

**Dealing With
Various Flavors of Missing Data in
Ex-Post Survey Harmonization and
Beyond**

Inauguraldissertation
zur Erlangung des akademischen Grades
einer Doktorin der Sozialwissenschaften
der Universität Mannheim

Vorgelegt von
Anna-Carolina Haensch

Hauptamtlicher Dekan der Fakultät für Sozialwissenschaften:

Prof. Dr. Michael Diehl

Erstbetreuer:

Prof. Dr. Jörg Drechsler

Zweitbetreuer:

Prof. Dr. Florian Keusch

Erstgutachter:

Prof. Dr. Florian Keusch

Zweitgutachterin:

Prof. Dr. Frauke Kreuter

Tag der Disputation:

15.06.2021

Acknowledgement

A lot of people have supported me during the work on this dissertation. First of all, I would like to thank my supervisor Jörg Drechsler. His encouragement, feedback, and support are the most important contribution to the completion of this dissertation. His great joy and expertise in tackling missing data problems have inspired me since my master's studies.

I wish to express my sincerest gratitude to Bernd Weiß, who was my mentor, project leader in the HaSpaD project at GESIS, and co-author and who widely opened up the field of ex-post survey harmonization for me. His interest and expertise in research synthesis, open science, and new forms of data analysis have encouraged and enabled me to write this dissertation.

Furthermore, I would like to thank Florian Keusch and Frauke Kreuter for reviewing and offering many opportunities and helpful advice during my PhD. Many thanks to my GESIS colleagues in Mannheim and Cologne, especially those from the DFG-project HaSpaD: Bernd Weiß, Sonja Schulz, Sebastian Sterl, Lisa Schmid, and Antonia May. It was a good time! My GESIS Panel colleagues for lots of coffee chats and lunch breaks. Thanks also go to Reinhard Schunck for being the co-author on one of the papers. Thanks to the PhD speakers at GESIS, Hannah Bucher and Anne Stroppe, and all other PhD students for creating a strong bond among PhD students at GESIS. Thanks for countless pomodoro sessions on Zoom, feedback in the research lab sessions, and the willingness to share advice and experiences. Thanks to Verena Ortmanns and Isabella Minderop for the many Friday afternoons spent writing together and for your great feedback. My sincerest thanks also go to my colleagues and students at the University of Mannheim, especially the FK2RG research group and IPSDS!

Finally, I would like to deeply thank my family and friends for their continuous love, help and support. Special thanks go to Mathis, who supported me all along the way even though he had his own PhD thesis to write. You are the best!

What Statistical Problems Are Not Missing-Data Problems?

Xiao-Li Meng, Joint Statistical Meetings (JSM) 2013

Contents

1. Introduction	1
1.1. Ex-post survey harmonization	2
1.2. Various flavors of missing data	6
1.2.1. Item nonresponse: Sporadically missing data	7
1.2.2. Unit nonresponse: Completely missing data for a unit	9
1.2.3. Systematically missing data: Completely missing data for a variable	11
1.3. Why this dissertation?	12
1.4. Extended summary of chapters	15
2. Multiple Imputation of Partially Observed Covariates in Discrete-Time	
Survival Analysis	30
2.1. Introduction	30
2.2. Discrete-time survival analysis model	32
2.2.1. The model	32
2.2.2. The data: Transformations and the person-period format	34
2.2.3. Implications of missing data for discrete-time survival analysis . .	36
2.3. Multiple imputation	38
2.3.1. Multiple imputation in general	38
2.3.2. Handling missing covariates values in case of a DTSAM as a substantive model	40

Contents

2.4. Simulation study	44
2.4.1. Data-generating mechanisms	44
2.4.2. Missingness	45
2.4.3. Methods compared and performance measures	46
2.4.4. Simulation results	46
2.5. An applied example with the German Family Panel pairfam	50
2.6. Discussion	53
3. Systematically Missing Partner Variables and Multiple Imputation Strategies: A Case Study With German Relationship Data	62
3.1. Introduction	62
3.2. Literature overview: Systematically missing data in related fields	65
3.3. Motivational example: Life satisfaction of partners in Germany	68
3.4. Imputation strategies for systematically missing partner variables	69
3.4.1. Multiple imputation	69
3.4.2. Multiple imputation for systematically missing partner variables – Assumptions and bridging studies	71
3.5. Data and simulation	75
3.5.1. Data sets: GSOEP, SHARE and pairfam	75
3.5.2. Simulation conditions	77
3.6. Results	78
3.6.1. Correlation results	79
3.6.2. Regression coefficient results	81
3.7. Study heterogeneity	86
3.8. Discussion	88

