (0.08)	-4.49 (0.24)
7	44692	-0.23 (0.07)	-13.53 (1.91)	0.16 (0.10)	-0.50 (0.04)	0.09 (0.09)	-3.94 (0.19)
8	41824	-0.24 (0.08)	-13.15 (1.81)	0.16 (0.10)	-0.49 (0.04)	0.10 (0.09)	-3.59 (0.16)
9	39272	-0.25 (0.08)	-12.88 (1.74)	0.16 (0.10)	-0.47 (0.04)	0.11 (0.09)	-3.33 (0.14)
10	37028	-0.26 (0.08)	-12.67 (1.68)	0.16 (0.10)	-0.48 (0.03)	0.11 (0.09)	-3.12 (0.13)

Table 36: Bias and MSE of the panel model estimators in Scenario C ($\kappa = \gamma = \rho = \phi = 0.70$) for fix non-response parameters $\alpha = -4.50, \beta = 3.20$ and attrition parameters $\alpha^\star = 0.90, \beta^\star = 0.90$.

Length	Case numbers	Relative bias (RB)*100					
		Bias($\hat{b}^{Pooled}_{Full}$)	Bias($\hat{b}^{Pooled}_{Resp}$)	Bias($\hat{b}^{Fixone}_{Full}$)	Bias($\hat{b}^{Fixone}_{Resp}$)	Bias($\hat{b}^{Ranone}_{Full}$)	Bias($\hat{b}^{Ranone}_{Resp}$)
2	62344	0.10 (0.09)	-27.58 (7.75)	0.11 (0.10)	-0.23 (0.16)	0.09 (0.05)	-15.02 (2.35)
3	59251	-0.04 (0.08)	-25.50 (6.66)	0.11 (0.10)	-0.23 (0.09)	0.04 (0.06)	-9.92 (1.05)
4	56250	-0.13 (0.08)	-24.05 (5.94)	0.11 (0.10)	-0.49 (0.08)	0.03 (0.07)	-7.60 (0.64)
5	53225	-0.19 (0.07)	-23.05 (5.46)	0.11 (0.10)	-0.56 (0.05)	0.03 (0.07)	-6.24 (0.43)
6	50456	-0.23 (0.07)	-22.36 (5.15)	0.11 (0.10)	-0.65 (0.05)	0.03 (0.08)	-5.42 (0.33)
7	47761	-0.25 (0.07)	-21.87 (4.92)	0.11 (0.10)	-0.58 (0.05)	0.04 (0.08)	-4.75 (0.26)
8	45334	-0.27 (0.08)	-21.53 (4.77)	0.11 (0.10)	-0.58 (0.04)	0.05 (0.09)	-4.31 (0.22)
9	43101	-0.28 (0.08)	-21.27 (4.65)	0.11 (0.10)	-0.60 (0.04)	0.05 (0.09)	-3.99 (0.19)
10	41043	-0.29 (0.08)	-21.05 (4.55)	0.11 (0.10)	-0.62 (0.04)	0.06 (0.09)	-3.75 (0.18)

Table 37: Bias and MSE of the panel model estimators in Scenario D ($\kappa = \gamma = \rho = \phi = 0.90$) for fix non-response parameters $\alpha = -4.50, \beta = 3.20$ and attrition parameters $\alpha^\star = 0.90, \beta^\star = 0.90$.

Length	Case numbers	Relative bias (RB)*100					
		Bias($\hat{b}^{Pooled}_{Full}$)	Bias($\hat{b}^{Pooled}_{Resp}$)	Bias($\hat{b}^{Fixone}_{Full}$)	Bias($\hat{b}^{Fixone}_{Resp}$)	Bias($\hat{b}^{Ranone}_{Full}$)	Bias($\hat{b}^{Ranone}_{Resp}$)
2	62293	-0.04 (0.10)	-30.65 (9.56)	0.05 (0.10)	0.10 (0.15)	0.002 (0.05)	-15.61 (2.52)
3	59505	-0.09 (0.10)	-30.46 (9.45)	0.05 (0.10)	-0.05 (0.10)	-0.002 (0.06)	-10.65 (1.21)
4	56869	-0.14 (0.10)	-30.33 (9.37)	0.05 (0.10)	0.03 (0.06)	-0.002 (0.06)	-8.11 (0.71)
5	54286	-0.18 (0.09)	-30.21 (9.30)	0.05 (0.10)	0.09 (0.05)	-0.001 (0.07)	-6.58 (0.49)
6	51984	-0.21 (0.09)	-30.11 (9.24)	0.05 (0.10)	0.28 (0.05)	0.001 (0.07)	-5.44 (0.35)
7	49823	-0.23 (0.09)	-30.03 (9.20)	0.05 (0.10)	0.35 (0.06)	0.002 (0.08)	-4.66 (0.27)
8	47770	-0.26 (0.08)	-29.98 (9.17)	0.05 (0.10)	0.36 (0.06)	0.004 (0.08)	-4.13 (0.23)
9	45924	-0.28 (0.08)	-29.93 (9.14)	0.05 (0.10)	0.39 (0.07)	0.007 (0.08)	-3.67 (0.20)
10	44172	-0.29 (0.08)	-29.87 (9.11)	0.05 (0.10)	0.42 (0.08)	0.009 (0.08)	-3.31 (0.17)

B. Appendix of Chapter 4

B.1. Analysis tables for wage equation

A.4. Comparison of the weighted with un-weighted cross-sectional ordered logit estimator in Scenario A-D

Table 38: Fade-away effect for the weighted and un-weighted cross-sectional ordered logit model estimators in Scenario A ($\kappa = \gamma = \rho = \phi = 0.10$) for fix non-response parameters $\alpha = -4.50, \beta = 3.20$ and attrition parameters $\alpha^{\star} = 0.90, \beta^{\star} = 0.90$.

Panel wave	Case numbers	Relative bias (RB)*100					
		Bias($\hat{b}_{Full}^{UOL_1}$)	Bias($\hat{b}_{Resp}^{UOL_1}$)	Bias($\hat{b}_{Resp}^{UOL_2}$)	Bias($\hat{b}_{Resp}^{WOL_1}$)	Bias($\hat{b}_{Resp}^{WOL_2}$)	Bias($\hat{b}_{Resp}^{WOL_3}$)
1	77308	0.00 (0.00)	-15.43 (2.70)	-15.43 (2.70)	-16.21 (3.26)	-16.21 (3.26)	1.17 (1.18)
2	71680	0.00 (0.00)	-3.70 (0.95)	-4.05 (1.12)	-0.95 (1.09)	-1.17 (1.04)	-2.78 (1.94)
3	66619	0.00 (0.00)	-1.90 (0.71)	0.11 (0.63)	2.41 (0.91)	1.94 (0.77)	2.72 (2.40)
4	62125	0.00 (0.00)	-1.74 (0.70)	-0.02 (0.63)	2.49 (0.91)	1.60 (0.71)	3.81 (2.89)
5	57953	0.00 (0.00)	-1.74 (0.69)	-0.14 (0.88)	2.14 (1.32)	1.11 (0.95)	3.64 (4.78)
6	54275	0.00 (0.00)	-1.73 (0.69)	1.65 (0.86)	3.75 (1.60)	3.08 (0.97)	2.96 (6.50)
7	50903	0.00 (0.00)	-1.73 (0.69)	1.28 (0.84)	3.94 (2.01)	2.75 (0.98)	4.53 (7.77)
8	47806	0.00 (0.00)	-1.73 (0.69)	0.12 (0.93)	2.41 (2.49)	1.38 (0.96)	2.98 (10.77)
9	45131	0.00 (0.00)	-1.73 (0.69)	0.41 (0.73)	5.17 (3.29)	1.62 (0.86)	7.76 (12.12)
10	42657	0.00 (0.00)	-1.73 (0.69)	0.53 (0.92)	5.11 (3.55)	1.76 (1.06)	9.20 (15.50)

Note: The first column represents wave of the panel. The second column refers to the number of respondents. Column third and fourth report the percent relative bias of the weighted and un-weighted ordered logit model estimators. Under this aspect "UOL" stands for the un-weighted ordered logit and "WOL" stands for the weighted ordered logit. The MSE of the estimates multiplied by 100 are given in the parenthesis.

Table 39: Fade-away effect for the weighted and un-weighted cross-sectional ordered logit model estimators in Scenario B ($\kappa = \gamma = \rho = \phi = 0.50$) for fix non-response parameters $\alpha = -4.50, \beta = 2.00$ and attrition parameters $\alpha^{\star} = 0.01, \beta^{\star} = 0.70$.

Panel wave	Case numbers	Relative bias (RB)*100					
		Bias($\hat{b}_{Full}^{UOL_1}$)	Bias($\hat{b}_{Resp}^{UOL_1}$)	Bias($\hat{b}_{Resp}^{UOL_2}$)	Bias($\hat{b}_{Resp}^{WOL_1}$)	Bias($\hat{b}_{Resp}^{WOL_2}$)	Bias($\hat{b}_{Resp}^{WOL_3}$)
1	77210	0.00 (0.00)	-15.39 (2.69)	-15.39 (2.69)	-15.79 (3.24)	-15.79 (3.24)	0.92 (1.02)
2	71605	0.00 (0.00)	-11.83 (2.03)	-12.61 (2.28)	-4.18 (1.12)	-5.46 (1.13)	3.34 (2.12)
3	66509	0.00 (0.00)	-7.46 (0.99)	13.50 (2.52)	2.29 (6.31)	20.69 (5.14)	28.40 (10.60)
4	62062	0.00 (0.00)	-5.59 (0.65)	19.52 (4.60)	2.92 (9.60)	26.30 (7.76)	32.24 (13.62)
5	57917	0.00 (0.00)	-4.79 (0.53)	20.62 (4.81)	2.90 (9.39)	26.45 (7.64)	31.20 (13.22)
6	54323	0.00 (0.00)	-4.42 (0.48)	20.42 (4.84)	2.89 (9.42)	26.02 (7.46)	31.32 (14.51)
7	50953	0.00 (0.00)	-4.25 (0.46)	20.25 (4.88)	2.75 (8.97)	25.26 (7.18)	28.35 (15.32)
8	47964	0.00 (0.00)	-4.16 (0.45)	20.55 (4.99)	2.86 (10.32)	25.83 (7.66)	28.59 (17.33)
9	45188	0.00 (0.00)	-4.12 (0.44)	20.51 (5.02)	2.83 (10.36)	25.74 (7.52)	26.02 (15.31)
10	42732	0.00 (0.00)	-4.10 (0.44)	20.04 (4.99)	2.81 (9.97)	25.15 (7.29)	25.34 (13.15)

Table 40: Fade-away effect for the weighted and un-weighted cross-sectional ordered logit model estimators in Scenario C ($\kappa = \gamma = \rho = \phi = 0.70$) for fix non-response parameters $\alpha = -4.50, \beta = 2.00$ and attrition parameters $\alpha^\star = 0.01, \beta^\star = 0.70$.

Panel wave	Case numbers	Relative bias (RB)*100					
		Bias($\hat{b}^{UOL_1}_{Full}$)	Bias($\hat{b}^{UOL_1}_{Resp}$)	Bias($\hat{b}^{UOL_2}_{Resp}$)	Bias($\hat{b}^{WOL_1}_{Resp}$)	Bias($\hat{b}^{WOL_2}_{Resp}$)	Bias($\hat{b}^{WOL_3}_{Resp}$)
1	77152	0.00 (0.00)	-15.55 (2.72)	-15.55 (2.72)	-15.73 (3.10)	-15.73 (3.10)	0.12 (0.87)
2	71502	0.00 (0.00)	-14.00 (2.55)	-14.73 (2.89)	-9.83 (2.17)	-11.06 (2.29)	1.89 (2.17)
3	66464	0.00 (0.00)	-11.24 (1.81)	13.62 (2.65)	21.09 (5.82)	18.74 (4.57)	30.46 (12.24)
4	61950	0.00 (0.00)	-9.04 (1.26)	28.83 (9.29)	38.09 (16.37)	34.81 (13.36)	46.20 (25.71)
5	57759	0.00 (0.00)	-7.61 (0.96)	35.70 (13.70)	46.22 (23.51)	41.88 (18.78)	54.47 (35.08)
6	54143	0.00 (0.00)	-6.71 (0.79)	39.13 (16.37)	50.23 (27.97)	45.34 (21.90)	54.97 (37.67)
7	50866	0.00 (0.00)	-6.13 (0.69)	40.70 (17.63)	52.23 (30.25)	46.72 (23.14)	55.20 (38.47)
8	47848	0.00 (0.00)	-5.75 (0.62)	42.18 (18.79)	54.48 (33.01)	48.15 (24.45)	55.72 (39.70)
9	45167	0.00 (0.00)	-5.49 (0.58)	43.36 (19.77)	53.50 (31.38)	48.90 (25.03)	54.12 (37.51)
10	42734	0.00 (0.00)	-5.32 (0.56)	43.42 (19.98)	53.26 (30.96)	49.02 (25.29)	52.48 (36.38)

Table 41: Fade-away effect for the weighted and un-weighted cross-sectional ordered logit model estimators in Scenario D ($\kappa = \gamma = \rho = \phi = 0.90$) for fix non-response parameters $\alpha = -4.50, \beta = 2.00$ and attrition parameters $\alpha^\star = 0.01, \beta^\star = 0.70$.

Panel wave	Case numbers	Relative bias (RB)*100					
		Bias($\hat{b}^{UOL_1}_{Full}$)	Bias($\hat{b}^{UOL_1}_{Resp}$)	Bias($\hat{b}^{UOL_2}_{Resp}$)	Bias($\hat{b}^{WOL_1}_{Resp}$)	Bias($\hat{b}^{WOL_2}_{Resp}$)	Bias($\hat{b}^{WOL_3}_{Resp}$)
1	77226	0.00 (0.00)	-16.34 (2.91)	-16.34 (2.91)	-17.00 (3.54)	-17.00 (3.54)	0.70 (0.92)
2	71604	0.00 (0.00)	-15.97 (2.83)	-17.23 (3.38)	-15.78 (3.41)	-17.11 (3.74)	1.73 (1.85)
3	66546	0.00 (0.00)	-15.32 (2.74)	-10.02 (1.65)	-7.05 (1.79)	-9.55 (1.92)	9.20 (3.51)
4	62083	0.00 (0.00)	-14.53 (2.59)	0.92 (0.77)	3.89 (1.73)	0.79 (1.11)	19.39 (7.50)
5	57937	0.00 (0.00)	-13.70 (2.39)	11.28 (2.14)	15.33 (4.05)	11.88 (2.56)	31.13 (14.14)
6	54347	0.00 (0.00)	-12.87 (2.17)	21.33 (5.46)	26.25 (9.10)	22.20 (6.11)	40.67 (22.97)
7	51006	0.00 (0.00)	-12.10 (1.96)	30.54 (10.43)	35.85 (15.27)	31.77 (11.33)	48.64 (31.51)
8	48045	0.00 (0.00)	-11.39 (1.77)	37.65 (15.42)	41.93 (19.68)	38.78 (16.32)	52.70 (34.18)
9	45308	0.00 (0.00)	-10.76 (1.60)	44.39 (20.94)	50.03 (27.58)	45.84 (22.34)	62.23 (46.42)
10	42822	0.00 (0.00)	-10.20 (1.45)	49.34 (25.64)	56.62 (35.20)	51.38 (27.79)	68.42 (55.50)

Table 42: Log earnings regression empirical results for the Full-Sample using cross-sectioanl OLS estimator.

Variable	Panel wave									
	1	2	3	4	5	6	7	8	9	10
Intercept	0.6657	0.6904	0.7350	0.7550	0.7787	0.7973	0.8034	0.8283	0.8338	0.8400
Age	0.0041	0.0044	0.0045	0.0042	0.0045	0.0043	0.0046	0.0045	0.0047	0.0048
Years of education (Edu)	0.0716	0.0685	0.0672	0.0670	0.0661	0.0657	0.0662	0.0658	0.0666	0.0674
Single (D)	-0.3500	-0.3292	-0.3021	-0.2737	-0.2408	-0.2234	-0.2080	-0.2005	-0.1923	-0.1861
Firm size 20-199 (F_3)	0.2118	0.2185	0.2142	0.2252	0.2225	0.2291	0.2285	0.2289	0.2266	0.2268
Firm size 200-1999 (F_4)	0.3422	0.3374	0.3311	0.3412	0.3296	0.3246	0.3180	0.3160	0.3118	0.3093
Firm size 2000+ (F_5)	0.3291	0.3382	0.3313	0.3504	0.3544	0.3558	0.3530	0.3547	0.3514	0.3500
Male (G)	0.2699	0.2721	0.2700	0.2704	0.2741	0.2746	0.2745	0.2754	0.2768	0.2755
Tenure	0.0055	0.0066	0.0057	0.0051	0.0045	0.0045	0.0044	0.0044	0.0044	0.0044

Table 43: Log earnings regression simulation results for the Resp-Samples using cross-sectional OLS estimator. Number of Monte Carlo replications is $R = 100$. Initial non-response rate is 30%, with no panel attrition after initial wave.

Variable	Panel wave									
	1	2	3	4	5	6	7	8	9	10
Intercept	1.0313	0.9880	0.9747	0.9716	0.9752	0.9727	0.9618	0.9754	0.9765	0.9755
Age	0.0018	0.0026	0.0029	0.0025	0.0029	0.0030	0.0034	0,0035	0.0038	0.0039
Years of education (Edu)	0.0658	0.0644	0.0652	0.0662	0.0662	0.0665	0.0676	0.0676	0.0684	0.0693
Single (D)	-0.2296	-0.2139	-0.1884	-0.1706	-0.1549	-0.1427	-0.1355	-0.1325	-0.1288	-0.1262
Firm size 20-199 (F_3)	0.1028	0.1216	0.1340	0.1540	0.1606	0.1717	0.1749	0.1761	0.1739	0.1750
Firm size 200-1999 (F_4)	0.2068	0.2157	0.2251	0.2463	0.2460	0.2503	0.2472	0.2462	0.2420	0.2405
Firm size 2000+ (F_5)	0.1851	0.2123	0.2241	0.2544	0.2655	0.2735	0.2743	0.2770	0.2749	0.2769
Male (G)	0.2187	0.2295	0.2308	0.2292	0.2285	0.2297	0.2308	0.2305	0.2317	0.2310
Tenure	0.0061	0.0068	0.0061	0.0055	0.0051	0.0049	0.0047	0.0045	0.0045	0.0046

Table 44: Fade-away effect of the initial non-response bias of the cross-sectional OLS estimator with SOEP and artificial initial non-response, with no panel attrition.

Difference of:	Bias$(\hat{b}_{p,t}) = (\hat{b}_{p,t}^{Full} - \hat{b}_{p,t}^{Resp})$ in panel wave									
	1	2	3	4	5	6	7	8	9	10
Intercept	-0.3657	-0.2977	-0.2397	-0.2166	-0.1964	-0.1754	-0.1583	-0.1471	-0.1427	-0.1355
Age	0.0024	0.0018	0.0016	0.0016	0.0016	0.0013	0.0012	0.0010	0.0009	0.0009
Years of education (Edu)	0.0059	0.0041	0.0020	0.0008	-0.0001	-0.0007	-0.0014	-0.0019	-0.0018	-0.0019
Single (D)	-0.1204	-0.1153	-0.1138	-0.1031	-0.0860	-0.0807	-0.0725	-0.0679	-0.0635	-0.0599
Firm size 20-199 (F_3)	0.1090	0.0969	0.0802	0.0712	0.0619	0.0574	0.0537	0.0527	0.0528	0.0518
Firm size 200-1999 (F_4)	0.1354	0.1217	0.1060	0.0949	0.0836	0.0743	0.0708	0.0698	0.0698	0.0688
Firm size 2000+ (F_5)	0.1439	0.1259	0.1072	0.0959	0.0889	0.0823	0.0787	0.0776	0.0765	0.0731
Male (G)	0.0512	0.0426	0.0391	0.0412	0.0456	0.0449	0.0437	0.0449	0.0452	0.0445
Tenure	-0.0006	-0.0002	-0.0004	-0.0004	-0.0006	-0.0004	-0.0003	-0.0001	-0.0002	-0.0002

Table 45: Empirical results for the Full-Sample, using cross-sectional OLS estimator with lagged $W_{i,t-1}$.

Variable	Panel wave								
	2	3	4	5	6	7	8	9	10
Intercept	0.4777	0.5162	0.5257	0.5244	0.5417	0.5204	0.5354	0.5170	0.5128
Lag hourly wage	0.4205	0.3952	0.3943	0.4217	0.4161	0.4386	0.4508	0.4692	0.4810
Age	0.0024	0.0025	0.0018	0.0021	0.0018	0.0020	0.0018	0.0018	0.0018
Years of education (Edu)	0.0368	0.0399	0.0410	0.0380	0.0385	0.0380	0.0365	0.0363	0.0360
Single (D)	-0.1401	-0.1237	-0.1150	-0.0882	-0.0873	-0.0765	-0.0727	-0.0653	-0.0619
Firm size 20-199 (F_3)	0.1131	0.1237	0.1379	0.1329	0.1385	0.1294	0.1248	0.1188	0.1174
Firm size 200-1999 (F_4)	0.1687	0.1874	0.2051	0.1869	0.1848	0.1706	0.1658	0.1569	0.1529
Firm size 2000+ (F_5)	0.1879	0.1935	0.2197	0.2120	0.2112	0.1998	0.1955	0.1849	0.1814
Male (G)	0.1520	0.1518	0.1519	0.1495	0.1503	0.1443	0.1417	0.1385	0.1344
Tenure	0.0049	0.0029	0.0023	0.0015	0.0018	0.0016	0.0014	0.0014	0.0014

Table 46: Simulation results for the Resp-Samples, using cross-sectional OLS estimator with lagged $W_{i,t-1}$. Number of Monte Carlo replications is $R = 100$. Initial non-response rate is 30%, with no panel attrition after initial wave.