4. TippingSens: An R Shiny Application to Facilitate Sensitivity Analysis for Causal Inference Under Confounding	97
4.1. Introduction	97
4.2. Rubin’s Causal Model and the assumption of unconfoundedness	101
4.2.1. The Rubin Causal Model	101
4.2.2. Quasi-experiments and the assumption of unconfoundedness	102
4.3. Sensitivity analysis in the context of causal inference	103
4.4. The Rosenbaum-Rubin sensitivity analysis and the TippingSens App	107
4.4.1. Technical details	107
4.4.2. The TippingSens app as a visualization tool	108
4.5. A practical example: Sensitivity analysis for a quasi-experimental evaluation of a German vocational training program for the unemployed	110
4.5.1. Study details	110
4.5.2. Sensitivity analysis with the TippingSens application	113
4.6. Discussion	119
5. Better Together? Regression Analysis of Complex Survey Data After Ex-post Survey Harmonization	128
5.1. Introduction	128
5.2. Regression analysis of complex survey-based data after harmonization	130
5.2.1. Overview and literature review	130
5.2.2. Two-stage IPD meta-analytical approaches	132
5.2.3. One-stage IPD meta-analytical approaches	133
5.2.4. Survey weights and the use of weights in regression analysis	135
5.3. A comparison of one-stage and two-stage approaches in case of pooled complex survey data	143
5.3.1. Introduction to the simulation design	143

5.3.2. Methodological decisions and differences between one-stage/two-stage meta-analysis	146
5.3.3. Study heterogeneity	148
5.4. A practical example: Same-sex couples and their satisfaction with family life in Germany	152
5.5. Discussion	155
6. Conclusion and Discussion	167
A. Appendix	174

1. Introduction

This dissertation is dedicated to overcoming missing data problems in ex-post survey harmonization. Although ex-post survey harmonization projects have become more common in the social sciences in recent years, methodological research on this topic is still relatively rare.

The present introductory chapter is organized as follows: In the first section, I provide a general introduction to ex-post survey harmonization, and in the second section, I give an overview of missing data problems in this field. I then further elucidate the motivation for the dissertation's specific topics and provide a summary of the main chapters. The dissertation's main body consists of four chapters, each of which presents a study exploring solutions for missing data problems not only in ex-post survey harmonization but also in related research fields. The study presented in Chapter 2 deals with multiple imputation approaches for discrete-time survival analysis. Chapter 3 presents a study examining multiple imputation approaches to handling systematically missing partner variables. The study presented in Chapter 4 is on the topic of sensitivity analysis in the case of unobserved confounders. And finally, the study in Chapter 5 is on the topic of weighting pooled survey data. The last chapter concludes the dissertation with a discussion of the results, scientific contributions, and limitations of the aforementioned studies, plus a broader outlook on the issue of study heterogeneity.

1.1. Ex-post survey harmonization

Comparing populations or specific social groups across time and space is an important research approach in the social sciences (Friedrichs and Nonnenmacher 2010). However, studying how certain aspects of the collective environment – such as socioeconomic conditions, the laws in force, institutions, or values – influence individuals’ characteristics is possible only with high-quality data spanning the time and space of interest. In many cases, these requirements regarding time, space, or analytical power cannot be fulfilled by a single survey. Rather, data from different surveys must be combined into a single pooled data set and prepared for analysis. This is referred to as survey harmonization (Dubrow and Tomescu-Dubrow 2015; Hussong et al. 2013; Granda et al. 2010), of which there are two types: *ex-ante* and *ex-post* (Wolf et al. 2016). The main difference between the two types is that *ex-ante* harmonization is part of the survey design; thus, the pooling of data is anticipated before data collection. By contrast, *ex-post* harmonization is done on pre-existing, already collected data; these data were not originally intended to be combined.

In this dissertation, I present methodological research that I conducted for the field of *ex-post* survey harmonization. As a rule, *ex-post* survey harmonization is not simply a matter of combining individual data sets into one pooled data set. Rather, it requires, for example, thorough documentation of the differences between and the peculiarities of the individual data sets. This enables secondary users of the harmonized data set to model differences in their analyses. Another important part of *ex-post* survey harmonization is variable harmonization – that is, the process of making (latent) variables measured with different instruments comparable – for example, by means of equating or simpler methods, such as linear stretching (Singh 2020).