Variable	Panel wave								
	2	3	4	5	6	7	8	9	10
Intercept	0.6157	0.6504	0.6523	0.6471	0.6503	0.6148	0.6221	0.6004	0.5860
Lag hourly wage	0.3032	0.3051	0.3210	0.3472	0.3489	0.3764	0.3969	0.4201	0.4380
Age	0.0029	0.0024	0.0016	0.0020	0.0019	0.0021	0.0019	0.0019	0.0019
Years of education (Edu)	0.0441	0.0450	0.0452	0.0428	0.0431	0.0425	0.0406	0.0397	0.0392
Single (D)	-0.1144	-0.0806	-0.0733	-0.0614	-0.0580	-0.0530	-0.0503	-0.0463	-0.0451
Firm size 20-199 (F_3)	0.1115	0.1134	0.1272	0.1243	0.1272	0.1217	0.1160	0.1087	0.1057
Firm size 200-1999 (F_4)	0.1634	0.1757	0.1913	0.1771	0.1754	0.1637	0.1572	0.1468	0.1411
Firm size 2000+ (F_5)	0.1830	0.1849	0.2084	0.2013	0.1995	0.1884	0.1822	0.1715	0.1680
Male (G)	0.1725	0.1685	0.1574	0.1490	0.1489	0.1435	0.1370	0.1327	0.1275
Tenure	0.0059	0.0042	0.0034	0.0028	0.0027	0.0024	0.0021	0.0021	0.0021

Table 47: Cross-sectional OLS estimator with lagged $W_{i,t-1}$: Fade-away effect of the initial non-response bias with SOEP and artificial initial non-response, with no panel attrition.

Difference of:	Bias$(\hat{b}_{p,t}) = (\hat{b}_{p,t}^{Full} - \hat{b}_{p,t}^{Resp})$ in panel wave								
	2	3	4	5	6	7	8	9	10
Intercept	-0.1380	-0.1342	-0.1266	-0.1227	-0.1086	-0.0944	-0.0867	-0.0834	-0.0732
Lag hourly wage	0.1172	0.0901	0.0733	0.0745	0.0671	0.0623	0.0539	0.0491	0.0430
Age	-0.0005	-0.0001	0.0002	0.0001	-0.0001	-0.0002	-0.0002	-0.0002	-0.0002
Years of education (Edu)	-0.0073	-0.0052	-0.0042	-0.0048	-0.0045	-0.0045	-0.0041	-0.0035	-0.0033
Single (D)	0.0257	-0.0431	-0.0416	-0.0268	-0.0293	-0.0235	-0.0224	-0.0190	-0.0167
Firm size 20-199 (F_3)	0.0017	0.0104	0.0107	0.0086	0.0113	0.0077	0.0087	0.0101	0.0118
Firm size 200-1999 (F_4)	0.0054	0.0116	0.0138	0.0098	0.0094	0.0069	0.0087	0.0101	0.0118
Firm size 2000+ (F_5)	0.0049	0.0085	0.0114	0.0107	0.0117	0.0113	0.0133	0.0134	0.0134
Male (G)	-0.0205	-0.0167	-0.0055	0.0005	0.0015	0.0009	0.0047	0.0058	0.0069
Tenure	0.0010	-0.0013	-0.0011	-0.0013	-0.0009	-0.0008	-0.0007	-0.0007	-0.0007

Table 48: Log earnings regression empirical results for the Full-Sample using OLS estimator for cross-sectional data.

Variable	Panel wave									
	1	2	3	4	5	6	7	8	9	10
Intercept	0.6657	0.6904	0.7350	0.7550	0.7787	0.7974	0.8034	0.8283	0.8338	0.8400
Age	0.0041	0.0044	0.0045	0.0042	0.0045	0.0043	0.0046	0.0045	0.0047	0.0048
Years of education (Edu)	0.0716	0.0685	0.0672	0.0670	0.0661	0.0657	0.0662	0.0658	0.0666	0.0674
Single (D)	-0.3500	-0.3292	-0.3021	-0.2737	-0.2408	-0.2234	-0.2080	-0.2005	-0.1923	-0.1861
Firm size 20-199 (F_3)	0.2118	0.2185	0.2142	0.2252	0.2225	0.2291	0.2285	0.2289	0.2266	0.2268
Firm size 200-1999 (F_4)	0.3422	0.3374	0.3311	0.3412	0.3296	0.3246	0.3180	0.3160	0.3118	0.3093
Firm size 2000+ (F_5)	0.3291	0.3382	0.3313	0.3504	0.3544	0.3558	0.3530	0.3547	0.3514	0.3500
Male (G)	0.2699	0.2721	0.2700	0.2704	0.2741	0.2746	0.2745	0.2754	0.2768	0.2755
Tenure	0.0055	0.0066	0.0057	0.0051	0.0045	0.0045	0.0044	0.0044	0.0044	0.0044

Table 49: Log earnings regression simulation results for the Resp-Samples using IPW estimator for cross-sectional data. Number of Monte Carlo replications is $R = 100$. Initial non-response rate is 30%, with no panel attrition after initial wave.

Variable	Panel wave									
	1	2	3	4	5	6	7	8	9	10
Intercept	0.6900	0.7000	0.7487	0.7894	0.8120	0.8213	0.8143	0.8400	0.8481	0.8520
Age	0.0038	0.0043	0.0042	0.0035	0.0038	0.0038	0.0041	0.0042	0.0044	0.0045
Years of education (Edu)	0.0706	0.0692	0.0680	0.0669	0.0663	0.0664	0.0676	0.0672	0.0679	00688
Single (D)	-0.2936	-0.2729	-0.2326	-0.2090	-0.1845	-0.1692	-0.1567	-0.1520	-0.1533	-0.1490
Firm size 20-199 (F_3)	0.2184	0.2127	0.2106	0.2333	0.2317	0.2377	0.2391	0.2374	0.2276	0.2270
Firm size 200-1999 (F_4)	0.3285	0.3176	0.3158	0.3413	0.3327	0.3322	0.3277	0.3256	0.3184	0.3150
Firm size 2000+ (F_5)	0.3144	0.3163	0.3158	0.3487	0.3520	0.3551	0.3533	0.3548	0.3496	0.3500
Male (G)	0.2624	0.2663	0.2656	0.2630	0.2620	0.2623	0.2626	0.2608	0.2638	0.2629
Tenure	0.0066	0.0070	0.0060	0.0049	0.0044	0.0044	0.0042	0.0039	0.0040	0.0041

Table 50: Fade-away effect of the initial non-response bias of the IPW estimator for cross-sectional data, with SOEP and artificial initial non-response, with no panel attrition.

Difference of:	Bias$(\hat{b}_{p,t}) = (\hat{b}_{p,t}^{Full} - \hat{b}_{p,t}^{Resp})$ in panel wave									
	1	2	3	4	5	6	7	8	9	10
Intercept	-0.0243	-0.0096	-0.0136	-0.0344	-0.0333	-0.0239	-0.0109	-0.0116	-0.0143	-0.0120
Age	0.0004	0.0001	0.0003	0.0007	0.0007	0.0005	0.0004	0.0003	0.0003	0.0003
Years of education (Edu)	0.0011	-0.0007	-0.0008	0.00004	-0.0002	-0.0007	-0.0014	-0.0015	-0.0013	-0.0014
Single (D)	-0.0564	-0.0563	-0.0696	-0.0647	-0.0564	-0.0542	-0.0513	-0.0485	-0.0390	-0.0371
Firm size 20-199 (F_3)	-0.0066	0.0058	0.0036	-0.0081	-0.0092	-0.0086	-0.0106	-0.0085	-0.0010	-0.0002
Firm size 200-1999 (F_4)	0.0137	0.0198	0.0153	-0.0001	-0.0031	-0.0076	-0.0097	-0.0096	-0.0066	-0.0057
Firm size 2000+ (F_5)	0.0147	0.0219	0.0156	0.0017	0.0024	0.0007	-0.0003	-0.0002	0.0018	0.00004
Male (G)	0.0075	0.0058	0.0044	0.0074	0.0121	0.0123	0.0119	0.0146	0.0130	0.0126
Tenure	-0.0011	-0.0005	-0.0003	0.0002	0.0001	0.0001	0.0002	0.0005	0.0003	0.0003

Table 51: Empirical results for the Full-Sample, using OLS estimator with lagged $W_{i,t-1}$ for cross-sectional data.

Variable	Panel wave								
	2	3	4	5	6	7	8	9	10
Intercept	0.4777	0.5162	0.5257	0.5244	0.5417	0.5204	0.5354	0.5170	0.5128
Lag hourly wage	0.4205	0.3952	0.3943	0.4217	0.4161	0.4386	0.4508	0.4692	0.4810
Age	0.0024	0.0025	0.0018	0.0021	0.0018	0.0020	0.0018	0.0018	0.0018
Years of education (Edu)	0.0368	0.0399	0.0410	0.0380	0.0385	0.0380	0.0365	0.0363	0.0359
Single (D)	-0.1401	-0.1237	-0.1150	-0.0882	-0.0873	-0.0765	-0.0727	-0.0653	-0.0619
Firm size 20-199 (F_3)	0.1131	0.1237	0.1379	0.1329	0.1385	0.1294	0.1248	0.1188	0.1174
Firm size 200-1999 (F_4)	0.1687	0.1874	0.2051	0.1869	0.1848	0.1706	0.1658	0.1569	0.1529
Firm size 2000+ (F_5)	0.1879	0.1935	0.2197	0.2119	0.2112	0.1998	0.1955	0.1849	0.1814
Male (G)	0.1520	0.1518	0.1519	0.1495	0.1503	0.1443	0.1417	0.1385	0.1344
Tenure	0.0049	0.0029	0.0023	0.0015	0.0018	0.0016	0.0014	0.0014	0.0014

Table 52: Simulation results for the Resp-Samples, using IPW estimator with lagged $W_{i,t-1}$ for cross-sectional data. Number of Monte Carlo replications is $R = 100$. Initial non-response rate is 30%, with no panel attrition after initial wave.

Variable	Panel wave								
	2	3	4	5	6	7	8	9	10
Intercept	0.4897	0.5555	0.5768	0.5793	0.5867	0.5552	0.5688	0.5532	0.5433
Lag hourly wage	0.3912	0.3628	0.5657	0.3906	0.3859	0.4097	0.4271	0.4482	0.4626
Age	0.0027	0.0023	0.0015	0.0018	0.0016	0.0019	0.0017	0.0018	0.0017
Years of education (Edu)	0.0397	0.0417	0.0425	0.0397	0.0405	0.0402	0.0384	0.0376	0.0373
Single (D)	-0.1284	-0.0907	-0.0820	-0.0654	-0.0626	-0.0535	-0.0516	-0.0457	-0.0442
Firm size 20-199 (F_3)	0.1149	0.1228	0.1359	0.1336	0.1372	0.1324	0.1251	0.1161	0.1138
Firm size 200-1999 (F_4)	0.1661	0.1859	0.2017	0.1877	0.1873	0.1764	0.1696	0.1577	0.1528
Firm size 2000+ (F_5)	0.1869	0.1964	0.2191	0.2119	0.2109	0.2005	0.1943	0.1823	0.1796
Male (G)	0.1567	0.1640	0.1579	0.1510	0.1524	0.1473	0.1405	0.1362	0.1318
Tenure	0.0055	0.0035	0.0029	0.0022	0.0024	0.0021	0.0018	0.0018	0.0018

Table 53: IPW estimator with lagged $W_{i,t-1}$ for cross-sectional data: Fade-away effect of the initial non-response bias with SOEP and artificial initial non-response, with no panel attrition.