Although these processes require great effort, *ex-post* survey harmonization projects have become more popular in the social sciences over the last few years. The substantive

1. Introduction

fields in which these projects have been conducted show the wide applicability and new popularity of the approach: IPUMS, the Integrated Public Use Microdata Series, harmonizes census and survey data from around the world; HaSpaD harmonizes and synthesizes partnership histories from different German surveys; ONBound focuses on combining individual-level data on national and religious identities from 20 international and national surveys as well as contextual data on included countries; and SDR, the Survey Data Recycling project, has harmonized measures of social capital, well-being, and political participation from several hundred national surveys all over the world. The aforementioned projects provide broad spatial and temporal data coverage in their respective fields. Other projects focus on the study of rare subpopulations. They include, for example, the International Ethnic and Immigrant Minorities' Survey Data Network (ethmig survey data), which harmonizes survey data on the integration of ethnic and migrant minorities, and InGRID-2, a project aimed *inter alia* at analyzing vulnerable and hard-to-reach groups in the labor market through ex-post survey harmonization. These six projects point to the advantages and possibilities of ex-post survey harmonization. Three particular benefits of this approach stand out: First, through the reuse, recycling, and combination of data (Law 2006; Slomczynski and Tomescu-Dubrow 2018), the robustness and reliability of research results can be increased. Second, through the increase in sample size (Wang et al. 2016), ex-post survey data harmonization allows the study of smaller subpopulations. And finally, third, ex-post survey harmonization enables the bridging (Singh 2020) of previous data gaps as far as the temporal and spatial dimensions are concerned. Quite literally, the whole is more than the sum of its parts. Regarding the increased robustness and reliability, ex-post survey harmonization is an important part of a larger movement to further advance secondary data analysis (Law 2006; Slomczynski and Tomescu-Dubrow 2018). Current research is often focused on producing innovative but singular research results instead of replications or re-analyses;

1. Introduction

cumulative knowledge growth is difficult under such conditions (Ioannidis 2005; Abbott 2007; Freese 2007; Wagner and Weiß 2014). However, a crucial part of modern statistics is to develop models that work well (under realistic violations of assumptions), improve inferential stability, and allow robust inference (Gelman and Vehtari 2021). Regarding inferential stability, ex-post survey harmonization allows more robust scientific results by testing them in several studies at once and achieving greater statistical power. Through multi-level modeling, researchers can incorporate information from different studies and reflect the heterogeneity of data sources simultaneously.

Ex-post survey harmonization is closely related to individual participant data (IPD) meta-analysis – that is, the *statistical analysis* of raw data from multiple scientific studies (Stewart and Tierney 2002; Stewart et al. 2012; Debray et al. 2015). One of the main differences between the two terms is their use in different scientific fields. While the term “ex-post survey harmonization” is used mainly in sociology and cross-cultural research (Wolf et al. 2016), the term “IPD meta-analysis” originated and is more common in medicine and psychology (Curran and Hussong 2009; Stewart et al. 2012). Many ex-post survey harmonization projects (e.g., IPUMS, ONBound, and SDR) focus on providing documentation and harmonization code or already harmonized data, thereby allowing other secondary users to conduct their own analyses with the harmonized data. This is less common in the field of IPD meta-analysis.

IPD meta-analyses have an established place in the so-called “evidence pyramid,” which ranks research results according to their reliability, see also Figure 1.1. With the rise of evidence-based medicine, the evidence level assigned to specific study designs has changed since the pyramid was first introduced (Murad et al. 2016; Shaneyfelt 2016). However, because of their strong reliability, systematic reviews and meta-analyses are usually placed at the top of the pyramid. The assignment of the highest evidence level to meta-analyses (or ex-post survey harmonization) has been challenged by some authors,

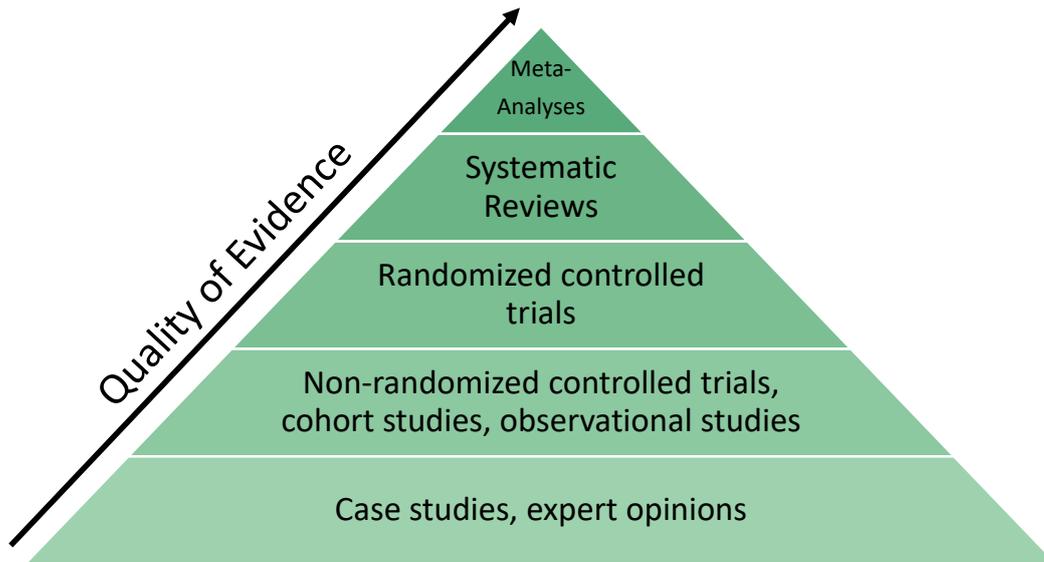


Figure (1.1) Evidence pyramid. Own depiction modeled after the “traditional” evidence pyramid in Murad et al. (2016).

who have expressed concern that researchers might be comparing apples and oranges due to the possibly great heterogeneity of the studies included in their meta-analyses (Paul and Leibovici 2014). I will return to this important concern and limitation in the concluding chapter.

As mentioned above, besides allowing a critical look at existing research findings, harmonized data sets also have a second potential benefit, namely, increased sample size, which can open up new research avenues. For example, harmonized data sets can enable researchers to conduct studies on specific or rare subpopulations, such as people with a less common ethnicity (e.g., the *ethmigsurveydata* project), vulnerable groups in the labor market (e.g., *InGRID2*), or people with a specific sexual orientation (e.g., *HaSpaD*). General social surveys usually have too few cases for these groups to obtain reliable, precise estimates. By combining data from several survey programs, researchers can achieve adequate sample sizes, even for rare subpopulations (Wang et al. 2016).

The third important benefit of ex-post survey harmonization is its bridging function –

that is, the fact that it can help to bridge gaps in temporal (or spatial) coverage. Because survey programs may have started too late or ended too early to cover all the years to be analyzed, or may not cover all the countries to be analyzed, the ex-post harmonization of data from several surveys may be required. The resulting harmonized data set spans the entire desired temporal or geographical space (see, e.g., the HaSpaD, IPUMS, ONBound, and SDR projects referenced above).

Having sketched the advantages of ex-post survey harmonization, I will now describe in general the missing data problems that can arise in this process and then provide more details on the exact methodological problems and solutions addressed in this dissertation.

1.2. Various flavors of missing data

One of the challenges researchers face in ex-post survey harmonization projects is one of the biggest challenges for survey researchers in general, namely, missing data. As Little and Rubin (2002, 4) put it: “Missing data are unobserved values that would be meaningful for analysis if observed; in other words, a missing value hides a meaningful value.” Carpenter and Kenward (2013) noted that the ubiquity of missing data, and the losses and problems it poses for analysis and inferences, have given rise to extensive statistical literature since the 1950s. These problems are also the central topic of this dissertation, mainly in the context of ex-post survey harmonization.

It is useful to distinguish between different types of data missingness. All data sets consist of *units*, each of which provides information on a series of *items*. For example, in a cross-sectional personal survey, individuals are the relevant units, and items are the individuals’ answers to the questions in the questionnaire. In a household survey, the units are households, and the items are answers to questions about the household and its members. In survey harmonization and IPD meta-analysis, researchers speak of *item nonresponse* or *partially* or *sporadically missing data* when data for some units are

missing for some items (Resche-Rigon and White 2018). The missing data then create a pattern in the data set that resembles a Swiss cheese. Variables that are missing for all units in one or more surveys are called *systematically missing variables* (Burgess et al. 2013; Resche-Rigon et al. 2013). If all items are missing for a unit, this is described as *unit nonresponse*. This dissertation offers solutions to missing data problems that belong to these three different types of missingness. To better explain the motivation and also the need for my dissertation, I will give a brief overview of the problematic consequences of the various missing data types before moving on to the specific missing data problems addressed.