Difference of:	Bias$(\hat{b}_{p,t}) = (\hat{b}_{p,t}^{Full} - \hat{b}_{p,t}^{Resp})$ in panel wave								
	2	3	4	5	6	7	8	9	10
Intercept	-0.0120	-0.0394	-0.0511	-0.0548	-0.0450	-0.0348	-0.0334	-0.0362	-0.0305
Lag hourly wage	0.0292	0.0324	0.0286	0.0311	0.0301	0.0290	0.0236	0.0211	0.0183
Age	-0.0003	0.0001	0.0003	0.0003	0.0001	0.0001	0.0001	0.0001	0.0001
Years of education (Edu)	-0.0029	-0.0019	-0.0014	-0.0017	-0.0020	-0.0022	-0.0019	-0.0013	-0.0014
Single (D)	-0.0117	-0.0330	-0.0329	-0.0229	-0.0247	-0.0230	-0.0211	-0.0196	-0.0176
Firm size 20-199 (F_3)	-0.0018	0.0010	0.0020	-0.0007	0.0013	-0.0030	-0.0003	0.0026	0.0036
Firm size 200-1999 (F_4)	0.0027	0.0015	0.0035	-0.0007	-0.0025	-0.0058	-0.0037	-0.0008	0.0001
Firm size 2000+ (F_5)	0.0010	-0.0030	0.0007	0.0001	0.0003	-0.0008	0.0012	0.0026	0.0018
Male (G)	-0.0047	-0.0122	-0.0060	-0.0016	-0.0021	-0.0030	0.0012	0.0023	0.0026
Tenure	-0.0006	-0.0006	-0.0006	-0.0006	-0.0006	-0.0005	-0.0003	-0.0004	-0.0004

Table 54: Empirical results for the Full-Sample, using RE model estimator.

Variable	Length									
	1	2	3	4	5	6	7	8	9	10
Intercept	0.6657	0.6657	0.6750	0.6647	0.6577	0.6797	0.6867	0.7112	0.7082	0.7168
Age	0.0041	0.0043	0.0048	0.0047	0.0049	0.0046	0.0044	0.0040	0.0037	0.0035
Years of education (Edu)	0.0716	0.0680	0.0665	0.0671	0.0667	0.668	0.0672	0.0669	0.0682	0.0689
Single (D)	-0.3500	-0.3271	-0.2891	-0.2567	-0.2178	-0.2037	-0.1957	-0.1956	-0.1819	-0.1907
Firm size 20-199 (F_3)	0.2118	0.2175	0.2044	0.2027	0.1905	0.1730	0.1598	0.1483	0.1454	0.1392
Firm size 200-1999 (F_4)	0.3422	0.3371	0.3251	0.3287	0.3110	0.2685	0.2469	0.2270	0.2169	0.2061
Firm size 2000+ (F_5)	0.3291	0.3389	0.3172	0.3266	0.3330	0.2978	0.2776	0.2629	0.2534	0.2430
Male (G)	0.2699	0.2762	0.2783	0.2808	0.2901	0.2969	0.3034	0.3140	0.3183	0.3216
Tenure	0.0055	0.0069	0.0058	0.0052	0.0044	0.0047	0.0045	0.0043	0.0042	0.0039

Table 55: Simulation results for the Resp-Samples, using RE model estimator. Number of Monte Carlo replications is $R = 100$. Initial non-response rate is 30%, with no panel attrition after intial wave.

Variable	Length									
	1	2	3	4	5	6	7	8	9	10
Intercept	1.0413	0.9901	0.9779	0.9695	0.9590	0.9617	0.9516	0.9726	0.9763	0.9809
Age	0.0016	0.0024	0.0027	0.0024	0.0027	0.0026	0.0027	0.0025	0.0023	0.0021
Years of education (Edu)	0.0659	0.0647	0.0647	0.0657	0.0656	0.0600	0.0671	0.0672	0.0680	0.0688
Single (D)	-0.2201	-0.2041	-0.1661	-0.1422	-0.1200	-0.1088	-0.1031	-0.1045	-0.1021	-0.1062
Firm size 20-199 (F_3)	0.1029	0.1193	0.1204	0.1316	0.1330	0.1340	0.1294	0.1200	0.1165	0.1137
Firm size 200-1999 (F_4)	0.2063	0.2126	0.2164	0.2319	0.2262	0.2111	0.1981	0.1844	0.1739	0.1675
Firm size 2000+ (F_5)	0.1841	0.2076	0.2076	0.2301	0.2388	0.2259	0.2124	0.2023	0.1939	0.1903
Male (G)	0.2119	0.2243	0.2305	0.2285	0.2299	0.2340	0.2397	0.2419	0.2447	0.2457
Tenure	0.0063	0.0071	0.0063	0.0058	0.0050	0.0049	0.0045	0.0039	0.0039	0.0038

Table 56: RE model estimator: Fade-away effect of the initial non-response bias with SOEP and artificial initial non-response, with no panel attrition.

Difference of:	$\text{Bias}(\hat{b}_{p,t}) = (\hat{b}_{p,t}^{Full} - \hat{b}_{p,t}^{Resp})$ in length									
	1	2	3	4	5	6	7	8	9	10
Intercept	-0.3756	-0.3243	-0.3029	-0.3047	-0.3013	-0.2820	-0.2649	-0.2614	-0.2681	-0.2640
Age	0.0026	0.0019	0.0021	0.0023	0.0022	-0.0020	-0.0017	-0.0014	0.0014	0.0014
Years of education (Edu)	0.0057	0.0034	0.0018	0.0014	0.0012	0.0008	0.0001	-0.0003	0.0003	0.0010
Single (D)	-0.1299	-0.1230	-0.1230	-0.1145	-0.0978	-0.0949	-0.0926	-0.0910	-0.0869	-0.0845
Firm size 20-199 (F_3)	0.1089	0.0982	0.0840	0.0711	0.0575	0.0390	0.0304	0.0283	0.0289	0.0255
Firm size 200-1999 (F_4)	0.1358	0.1246	0.1088	0.0969	0.0848	0.0574	0.0488	0.0426	0.0429	0.0386
Firm size 2000+ (F_5)	0.1449	0.1313	0.1096	0.0965	0.0942	0.0719	0.0652	0.0606	0.0596	0.0528
Male (G)	0.0580	0.0519	0.0478	0.0524	0.0603	0.0629	0.0638	0.0721	0.0736	0.0759
Tenure	-0.0008	-0.0001	-0.0005	-0.0005	-0.0006	-0.0003	-0.00003	0.0003	0.0003	0.0001

Table 57: Empirical results for the Full-Sample, using RE model estimator with auto-correlated errors.

Variable	Length									
	1	2	3	4	5	6	7	8	9	10
Age	0.0041	0.0043	0.0045	0.0042	0.0045	0.0040	0.0041	0.0035	0.0034	0.0031
Years of education (Edu)	0.0716	0.0680	0.0667	0.0666	0.0651	0.0647	0.0650	0.0638	0.0648	0.0650
Single (D)	-0.3500	-0.3271	-0.2938	-0.2635	-0.2242	-0.2130	-0.1977	-0.1933	-0.1840	-0.1792
Firm size 20-199 (F_3)	0.2118	0.2175	0.2074	0.2147	0.2079	0.1957	0.1912	0.1832	0.1770	0.1731
Firm size 200-1999 (F_4)	0.3422	0.3371	0.3262	0.3339	0.3156	0.2859	0.2743	0.2622	0.2531	0.2448
Firm size 2000+ (F_5)	0.3291	0.3389	0.3229	0.3422	0.3441	0.3205	0.3144	0.3044	0.2969	0.2893
Male (G)	0.2699	0.2762	0.2743	0.2766	0.2841	0.2862	0.2881	0.2937	0.2975	0.2974
Tenure	0.0055	0.0069	0.0057	0.0053	0.0047	0.0053	0.0053	0.0054	0.0054	0.0055

Table 58: Simulation results for the Resp-Samples, using RE model estimator with auto-correlated errors. Number of Monte Carlo replications is $R = 100$. Initial non-response rate is 30%, with no panel attrition after intial wave.

Variable	Length									
	1	2	3	4	5	6	7	8	9	10
Age	0.0016	0.0024	0.0025	0.0022	0.0026	0.0025	0.0027	0.0024	0.0023	0.0020
Years of education (Edu)	0.0659	0.0647	0.0648	0.0657	0.0650	0.0652	0.0666	0.0659	0.0666	0.0647
Single (D)	-0.2201	-0.2041	-0.1686	-0.1513	-0.1336	-0.1240	-0.1169	-0.1143	-0.1092	-0.1076
Firm size 20-199 (F_3)	0.1029	0.1193	0.1226	0.1411	0.1434	0.1477	0.1491	0.1444	0.1392	0.1386
Firm size 200-1999 (F_4)	0.2063	0.2126	0.2182	0.2384	0.2308	0.2258	0.2201	0.2136	0.2058	0.2030
Firm size 2000+ (F_5)	0.1841	0.2076	0.2126	0.2438	0.2494	0.2471	0.2460	0.2419	0.2372	0.2376
Male (G)	0.2119	0.2243	0.2286	0.2256	0.2260	0.2284	0.2317	0.2319	0.2351	0.2343
Tenure	0.0063	0.0071	0.0061	0.0056	0.0050	0.0050	0.0047	0.0044	0.0044	0.0044

Table 59: RE model estimator with auto-correlated errors: Fade-away effect of the initial non-response bias with SOEP and artificial initial non-response, with no panel attrition.

Difference of:	$\text{Bias}(\hat{b}_{p,t}) = (\hat{b}_{p,t}^{Full} - \hat{b}_{p,t}^{Resp})$ in length									
	1	2	3	4	5	6	7	8	9	10
Age	0.0026	0.0019	0.0020	0.0020	0.0019	0.0015	0.0014	0.0011	0.0011	0.0011
Years of education (Edu)	0.0057	0.0034	0.0019	0.0009	0.0001	-0.0005	-0.0016	-0.0022	-0.0019	-0.0024
Single (D)	-0.1299	-0.1230	-0.1252	-0.1121	-0.0907	-0.0889	-0.0808	-0.0791	-0.0749	-0.0716
Firm size 20-199 (F_3)	0.1089	0.0982	0.0848	0.0736	0.0645	0.0480	0.0421	0.0388	0.0378	0.0345
Firm size 200-1999 (F_4)	0.1358	0.1246	0.1080	0.0956	0.0849	0.0601	0.0542	0.0486	0.0474	0.0418
Firm size 2000+ (F_5)	0.1449	0.1313	0.1103	0.0984	0.0947	0.0734	0.0684	0.0625	0.0597	0.0517
Male (G)	0.0580	0.0519	0.0458	0.0511	0.0582	0.0578	0.0565	0.0618	0.0624	0.0632
Tenure	-0.0008	-0.0001	-0.0005	-0.0003	-0.0003	0.0003	0.0006	0.0010	0.0010	0.0011

Table 60: Empirical results for the Full-Sample, using RE model estimator with lagged $W_{i,t-1}$.