1.2.1. Item nonresponse: Sporadically missing data

The first case of missing data that will be covered is sporadically missing data or item nonresponse. The consequences of missing data for analysis and inference were first laid out in a groundbreaking paper by Rubin (1976). In most software packages, the default approach to handling item nonresponse is to restrict the analysis to complete observations. Depending on the missingness pattern, the resulting information loss can be substantial. Even more importantly, however, the resulting subset of complete cases may no longer represent the target population (Carpenter and Kenward 2013).

Rubin (1976) also introduced a classification of missing data mechanisms that allows us to better understand the consequences of missing data for inference. He divided missing data mechanisms into three categories. If the missingness is independent of both observed and unobserved data, the observed data will still represent the target population. The missing data are then called *missing completely at random (MCAR)*.

If, given the observed data, the probability of data being missing is independent of the unobserved data, we speak of data *missing at random (MAR)*, and the analysis of the complete cases will likely be biased. Under MAR, it is generally possible to conduct a

1. Introduction

valid analysis that does not require the explicit inclusion of a model for the missingness mechanism (assumption of ignorability). Modern approaches to dealing with missing data, such as multiple imputation, generally start from the assumption of ignorability (Carpenter and Kenward 2013).

If data are neither MAR nor MCAR – that is, if the chance of an observation being missing is dependent on the value itself – even given the observed data – this is referred to as data *missing not at random* (MNAR, Rubin 1976) or not missing at random (NMAR). Strategies to handle data that are MNAR rely on finding more data about the reason for missingness or on performing sensitivity analyses. I will focus in this dissertation on data that are MCAR or MAR – that is, missing data that fulfill the assumption of ignorability.

Item nonresponse (i.e., sporadically missing data) can be a significant problem in the analysis of data from a single survey. This problem does not simply vanish if data from multiple surveys are pooled during ex-post survey harmonization. The tools used to handle sporadically missing data in pooled data sets are in principle the same as those for handling a data set from a single survey with a multi-level structure (see Resche-Rigon and White 2018). Therefore, I will now briefly introduce two of the most important methods for dealing with item nonresponse, namely, multiple imputation (MI) and maximum likelihood estimation (ML, Little and Rubin 2002; Schafer 1997).¹ These methods lead to unbiased estimates when assumptions are met and to variance estimates that account for increased variability due to missing data.

The MI approach consists of three main stages: first, an imputation step in which missing values are replaced with multiple sets of values to complete the data; second, a data analysis step, in which standard analyses are applied to each completed data set; third, a pooling step, in which the obtained parameters are pooled and combined, usually

¹Other principled methods are fully Bayesian methods (Clayton et al. 1998) and methods that model the missingness mechanism (Kenward 1998).

1. Introduction

using Rubin's rules (Rubin 1987). The goal of MI is not to impute values that are as close as possible to the missing values but rather to impute in a way that leads to valid statistical inference (Carpenter and Kenward 2013). The closest competitor to multiple imputation is maximum likelihood estimation (ML). In ML, dealing with the missing data and estimating parameters and standard errors are all done in a single step (Schafer 1997; Allison 2002; Little and Rubin 2002). A growing body of literature has dealt with differences in the results and the advantages and disadvantages of both approaches. However, in many applications, they lead to similar results (Schafer and Graham 2002).

1.2.2. Unit nonresponse: Completely missing data for a unit

Another of the most ubiquitous missing data problems in survey research is unit nonresponse. Like high item nonresponse rates, high unit nonresponse rates can lead to issues regarding the precision of estimates and potential nonresponse bias. Due to failed contact attempts or refusals of target units, it is almost impossible and practically unfeasible to reach and motivate all selected target units in a survey. During the last decades, nonresponse rates have risen, and participation rates have dwindled in most countries (Curtin et al. 2005; Singer 2006; De Leeuw and Hox 2018). Although the response rate often serves as a proxy for possible nonresponse bias, it is only one of two factors that influence the magnitude of nonresponse bias. The other one is the correlation between participation and the target variable. Whether one understands the division into respondents and non-respondents as a deterministic one or employs a stochastic model of survey participation, the nonresponse rate does not automatically determine the magnitude of the nonresponse bias (Kalton and Maligalig 1991; Valliant et al. 2013). There are several other methods to assess the risk of nonresponse bias (Wagner 2012). Standard methods include comparing survey estimates with estimates obtained from official statistics, such as, for example, the Microcensus in Germany (Blohm and Koch

1. Introduction

2015). For reasons of availability, these comparisons are often limited to sociodemographic variables. Other possibilities include comparisons between respondents and non-respondents regarding data such as paradata that are available for both groups (Kreuter et al. 2010).