Variable	Length								
	2	3	4	5	6	7	8	9	10
Intercept	0.4958	0.5475	0.5698	0.5478	0.5855	0.5634	0.5841	0.5626	0.5648
Lag hourly wage	0.1649	0.1409	0.1226	0.1385	0.1225	0.1408	0.1438	0.1592	0.1630
Age	0.0037	0.0039	0.0036	0.0039	0.0035	0.0034	0.0032	0.0028	0.0026
Years of education (Edu)	0.0563	0.0573	0.0593	0.0573	0.0587	0.0585	0.0575	0.0575	0.0578
Single (D)	-0.2686	-0.2315	-0.2167	-0.1825	-0.1777	-0.1660	-0.1629	-0.1536	-0.1536
Firm size 20-199 (F_3)	0.1702	0.1723	0.1826	0.1752	0.1649	0.1469	0.1362	0.1331	0.1262
Firm size 200-1999 (F_4)	0.2619	0.2675	0.2851	0.2677	0.2408	0.2168	0.1997	0.1902	0.1797
Firm size 2000+ (F_5)	0.2784	0.2740	0.2987	0.2986	0.2746	0.2530	0.2389	0.2273	0.2169
Male (G)	0.2504	0.2396	0.2411	0.2443	0.2510	0.2504	0.2557	0.2557	0.2576
Tenure	0.0070	0.0051	0.0042	0.0031	0.0034	0.0031	0.0027	0.0028	0.0027

Table 61: Simulation results for the Resp-Samples, using RE model estimator with lagged $W_{i,t-1}$. Number of Monte Carlo replications is $R = 100$. Initial non-response rate is 30%, with no panel attrition after initial wave.

Variable	Length								
	2	3	4	5	6	7	8	9	10
Intercept	0.8435	0.8908	0.8932	0.8728	0.8952	0.8611	0.8727	0.8609	0.8553
Lag hourly wage	0.1010	0.0675	0.0635	0.0778	0.0639	0.0827	0.0905	0.1011	0.1074
Age	0.0023	0.0025	0.0021	0.0024	0.0024	0.0025	0.0023	0.0021	0.0019
Years of education (Edu)	0.0578	0.0600	0.0614	0.0602	0.0617	0.0617	0.0611	0.0612	0.0618
Single (D)	-0.1915	-0.1488	-0.1303	-0.1119	-0.1050	-0.1001	-0.1002	-0.0974	-0.1005
Firm size 20-199 (F_3)	0.1172	0.1241	0.1377	0.1370	0.1332	0.1253	0.1156	0.1125	0.1082
Firm size 200-1999 (F_4)	0.1937	0.2084	0.2266	0.2167	0.2013	0.1865	0.1715	0.1619	0.1546
Firm size 2000+ (F_5)	0.2034	0.2135	0.2395	0.2415	0.2249	0.2083	0.1958	0.1863	0.1812
Male (G)	0.2246	0.2252	0.2209	0.2162	0.2204	0.2202	0.2202	0.2197	0.2188
Tenure	0.0080	0.0064	0.0055	0.0045	0.0044	0.0039	0.0032	0.0031	0.0030

Table 62: RE model estimator with lagged $W_{i,t-1}$: Fade-away effect of the initial non-response bias with SOEP and artificial initial non-response, with no panel attrition.

Difference of:	Bias$(\hat{b}_{p,t}) = (\hat{b}_{p,t}^{Full} - \hat{b}_{p,t}^{Resp})$ in length								
	2	3	4	5	6	7	8	9	10
Intercept	-0.3477	-0.3433	-0.3233	-0.3250	-0.3097	-0.2977	-0.2886	-0.2983	-0.2905
Lag hourly wage	0.0640	0.0734	0.0591	0.0607	0.0586	0.0581	0.0534	0.0581	0.0557
Age	0.0014	0.0014	0.0014	0.0014	0.0011	0.0010	0.0009	0.0007	0.0007
Years of education (Edu)	-0.0015	-0.0027	-0.0021	-0.0029	-0.0030	-0.0032	-0.0036	-0.0037	-0.0040
Single (D)	-0.0772	-0.0827	-0.0834	-0.0706	-0.0727	-0.0659	-0.0627	-0.0562	-0.0531
Firm size 20-199 (F_3)	0.0530	0.0481	0.0450	0.0383	0.0318	0.0216	0.0206	0.0207	0.0186
Firm size 200-1999 (F_4)	0.0682	0.0591	0.0585	0.0510	0.0395	0.0303	0.0282	0.0283	0.0251
Firm size 2000+ (F_5)	0.0751	0.0606	0.0592	0.0570	0.0500	0.0447	0.0430	0.0410	0.0357
Male (G)	0.0258	0.0144	0.0202	0.0282	0.0306	0.0302	0.0355	0.0360	0.0388
Tenure	-0.0010	-0.0013	-0.0013	-0.0014	-0.0010	-0.0008	-0.0005	-0.0003	-0.0004

Table 63: Empirical results for the Full-Sample, using FE Within estimator.

Variable	Length								
	2	3	4	5	6	7	8	9	10
Years of education (Edu)	0.0323	0.0852	0.1695	0.2097	0.2291	0.2339	0.2280	0.2027	0.1959
Single (D)	-0.0913	0.0152	-0.0547	-0.0970	-0.1257	-0.1410	-0.1564	-0.1681	-0.1814
Firm size 20-199 (F_3)	-0.1029	0.0062	0.0258	0.0118	0.0504	0.0498	0.0528	0.0608	0.0605
Firm size 200-1999 (F_4)	-0.0140	0.0900	0.1531	0.1493	0.1056	0.1007	0.0993	0.1046	0.0997
Firm size 2000+ (F_5)	0.0517	-0.0288	0.0588	0.0818	0.0886	0.0846	0.0963	0.1028	0.1049
Tenure	0.0116	0.0090	0.0050	-0.0014	-0.0010	-0.0003	-0.0015	-0.0010	-0.0016

Table 64: Simulation results for the Resp-Samples, using FE Within estimator. Number of Monte Carlo replications is $R = 100$. Initial non-response rate is 30%, with no panel attrition after initial wave.

Variable	Length								
	2	3	4	5	6	7	8	9	10
Years of education (Edu)	-0.1300	-0.0371	0.0401	0.0804	0.1066	0.1143	0.1187	0.1078	0.1109
Single (D)	-0.0186	0.0238	-0.0031	-0.0373	-0.0751	-0.0891	-0.1056	-0.1189	-0.1348
Firm size 20-199 (F_3)	-0.1411	-0.0293	-0.0031	-0.0085	0.0558	0.0568	0.0557	0.0637	0.0617
Firm size 200-1999 (F_4)	0.0573	0.0873	0.1311	0.1300	0.1026	0.0980	0.0914	0.0968	0.0911
Firm size 2000+ (F_5)	0.0044	-0.0813	0.0082	0.0437	0.0787	0.0744	0.0835	0.0910	0.0933
Tenure	0.0132	0.0122	0.0092	0.0031	0.00286	0.0030	0.0017	0.0022	0.0015

Table 65: FE Within estimator: Fade-away effect of the initial non-response bias with SOEP and artificial initial non-response, with no panel attrition.

Difference of:	Bias$(\hat{b}_{p,t}) = (\hat{b}_{p,t}^{Full} - \hat{b}_{p,t}^{Resp})$ in length								
	2	3	4	5	6	7	8	9	10
Years of education (Edu)	0.1623	0.1220	0.1294	0.1293	0.1225	0.1196	0.1093	0.0949	0.0850
Single (D)	-0.0727	-0.0087	-0.0516	-0.0597	-0.0506	-0.0518	-0.0508	-0.0492	-0.0466
Firm size 20-199 (F_3)	0.0382	0.0355	0.0289	0.0203	-0.0054	-0.0070	-0.0029	-0.0029	-0.0013
Firm size 200-1999 (F_4)	-0.0713	0.0028	0.0219	0.0193	0.0030	0.0028	0.0078	0.0078	0.0085
Firm size 2000+ (F_5)	0.0473	0.0525	0.0506	0.0381	0.0099	0.0102	0.0127	0.0118	0.0116
Tenure	-0.0016	-0.0033	-0.0042	-0.0044	-0.0038	-0.0033	-0.0032	-0.0031	-0.0031

Table 66: Empirical results for the Full-Sample, using OLS estimator with RE.

Variable	Length									
	1	2	3	4	5	6	7	8	9	10
Intercept	0.6657	0.6657	0.6750	0.6647	0.6577	0.6797	0.6867	0.7112	0.7082	0.7168
Age	0.0041	0.0043	0.0048	0.0047	0.0049	0.0046	0.0044	0.0040	0.0037	0.0035
Years of education (Edu)	0.0716	0.0680	0.0665	0.0671	0.0667	0.668	0.0672	0.0669	0.0682	0.0689
Single (D)	-0.3500	-0.3271	-0.2891	-0.2567	-0.2178	-0.2037	-0.1957	-0.1956	-0.1819	-0.1907
Firm size 20-199 (F_3)	0.2118	0.2175	0.2044	0.2027	0.1905	0.1730	0.1598	0.1483	0.1454	0.1392
Firm size 200-1999 (F_4)	0.3422	0.3371	0.3251	0.3287	0.3110	0.2685	0.2469	0.2270	0.2169	0.2061
Firm size 2000+ (F_5)	0.3291	0.3389	0.3172	0.3266	0.3330	0.2978	0.2776	0.2629	0.2534	0.2430
Male (G)	0.2699	0.2762	0.2783	0.2808	0.2901	0.2969	0.3034	0.3140	0.3183	0.3216
Tenure	0.0055	0.0069	0.0058	0.0052	0.0044	0.0047	0.0045	0.0043	0.0042	0.0039

Table 67: Simulation results for the Resp-Samples, using IPW estimator with RE. Number of Monte Carlo replications is $R = 100$. Initial non-response rate is 30%, with no panel attrition after initial wave.

Variable	Length									
	1	2	3	4	5	6	7	8	9	10
Intercept	1.0017	0.8897	0.8791	0.8732	0.8610	0.8641	0.8536	0.8704	0.8702	0.8706
Age	0.0018	0.0029	0.0032	0.0028	0.0031	0.0030	0.0030	0.0029	0.0027	0.0025
Years of education (Edu)	0.0663	0.0655	0.0654	0.0667	0.0668	0.0675	0.0688	0.0690	0.0699	0.0710
Single (D)	-0.2288	-0.2253	-0.1773	-0.1537	-0.1342	-0.1235	-0.1193	-0.1233	-0.1208	-0.1251
Firm size 20-199 (F_3)	0.1151	0.1452	0.1370	0.1437	0.1473	0.1446	0.1381	0.1326	0.1295	0.1271
Firm size 200-1999 (F_4)	0.2175	0.2349	0.2314	0.2458	0.2401	0.2237	0.2105	0.2006	0.1910	0.1858
Firm size 2000+ (F_5)	0.1977	0.2371	0.2285	0.2473	0.2563	0.2427	0.2309	0.2255	0.2188	0.2155
Male (G)	0.2186	0.2367	0.2404	0.2369	0.2371	0.2396	0.2439	0.2452	0.2473	0.2479
Tenure	0.0064	0.0072	0.0061	0.0054	0.0044	0.0043	0.0039	0.0032	0.0031	0.0030

Table 68: IPW estimator with RE: Fade-away effect of the initial non-response bias with SOEP and artificial initial non-response, with no panel attrition.

Difference of:	$\text{Bias}(\hat{b}_{p,t}) = (\hat{b}_{p,t}^{Full} - \hat{b}_{p,t}^{Resp})$ in length									
	1	2	3	4	5	6	7	8	9	10
Intercept	-0.3360	-0.2240	-0.2041	-0.2085	-0.2033	-0.1844	-0.1669	-0.1592	-0.1620	-0.1538
Age	0.0023	0.0015	0.0016	0.0019	0.0018	0.0016	0.0014	0.0011	0.0010	0.0010
Years of education (Edu)	0.0054	0.0025	0.0011	0.0004	-0.0001	-0.0007	-0.0016	-0.0022	-0.0017	-0.0021
Single (D)	-0.1212	-0.1018	-0.1118	-0.1030	-0.0836	-0.0802	-0.0764	-0.0723	-0.0682	-0.0657
Firm size 20-199 (F_3)	0.0967	0.0723	0.0674	0.0590	0.0432	0.0285	0.0217	0.0157	0.0159	0.0121
Firm size 200-1999 (F_4)	0.1247	0.1023	0.0937	0.0830	0.0709	0.0449	0.0364	0.0264	0.0258	0.0203
Firm size 2000+ (F_5)	0.1313	0.1018	0.0887	0.0794	0.0767	0.0551	0.0467	0.0373	0.0347	0.0275
Male (G)	0.0513	0.0394	0.0380	0.0440	0.0530	0.0573	0.0595	0.0689	0.0710	0.0737
Tenure	-0.0009	-0.0003	-0.0003	-0.0002	-0.00004	0.0003	0.0006	0.0011	0.0011	0.0010

Table 69: Empirical results for the Full-Sample, using OLS estimator with RE and auto-correlated errors.

Variable	Length									
	1	2	3	4	5	6	7	8	9	10
Age	0.0041	0.0043	0.0045	0.0042	0.0045	0.0040	0.0041	0.0035	0.0034	0.0031
Years of education (Edu)	0.0716	0.0680	0.0667	0.0666	0.0651	0.0647	0.0650	0.0638	0.0648	0.0650
Single (D)	-0.3500	-0.3271	-0.2938	-0.2635	-0.2242	-0.2130	-0.1977	-0.1933	-0.1840	-0.1792
Firm size 20-199 (F_3)	0.2118	0.2175	0.2074	0.2147	0.2079	0.1957	0.1912	0.1832	0.1770	0.1731
Firm size 200-1999 (F_4)	0.3422	0.3371	0.3262	0.3339	0.3156	0.2859	0.2743	0.2622	0.2531	0.2448
Firm size 2000+ (F_5)	0.3291	0.3389	0.3229	0.3422	0.3441	0.3205	0.3144	0.3044	0.2969	0.2893
Male (G)	0.2699	0.2762	0.2743	0.2766	0.2841	0.2862	0.2881	0.2937	0.2975	0.2974
Tenure	0.0055	0.0069	0.0057	0.0053	0.0047	0.0053	0.0053	0.0054	0.0054	0.0055

Table 70: Simulation results for the Resp-Samples, using IPW estimator with RE and auto-correlated errors. Number of Monte Carlo replications is $R = 100$. Initial non-response rate is 30%, with no panel attrition after initial wave.

Variable	Length									
	1	2	3	4	5	6	7	8	9	10
Age	0.0038	0.0040	0.0039	0.0033	0.0036	0.0034	0.0035	0.0032	0.0031	0.0028
Years of education (Edu)	0.0705	0.0684	0.0674	0.0674	0.0662	0.0664	0.0677	0.0668	0.0672	0.0679
Single (D)	-0.2941	-0.2751	-0.2242	-0.1979	-0.1734	-0.1613	-0.1480	-0.1455	-0.1378	-0.1350
Firm size 20-199 (F_3)	0.2144	0.2020	0.1876	0.1960	0.1935	0.1892	0.1882	0.1781	0.1698	0.1679
Firm size 200-1999 (F_4)	0.3271	0.3053	0.2924	0.3024	0.2879	0.2748	0.2668	0.2566	0.2457	0.2413
Firm size 2000+ (F_5)	0.3122	0.3061	0.2916	0.3113	0.3105	0.2993	0.2955	0.2878	0.2804	0.2791
Male (G)	0.2614	0.2620	0.2630	0.2584	0.2575	0.2591	0.2613	0.2607	0.2632	0.2624
Tenure	0.0064	0.0072	0.0058	0.0053	0.0045	0.0048	0.0045	0.0041	0.0041	0.0041

Table 71: IPW estimator with RE and auto-correlated errors: Fade-away effect of the initial non-response bias with SOEP and artificial initial non-response, with no panel attrition.

Difference of:	Bias$(\hat{b}_{p,t}) = (\hat{b}_{p,t}^{Full} - \hat{b}_{p,t}^{Resp})$ in length									
	1	2	3	4	5	6	7	8	9	10
Age	0.0004	0.0003	0.0006	0.0009	0.0009	0.0007	0.0006	0.0003	0.0003	0.0003
Years of education (Edu)	0.0011	-0.0004	-0.0007	-0.0009	-0.0011	-0.0018	-0.0028	-0.0030	-0.0024	-0.0029
Single (D)	-0.0559	-0.0520	-0.0697	-0.0656	-0.0509	-0.0517	-0.0497	-0.0478	-0.0463	-0.0442
Firm size 20-199 (F_3)	-0.0026	0.0155	0.0199	0.0187	0.0144	0.0066	0.0030	0.0051	0.0072	0.0053
Firm size 200-1999 (F_4)	0.0151	0.0318	0.0338	0.0315	0.0277	0.0111	0.0075	0.0056	0.0075	0.0035
Firm size 2000+ (F_5)	0.0168	0.0328	0.0314	0.0309	0.0336	0.0213	0.0189	0.0167	0.0165	0.0102
Male (G)	0.0085	0.0142	0.0113	0.0183	0.0266	0.0271	0.0269	0.0330	0.0343	0.0351
Tenure	-0.0009	-0.0002	-0.0002	0.0002	0.0002	0.0005	0.0007	0.0013	0.0013	0.0014

Table 72: Empirical results for the Full-Sample, using OLS estimator with RE and lagged $W_{i,t-1}$.

Variable	Length								
	2	3	4	5	6	7	8	9	10
Intercept	0.4958	0.5475	0.5698	0.5478	0.5855	0.5634	0.5841	0.5626	0.5648
Lag hourly wage	0.1654	0.1409	0.1226	0.1385	0.1225	0.1408	0.1438	0.1592	0.1630
Age	0.0037	0.0039	0.0036	0.0039	0.0035	0.0034	0.0032	0.0028	0.0026
Years of education (Edu)	0.0563	0.0573	0.0593	0.0573	0.0587	0.0585	0.0575	0.0575	0.0578
Single (D)	-0.2686	-0.2315	-0.2167	-0.1825	-0.1777	-0.1660	-0.1629	-0.1536	-0.1536
Firm size 20-199 (F_3)	0.1702	0.1723	0.1826	0.1752	0.1696	0.1469	0.1362	0.1331	0.1268
Firm size 200-1999 (F_4)	0.2619	0.2675	0.2851	0.2677	0.2408	0.2168	0.1997	0.1902	0.1797
Firm size 2000+ (F_5)	0.2784	0.2740	0.2987	0.2986	0.2746	0.2530	0.2389	0.2273	0.2169
Male (G)	0.2504	0.2396	0.2411	0.2443	0.2510	0.2504	0.2557	0.2557	0.2576
Tenure	0.0070	0.0051	0.0042	0.0031	0.0034	0.0031	0.0027	0.0028	0.0027

Table 73: Simulation results for the Resp-Samples, using IPW estimator with RE and lagged $W_{i,t-1}$. Number of Monte Carlo replications is $R = 100$. Initial non-response rate is 30%, with no panel attrition after initial wave.

Variable	Length								
	2	3	4	5	6	7	8	9	10
Intercept	0.7924	0.8407	0.8166	0.7670	0.7837	0.7428	0.7481	0.7286	0.7183
Lag hourly wage	0.0679	0.0314	0.0441	0.0754	0.0685	0.0904	0.0997	0.1137	0.1208
Age	0.0027	0.0029	0.0024	0.0027	0.0025	0.0026	0.0024	0.0022	0.0020
Years of education (Edu)	0.0602	0.0625	0.0632	0.0610	0.0623	0.0623	0.0615	0.0614	0.0620
Single (D)	-0.2214	-0.1662	-0.1435	-0.1224	-0.1152	-0.1095	-0.1101	-0.1064	-0.1091
Firm size 20-199 (F_3)	0.1405	0.1386	0.1489	0.1491	0.1401	0.1302	0.1228	0.1199	0.1158
Firm size 200-1999 (F_4)	0.2182	0.2263	0.2413	0.2289	0.2100	0.1939	0.1809	0.1705	0.1642
Firm size 2000+ (F_5)	0.2368	0.2369	0.2571	0.2568	0.2373	0.2208	0.2110	0.2014	0.1964
Male (G)	0.2351	0.2380	0.2311	0.2233	0.2247	0.2230	0.2216	0.2200	0.2186
Tenure	0.0081	0.0064	0.0054	0.0041	0.0041	0.0036	0.0028	0.0027	0.0025

Table 74: IPW estimator with RE and lagged $W_{i,t-1}$: Fade-away effect of the initial non-response bias with SOEP and artificial initial non-response, with no panel attrition.

Difference of:	$\text{Bias}(\hat{b}_{p,t}) = (\hat{b}_{p,t}^{Full} - \hat{b}_{p,t}^{Resp})$ in length								
	2	3	4	5	6	7	8	9	10
Intercept	-0.2966	-0.2932	-0.2468	-0.2191	-0.1982	-0.1794	-0.1640	-0.1660	-0.1535
Lag hourly wage	0.0971	0.1095	0.0785	0.0632	0.0540	0.0504	0.0441	0.0456	0.0423
Age	0.0010	0.0010	0.0012	0.0012	0.0010	0.0008	0.0007	0.0006	0.0005
Years of education (Edu)	-0.0039	-0.0052	-0.0039	-0.0037	-0.0035	-0.0037	-0.0040	-0.0039	-0.0042
Single (D)	-0.0473	-0.0654	-0.0732	-0.0601	-0.0625	-0.0565	-0.0527	-0.0472	-0.0445
Firm size 20-199 (F_3)	0.0296	0.0337	0.0338	0.0262	0.0248	0.0167	0.0135	0.0133	0.0110
Firm size 200-1999 (F_4)	0.0438	0.0412	0.0438	0.0388	0.0309	0.0229	0.0188	0.0197	0.0155
Firm size 2000+ (F_5)	0.0416	0.0372	0.0416	0.0418	0.0373	0.0323	0.0279	0.0260	0.0205
Male (G)	0.0153	0.0016	0.0100	0.0210	0.0263	0.0275	0.0341	0.0357	0.0391
Tenure	-0.0011	-0.0013	-0.0012	-0.0011	-0.0007	-0.0005	-0.0001	0.0001	0.0001

B.2. Analysis tables for satisfaction scores

Table 75: Empirical results for the Full-Sample, using ordered logit model estimator.