If survey providers or researchers conclude that nonresponse bias due to unit nonresponse is a potential problem, methods to correct this bias are available, most prominently different weighting approaches, which I will briefly cover in the following. The different methods of adjusting for unit nonresponse through weighting can be divided into two main groups: (1) weighting class/propensity score adjustments (Kalton and Maligalig 1991; Little and Vartivarian 2003, 2005) and (2) post-stratification or raking (Deville and Särndal 1992; Kott 2009).

In weighting class or propensity score adjustment, variables are used that are available for both respondents and non-respondents. In weighting class adjustments, weighting classes or cells are formed from predicted response probabilities. Ideally, the variables used to predict the response probabilities are highly correlated to responding and to the variables used in substantive research. In propensity score adjustments, the inverse of the respondent's predicted propensity score is used for weighting (Kalton and Maligalig 1991; Little and Vartivarian 2003, 2005; Valliant et al. 2013).

On the other hand, post-stratification and raking use auxiliary data from official sources, for example, censuses, administrative records, or published statistics. In post-stratification, weighting classes, also called post-strata, are formed by crossing all categories of the auxiliary variables. The goal is to construct weights that restore the class-specific population counts in the weighted estimates. If there are many important auxiliary variables available, the adjustment cells may create too many variables. In that case, iterative raking adjusts the weight for each unit until the distribution of the sample matches the population distribution for the auxiliary variables. In addition, it only

1. Introduction

requires knowing the marginal proportions for each auxiliary variable used for weighting (Deville and Särndal 1992; Kott 2009; Valliant et al. 2013).

Many survey providers offer survey weights to researchers. Most include post-stratification or raking weights for the correction of possible nonresponse bias. As most survey programs select the units to be surveyed with varying probabilities, they also provide design weights or weights combined from design and post-stratification weights. Bias resulting from unit nonresponse may not only affect analyses with single data sets but also analyses with pooled data sets, and appropriate weighting steps may be necessary.

1.2.3. Systematically missing data: Completely missing data for a variable

Before moving on to the motivation for this dissertation and a brief summary of the main chapters, I will discuss the last type of missing data – systematically missing data – which occurs when a variable is not observed at all for one or more surveys (Resche-Rigon et al. 2013) or when measurements differ between surveys. In this case, the missingness resembles that due to item nonresponse, except that it is prevalent in specific subsets of the combined data set. Such gaps are easiest to fill if there is at least one survey that covers all the measurements or variables that exist in the surveys and are systematically missing for another survey. Score transformation from one instrument to the other can then be done through linear stretching or linear or equipercetile equating (Singh 2020). Other options are multi-level multiple imputation approaches that allow to model the heterogeneity between studies and use all available albeit only partially observed variables (Jolani et al. 2015; Jolani 2017).

A second problem with systematically missing data occurs in projects where a variable to be included is not measured in *any* survey. This can be due to different reasons. For example, the variable may be of low interest to most researchers or survey providers,

1. Introduction

or it could be hard to measure or observe (e.g., sensitive information). Regardless of the reason for the missingness, systematically missing data can lead to substantial problems in causal inference. If a confounder is completely missing, the assumption of unconfoundedness in causal inference is violated. This means that the probability of treatment depends on the potential outcomes. And when the pre-treatment set is not rich enough to remove systematic biases between treatment and control units, the causal claim is not sustainable. In practice, it is impossible to test the adequateness of the assumption of unconfoundedness. If there is reason to believe that the unconfoundedness assumption might be violated, the main analysis can be supported or called into question through supporting and supplementary analysis. One possibility is to drop the unconfoundedness assumption and replace it with other assumptions about the confoundedness of the causal claim (Imbens and Rubin 2015). Most of the time, rather than dropping the unconfoundedness assumption altogether, it is relaxed, leading to a range of plausible causal estimates of interest instead of single-point estimates.