Variable	Panel wave t										
	1	2	3	4	5	6	7	8	9	10	11
Intercept 5	-1.3048	-1.1404	-0.7524	-0.6846	-0.6711	-0.7665	-0.8157	-0.7570	-0.7451	-0.7313	-0.7323
Intercept 6	-0.5855	-0.4197	-0.0106	0.0603	0.0696	-0.0290	-0.0773	-0.0109	-0.0033	0.0014	-0.0005
Intercept 7	0.4874	0.6925	1.1468	1.2338	1.2698	1.1721	1.1202	1.1929	1.2126	1.2169	1.2176
Intercept 8	2.2357	2.4834	2.9717	3.0404	3.0808	2.9992	2.9663	3.0560	3.0920	3.1022	3.1116
Intercept 9	3.5350	3.8561	4.3836	4.4858	4.5145	4.4786	4.4572	4.5865	4.6403	4.6492	4.6798
Age	0.0697	0.0670	0.0560	0.0553	0.0584	0.0648	0.0693	0.0672	0.0665	0.0666	0.0658
Age squared	-0.0010	-0.0010	-0.0008	-0.0008	-0.0008	-0.0009	-0.0009	-0.0009	-0.0009	-0.0009	-0.0009
Years of education	0.0002	-0.0115	-0.0213	-0.0227	-0.0224	-0.0264	-0.0283	-0.0287	-0.0296	-0.0297	-0.0295
Male	0.0845	0.1160	0.1072	0.1183	0.1322	0.1330	0.1288	0.1248	0.1230	0.1211	0.1240
Single	0.4472	0.3953	0.3499	0.3082	0.3060	0.3179	0.3239	0.3109	0.3157	0.3201	0.3359
Widowed	0.1997	0.3374	0.3282	0.3152	0.2726	0.2535	0.2542	0.2687	0.3222	0.3266	0.3453
Divorced	0.4283	0.3925	0.3285	0.3442	0.3217	0.3161	0.3298	0.3312	0.3335	0.3098	0.3159
Separated	0.6090	0.6965	0.8676	0.7796	0.8152	0.8516	0.9210	0.9543	0.8754	0.8420	0.8271
Doctor visits	-0.0006	-0.0011	-0.0011	-0.0017	-0.0023	-0.0017	-0.0019	-0.0021	-0.0022	-0.0021	-0.0023
Hospital stays	0.0473	0.0445	-0.1087	-0.0562	-0.0303	-0.0439	-0.0089	0.0103	0.0172	0.0092	0.0205
Household income	-1*e^{-}5	-1*e^{-}5	-1*e^{-}5	-8*e^{-}6	-1*e^{-}5	-1*e^{-}5	-1*e^{-}5	-8*e^{-}6	-8*e^{-}6	-8*e^{-}6	-8*e^{-}6
Health satisfaction 5	-0.0503	-0.5689	-0.6113	-0.6159	-0.6123	-0.5710	-0.5696	-0.5702	-0.5669	-0.5719	-0.5799
Health satisfaction 6	-0.7782	-0.8763	-0.9294	-0.9931	-1.0171	-0.9548	-0.9977	-1.0046	-1.0131	-1.0081	-1.0249
Health satisfaction 7	-1.2027	-1.2622	-1.3671	-1.4240	-1.4399	-1.4417	-1.4785	-1.5217	-1.5275	-1.5404	-1.5647
Health satisfaction 8	-1.7676	-1.8621	-1.9707	-1.9956	-2.0622	-2.0620	-2.0958	-2.1405	-2.1492	-2.1704	-2.2072
Health satisfaction 9	-2.4420	-2.4546	-2.6157	-2.6330	-2.6838	-2.6976	-2.7524	-2.8175	-2.8594	-2.8876	-2.9294
Health satisfaction 10	-2.6795	-2.8482	-2.9859	-3.0832	-3.1389	-3.1572	-3.2241	-3.2914	-3.3278	-3.3435	-3.3637

Dependent variable is life satisfaction (11-point scale) of individual's aged 17 and above over the sample period 2000-2010. Source: Own calculations with life satisfaction data from SOEP, waves 2000-2010 (see text).

Table 76: Simulation results for the Resp-Samples in Scenario $\alpha = -6.00$ and $\beta = 0.90$, using ordered logit model estimator.

Variable	Panel wave t										
	1	2	3	4	5	6	7	8	9	10	11
Intercept 5	-3.0070	-2.1036	-1.4007	-1.2323	-1.0848	-1.1172	-1.0887	-0.9828	-0.9842	-0.9428	-0.9280
Intercept 6	-2.1015	-1.3218	-0.6248	-0.4737	-0.3371	-0.3719	-0.3360	-0.2238	-0.2299	-0.2020	-0.1875
Intercept 7	-0.6798	-0.0361	0.6576	0.7875	0.9251	0.8779	0.9028	1.0204	1.0274	1.0512	1.0654
Intercept 8	1.2940	1.8865	2.5747	2.6653	2.7977	2.7638	2.8052	2.9349	2.9544	2.9760	3.0005
Intercept 9	2.6604	3.3166	4.0402	4.1593	4.2740	4.2777	4.3302	4.4962	4.5357	4.5529	4.5981
Age	0.0722	0.0683	0.0527	0.0508	0.0533	0.0587	0.0615	0.0583	0.0581	0.0578	0.0566
Age squared	-0.0011	-0.0010	-0.0008	-0.0008	-0.0008	-0.0008	-0.0009	-0.0008	-0.0008	-0.0008	-0.0008
Years of education	0.0100	-0.0090	-0.0228	-0.0247	-0.0249	-0.0301	-0.0324	-0.0323	-0.0326	-0.0328	-0.0319
Male	0.1579	0.1865	0.1830	0.1892	0.2005	0.1985	0.1986	0.1932	0.1930	0.1907	0.1928
Single	0.3106	0.3057	0.2346	0.2135	0.2127	0.2387	0.2560	0.2556	0.2639	0.2673	0.2786
Widowed	0.3350	0.4098	0.4303	0.4294	0.3880	0.3721	0.3687	0.3907	0.4419	0.4433	0.4509
Divorced	0.3264	0.3473	0.3251	0.3194	0.3003	0.2923	0.3123	0.3265	0.3302	0.2983	0.2971
Separated	0.3535	0.4718	0.7081	0.6567	0.7206	0.7605	0.8253	0.8522	0.7798	0.7504	0.7541
Doctor visits	-0.0016	-0.0014	-0.0019	-0.0028	-0.0034	-0.0025	-0.0028	-0.0030	-0.0032	-0.0030	-0.0030
Hospital stays	0.0151	-0.0444	-0.1748	-0.0813	-0.0429	-0.0452	-0.0236	-0.0001	0.0064	0.0009	0.0068
Household income	-1*e^-5	-1*e^-5	-8*e^-6	-7*e^-6	-7*e^-6	-7*e^-6	-7*e^-6	-7*e^-6	-6*e^-6	-6*e^-6	-6*e^-6
Health satisfaction 5	-0.4712	-0.6461	-0.6177	-0.5730	-0.6209	-0.5713	- 0.5819	-0.5939	-0.5810	-0.5949	-0.5928
Health satisfaction 6	-0.4334	-0.6187	-0.6985	-0.7747	-0.8471	-0.8203	- 0.8788	-0.9146	-0.9275	-0.9327	-0.9469
Health satisfaction 7	-0.8119	-1.0505	-1.1691	-1.2181	-1.2782	-1.2888	- 1.3526	-1.4246	-1.4326	-1.4497	-1.4783
Health satisfaction 8	-1.3092	-1.6327	-1.7613	-1.7948	-1.9040	-1.9200	- 1.9838	-2.0586	-2.0719	-2.0971	-2.1476
Health satisfaction 9	-1.9958	-2.2195	-2.4045	-2.4353	-2.5397	-2.5723	- 2.6684	-2.7634	-2.8096	-2.8355	-2.8824
Health satisfaction 10	-2.4173	-2.7253	-2.9075	-2.9991	-3.1166	-3.1334	- 3.2312	-3.3278	-3.3719	-3.3825	-3.4170

Dependent variable is life satisfaction (11-point scale) of individual's aged 17 and above over the sample period 2000-2010. Number of Monte Carlo replications $R = 100$ times. Initial non-response rate is 35%, with no panel attrition in later waves. Source: Own calculations with life satisfaction data from SOEP, waves 2000-2010 (see text).

Table 77: Fade-away effect of the ordered logit model estimator for cross-sectional data, with SOEP and artificial initial non-response, with no panel attrition.

Difference of:	Bias$(\hat{b}_{p,t}) = (\hat{b}_{p,t}^{Full} - \hat{b}_{p,t}^{Resp})$ in wave t										
	1	2	3	4	5	6	7	8	9	10	11
Intercept 5	1.7021	0.9632	0.6484	0.5477	0.4137	0.3507	0.2730	0.2259	0.2391	0.2115	0.1958
Intercept 6	1.5161	0.9021	0.6142	0.5340	0.4066	0.3429	0.2587	0.2128	0.2266	0.2034	0.1870
Intercept 7	1.1672	0.7286	0.4893	0.4463	0.3447	0.2941	0.2174	0.1726	0.1852	0.1657	0.1523
Intercept 8	0.9417	0.5969	0.3970	0.3751	0.2832	0.2353	0.1611	0.1211	0.1376	0.1262	0.1111
Intercept 9	0.8746	0.5396	0.3434	0.3265	0.2405	0.2009	0.1270	0.0903	0.1046	0.0964	0.0817
Age	-0.0024	-0.0013	0.0033	0.0045	0.0051	0.0061	0.0078	0.0089	0.0085	0.0088	0.0092
Age squared	6*e^-5	3*e^-5	-2*e^-5	-4*e^-5	-4*e^-5	-0.0001	-0.0001	-0.0001	-0.0001	-0.0001	-0.0001
Years of education	-0.0098	-0.0025	0.0016	0.0020	0.0025	0.0037	0.0041	0.0037	0.0030	0.0030	0.0024
Male	-0.0734	-0.0706	-0.0759	-0.0709	-0.0683	-0.0655	-0.0698	-0.0684	-0.0700	-0.0695	-0.0687
Single	0.1366	0.0896	0.1154	0.0947	0.0933	0.0792	0.0679	0.0553	0.0519	0.0528	0.0572
Widowed	-0.1353	-0.0724	-0.1020	-0.1142	-0.1154	-0.1186	-0.1145	-0.1220	-0.1198	-0.1167	-0.1056
Divorced	0.1018	0.0453	0.0034	0.0248	0.0217	0.0238	0.0175	0.0048	0.0033	0.0115	0.0188
Separated	0.2555	0.2247	0.1595	0.1229	0.0946	0.0911	0.0958	0.1020	0.0956	0.0916	0.0730
Doctor visits	0.0011	0.0003	0.0008	0.0011	0.0012	0.0010	0.0008	0.0009	0.0010	0.0009	0.0007
Hospital stays	0.0323	0.0890	0.0661	0.0250	0.0126	0.0013	0.0147	0.0104	0.0108	0.0084	0.0137
Household income	-2*e^-6	-1*e^-6	-1.6*e^-6	-1.7*e^-6	-1.6*e^-6	-1.7*e^-6	-1.7*e^-6	-1.7*e^-6	-1.7*e^-6	-1.7*e^-6	-1.7*e^-6
Health satisfaction 5	-0.0319	0.0772	0.0064	-0.0430	0.0087	0.0003	0.0124	0.0237	0.0141	0.0230	0.0129
Health satisfaction 6	-0.3448	-0.2575	-0.2309	-0.2185	-0.1700	-0.1345	-0.1189	-0.0900	-0.0856	-0.0754	-0.0780
Health satisfaction 7	-0.3908	-0.2117	-0.1980	-0.2059	-0.1618	-0.1528	-0.1259	-0.0970	-0.0950	-0.0907	-0.0864
Health satisfaction 8	-0.4583	-0.2293	-0.2094	-0.2008	-0.1582	-0.1420	-0.1120	-0.0819	-0.0773	-0.0733	-0.0596
Health satisfaction 9	-0.4462	-0.2351	-0.2112	-0.1977	-0.1441	-0.1254	-0.0840	-0.0542	-0.0498	-0.0520	-0.0470
Health satisfaction 10	-0.2622	-0.1229	-0.0784	-0.0842	-0.0224	-0.0238	0.0071	0.0363	0.0440	0.0390	0.0533