1.3. Why this dissertation?

In the first part of this introductory chapter, I explained the importance of ex-post survey harmonization. Subsequently, I sketched three different types of missing data that are potentially problematic, both for research with single data sets and harmonized data sets. These possible problematic effects of missing data point to the need for methodological research aimed at resolving these issues. This dissertation aims to offer solutions for the missing data problems that ex-post survey harmonization projects face. In the following, I will situate my dissertation and its scientific merit within this dual framework – ex-post survey harmonization and missing data – and, where appropriate, outline its relevance for other fields of sociology. An overview is given in Table 1.1.

As shown by the wide range of exemplary ex-post survey harmonization projects referenced

1. Introduction

Table (1.1) Overview of the research areas that this dissertation contributes to.

		Analyses with harmonized data sets		Analyses with harmonized <i>or</i> single data sets		
		Analyses of pooled data with survey weights	Analyses under mid to high study heterogeneity	Discrete time survival analyses	Analyses with data from multi-actor studies	Causal inference with unobserved confounders
Missing data flavor	Item nonresponse			Chapter 2		
	Unit nonresponse	Chapter 5				
	Systematically missing variables		Chapter 3		Chapter 3	Chapter 4, Chapter 3

Note:

Chapter 2 – Multiple Imputation of Partially Observed Covariates in Discrete-Time Survival Analysis

Chapter 3 – Systematically Missing Partner Variables and Multiple Imputation Strategies: A Case Study With German Relationship Data

Chapter 4 – TippingSens: An R Shiny Application to Facilitate Sensitivity Analysis for Causal Inference Under Confounding

Chapter 5 – Better Together? Regression Analysis of Complex Survey Data After Ex-Post Harmonization

in Section 1.1. above (e.g., HaSpaD, IPUMS, ONBound, and SDR), researchers can conduct ex-post survey harmonization projects in very different substantive fields. The specific nature and urgency of the missing data problems mentioned above will differ across substantive areas. The ex-post survey harmonization project that has influenced this dissertation most is HaSpaD (Harmonizing and Synthesizing Partnership Histories From Different Research Data Infrastructures), which is funded by the German Research Foundation (DFG). That project focuses on separation factors in marriages and other partnerships over time. The methodological problems with missing data and the applied exemplary analyses in Chapters 2, 3, and 5 of this dissertation are thus rooted in this sociological subfield.

For example, Chapter 2 deals with sporadically missing data/item nonresponse in covari-

1. Introduction

ates for discrete-time survival analysis models (DTSAM). These models are employed in family sociology to study relationship events, such as divorce. However, they are also used to address completely different research questions where time is recorded in a discrete manner. As the term "survival" analysis suggests, the event to be studied could also be a patient's death during a medical study. The study presented in Chapter 2 is thus also highly relevant for quite distant substantive fields that are nonetheless connected through similar analytical models such as, in the present case, DTSAMs.

Chapter 3 is also deeply rooted in ex-post survey data harmonization in partnership research. It deals with systematically missing data in multi-actor surveys – specifically, with systematically missing variables for the partners of anchor respondents. This is an important question because data from secondary respondents are often not as extensive as those from anchor respondents, thereby hampering studies of hetero- and homophily between anchor and partner respondents, for example. The research question addressed in Chapter 3 is whether data from other studies (so-called bridging studies) can help with the imputation of systematically missing data for partner variables. In one section of the chapter, I also briefly examine the more fundamental question of whether study heterogeneity prevents the transfer of correlations observed in other studies.

Chapter 4 addresses a research question related to systematically missing variables and the subsequent possible violation of the unconfoundedness assumption. It presents an R Shiny (Chang et al. 2020) web application that I wrote to allow the simple execution of sensitivity analyses and the visualization of the results of these analyses. Again, this is a very common problem in other studies of causal relationships, and it can also be a problem in ex-post survey harmonization projects. The research question addressed in Chapter 5 is relevant for ex-post survey harmonization: How should unit nonresponse and survey weighting be dealt with after pooling and harmonizing data sets? Two prominent possibilities arise from the literature on IPD meta-analysis (Burke et al.

2017): The first possibility is to analyze the data from the individual surveys separately and then to combine the estimate in a standard meta-analysis; the second possibility is to pool the data sets and then to analyze the combined data set while modeling the heterogeneity in the combined data set. Although Burke et al. (2017) offered a comparison of advantages and disadvantages of IPD meta-analyses for experimental data in psychology and medicine, the two options have not been compared for survey data, much less for survey data with unequal sampling or inclusion probabilities. However, from the results of the study presented in Chapter 5, I will be able to close this gap with regard to ex-post survey harmonization and to recommend the second option, as it avoids the problematic assumption of known variances that comes with a standard meta-analysis.