Table 78: Empirical results of the regression of life satisfaction under the Full-Sample using RE model estimator.

Variable	Length of the panel									
	1	2	3	4	5	6	7	8	9	10
Intercept	6.8856	6.5765	6.4868	6.4273	6.6498	6.4997	6.4198	6.4202	6.2903	6.3766
Age	-0.0433	-0.0345	-0.0313	-0.0327	-0.0349	-0.0334	-0.0270	-0.0263	-0.0253	-0.0221
Age squared	0.0006	0.0005	0.0005	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0003
Years of education	0.0147	0.0222	0.0227	0.0225	0.0258	0.0297	0.0301	0.0299	0.0309	0.0309
Male	-0.0766	-0.0644	-0.0741	-0.0831	-0.0834	-0.0815	-0.0801	-0.0805	-0.0807	-0.0825
Single	-0.2804	-0.2439	-0.1792	-0.1719	-0.1846	-0.1992	-0.1732	0.1807	-0.1803	-0.2037
Widowed	-0.1344	-0.0571	-0.0484	0.0071	0.0100	0.0091	-0.0528	-0.1119	-0.0814	-0.1026
Divorced	-0.2559	-0.2074	-0.1995	-0.1750	-0.1811	-0.1921	-0.1724	-0.1526	-0.1182	-0.1131
Separated	-0.5775	-0.6229	-0.5400	-0.4707	-0.5125	-0.5291	-0.5014	-0.4667	-0.4566	-0.4467
Doctor visits	-0.0004	-0.0010	-0.0009	-0.0010	-0.0016	-0.0014	-0.0011	-0.0011	-0.0010	-0.0011
Hospital stays	-0.0637	-0.0176	-0.0129	-0.0268	-0.0125	-0.0014	-0.0472	-0.0517	-0.0489	-0.0525
Household income	$6*e^{-}6$	$5*e^{-}6$	$4*e^{-}6$	$4*e^{-}6$	$4*e^{-}6$	$4*e^{-}6$	$4*e^{-}6$	$3*e^{-}6$	$3*e^{-}6$	$3*e^{-}6$
Health satisfaction 5	0.3656	0.3497	0.3159	0.3185	0.2807	0.2868	0.2857	0.2821	0.2908	0.3006
Health satisfaction 6	0.5698	0.5239	0.5401	0.5112	0.4458	0.4634	0.4663	0.4734	0.4676	0.4865
Health satisfaction 7	0.7968	0.7690	0.7486	0.7323	0.7069	0.7277	0.7454	0.7375	0.7330	0.7434
Health satisfaction 8	1.1579	1.0875	1.0410	1.0230	0.9823	0.9762	0.9784	0.9660	0.9640	0.9774
Health satisfaction 9	1.5030	1.4308	1.3475	1.3001	1.2388	1.2402	1.2366	1.2316	1.2280	1.2446
Health satisfaction 10	1.6855	1.5660	1.5089	1.4416	1.3911	1.3869	1.3882	1.3890	1.3886	1.4012

Dependent variable is life satisfaction (11-point scale) of individual's aged 17 and above over the sample period 2000-2010. Source: Own calculations with life satisfaction data from SOEP, waves 2000-2010 (see text).

Table 79: Simulation results of the regression of life satisfaction under the Resp-Samples, using RE model estimator.

Variable	Length of the panel									
	1	2	3	4	5	6	7	8	9	10
Intercept	7.2156	6.7610	6.6484	6.5645	6.7856	6.6023	6.5219	6.5341	6.3882	6.4665
Age	-0.0372	-0.0272	-0.0249	-0.0275	-0.0300	-0.0285	-0.0224	-0.0216	-0.0200	-0.0164
Age squared	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003	0.0003	0.0003
Years of education	0.0107	0.0198	0.0212	0.0216	0.0257	0.0291	0.0290	0.0284	0.0291	0.0285
Male	-0.1104	-0.1072	-0.1117	-0.1234	-0.1260	-0.1255	-0.1237	-0.1236	-0.1235	-0.1235
Single	-0.2040	-0.1481	-0.1068	-0.1114	-0.1358	-0.1507	-0.1378	-0.1477	-0.1454	-0.1690
Widowed	-0.1759	-0.1510	-0.1544	-0.1284	-0.1286	-0.1405	-0.2123	-0.2727	-0.2454	-0.2550
Divorced	-0.2020	-0.1839	-0.1524	-0.1393	-0.1298	-0.1438	-0.1367	-0.1195	-0.0941	-0.0864
Separated	-0.4153	-0.5601	-0.5252	-0.4762	-0.5168	-0.5373	-0.5100	-0.4693	-0.4697	-0.4674
Doctor visits	-0.0004	-0.0012	-0.0011	-0.0013	-0.0022	-0.0018	-0.0013	-0.0012	-0.0012	-0.0012
Hospital stays	-0.0195	0.0478	-0.0096	-0.0244	-0.0203	-0.0490	-0.0502	-0.0579	-0.0513	-0.0514
Household income	$1*e^{-}5$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Health satisfaction 5	0.3707	0.3411	0.3089	0.3340	0.2856	0.3006	0.3119	0.3022	0.3084	0.3062
Health satisfaction 6	0.4145	0.4249	0.4689	0.4582	0.4068	0.4349	0.4561	0.4627	0.4538	0.4688
Health satisfaction 7	0.6457	0.6495	0.6538	0.6660	0.6467	0.6906	0.7312	0.7260	0.7205	0.7298
Health satisfaction 8	0.9532	0.9308	0.9262	0.9392	0.9058	0.9267	0.9526	0.9397	0.9358	0.9531
Health satisfaction 9	1.2391	1.2445	1.2121	1.1949	1.1485	1.1781	1.1956	1.1932	1.1850	1.2025
Health satisfaction 10	1.4810	1.4178	1.3829	1.3521	1.3038	1.3317	1.3498	1.3521	1.3419	1.3549

Dependent variable is life satisfaction (11-point scale) of individual's aged 17 and above over the sample period 2000-2010. Number of Monte Carlo replications $R = 100$ times. Initial non-response rate is 35%, with no panel attrition in later waves. ource: Own calculations with life satisfaction data from SOEP, waves 2000-2010 (see text).

Table 80: Fade-away effect of the RE model estimator, with SOEP and artificial initial non-response, with no panel attrition.

Difference of:	$\text{Bias}(\hat{b}_{p,t}) = (\hat{b}_{p,t}^{Full} - \hat{b}_{p,t}^{Resp})$ in length									
	1	2	3	4	5	6	7	8	9	10
Intercept	-0.3301	-0.1845	-0.1616	-0.1372	-0.1359	-0.1026	-0.1020	-0.1139	-0.0979	-0.0898
Age	-0.0061	-0.0074	-0.0063	-0.0052	-0.0049	-0.0052	-0.0046	-0.0047	-0.0053	-0.0057
Age squared	0.0001	0.0001	0.0001	0.00004	0.00004	0.00004	0.00003	0.00004	0.00004	0.00005
Years of education	0.0040	0.0024	0.0016	0.0009	0.0001	0.0007	0.0011	0.0015	0.0018	0.0025
Male	0.0338	0.0428	0.0376	0.0403	0.0427	0.0440	0.0436	0.0431	0.0427	0.0410
Single	-0.0764	0.0958	-0.0724	-0.0605	-0.0488	-0.0486	-0.0354	-0.0330	-0.0349	-0.0346
Widowed	0.0416	0.0939	0.1060	0.1355	0.1386	0.1497	0.1595	0.1608	0.1640	0.1524
Divorced	-0.0540	-0.0235	-0.0471	-0.0357	-0.0513	-0.0483	-0.0357	-0.0330	-0.0242	-0.0268
Separated	-0.1623	-0.0628	-0.0148	0.0055	0.0043	0.0083	0.0086	0.0026	0.0131	0.0207
Doctor visits	0.00001	0.0002	0.0002	0.0003	0.0006	0.0004	0.0002	0.0001	0.0001	0.0001
Hospital stays	-0.0442	-0.0302	-0.0034	-0.0024	0.0078	0.0018	0.0030	0.0061	0.0024	-0.0012
Household income	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Health satisfaction 5	-0.0051	0.0087	0.0070	-0.0155	-0.0049	-0.0138	-0.0262	-0.0201	-0.0176	-0.0056
Health satisfaction 6	0.1553	0.0990	0.0712	0.0531	0.0391	0.0249	0.0103	0.0107	0.0137	0.0177
Health satisfaction 7	0.1512	0.1195	0.0949	0.0663	0.0602	0.0371	0.0142	0.0115	0.0125	0.0137
Health satisfaction 8	0.2047	0.1567	0.1149	0.0837	0.0765	0.0495	0.0258	0.0263	0.0283	0.0243
Health satisfaction 9	0.2640	0.1864	0.1355	0.1052	0.0904	0.0622	0.0410	0.0383	0.0429	0.0421
Health satisfaction 10	0.2045	0.1483	0.1260	0.0895	0.0874	0.0552	0.0384	0.0369	0.0467	0.0463

Statutory Declaration

Declaration of Authorship

I hereby solemnly declare that this thesis contains my original work, and that I have written this thesis independently. I also declare that the material of this work has not been accepted for the award of any other degree under my name concurrently or latterly to this University or any other tertiary institution. Besides that, and to the best of my knowledge and belief, this thesis doesn't contain any material which has been previously written or published by another person, except from due reference which has been made in the text.

Berlin, 11 February, 2020

Mursala Khan