In summary, the aim of this dissertation is to contribute to methodological research in ex-post survey harmonization (especially in family sociology) and to the more general methodological literature. Before moving on to the first study on multiple imputation of covariates in discrete-time survival analysis, extended summaries of the four main chapters are given in the next section.

1.4. Extended summary of chapters

In the following, each main chapter/study is presented with an extended summary.

Study 1 (Chapter 2): Multiple Imputation of Partially Observed Covariates in Discrete-Time Survival Analysis In Chapter 2, I examine a *sporadically missing data problem in the case of discrete-time survival analysis* (relevant both for the analysis of single and harmonized data sets with item nonresponse). Many phenomena in the social and medical sciences can be described as events – that is, qualitative changes that occur at a particular point in time. Typical research questions focus on whether, when, and under

1. Introduction

what circumstances events occur. The practical analysis of discrete-time survival data can be challenged by missing data in one or more covariates. The negative consequences of such missing data range from precision losses to bias. In this chapter, I will use the popular multiple imputation (MI) approach to circumvent these unwanted effects of missing data. With MI, it is generally crucial to include the outcome information in the imputation model for covariates of the substantive model. This is not straightforward in the case of discrete-time survival models for three reasons. First, the outcome is only partially observed because not all units experience the event – that is, the time-to-event data are censored. Second, there is not usually one outcome variable but two: the indicator of whether an event has happened and the variable in which the last observation is recorded. Third, the imputing researcher has to decide whether to impute while the data set is still in person-oriented format or after transformation into person-period format. In person-period format, which is also used for substantive analysis, there are multiple records for each person, one for each period observed.

In the literature, there is little guidance on how to incorporate the observed outcome information into the imputation model of missing covariates in discrete-time survival analysis for either single data sets or multiple combined data sets. Indeed, to my knowledge, the study presented in Chapter 2 is the first to investigate this problem systematically. Different approaches using fully conditional specification multiple imputation (FCS Buuren et al. 2006) and the newer substantive-model compatible fully conditional specification multiple imputation (SMC-FCS Bartlett et al. 2015). These approaches vary in their complexity and in the data format used during imputation. I compare the methods using Monte Carlo simulations and provide a practical example using data from the German Family Panel (pairfam). A compatible imputation model for SMC-FCS MI with data in person-period format proves to be the key to imputations with good performance results under different simulation conditions.

Study 2 (Chapter 3): Systematically Missing Partner Variables and Multiple Imputation Strategies: A Case Study With German Relationship Data

The study presented in Chapter 3 employs multiple imputation, not to tackle a sporadically missing data problem but rather to handle systematically missing data – that is, a variable missing for all units in a survey. I examine the particular case of *systematically missing information on partners*, a problem that can arise in multi-actor surveys. Multi-actor surveys collect information on persons who maintain a significant connection with each other, for example, by including the partner of an originally sampled anchor respondent as a so-called secondary respondent for relationship research (Kalmijn and Liefbroer 2010).

However, due to monetary and time restrictions in data collection, specific variables are often recorded only for anchor respondents but not for their partners. The variable for partners is thus *systematically missing* (Resche-Rigon et al. 2013), which gives rise to serious problems, for example, for studies researching homophily or heterophily between partners.

Systematically missing data are common in various settings, for example, after changes in measurement instruments in repeated surveys and in the context of ex-post survey harmonization if one or more surveys did not include the specific variable. Using MI techniques to impute the missing variables is a common approach in both cases (Schenker and Parker 2003; Resche-Rigon and White 2018).

I examine whether and how multiple imputation can be used when data on secondary respondents are systematically missing. I begin by reviewing previous research on systematically missing data and giving an overview of the critical differences in the pattern of these data. Building on this literature review, I outline MI approaches to handling systematically missing data from secondary respondents and illustrate these approaches by means of a simulation with data from the German Socio-Economic Panel

1. Introduction

(Goebel et al. 2019), pairfam – The German Family Panel (Brüderl et al. 2017), and SHARE, the Survey of Health, Ageing, and Retirement in Europe (Börsch-Supan et al. 2013; Börsch-Supan 2020). I will use the two latter ones as so-called bridging studies (Parker et al. 2004) in the simulation.

The results of the simulation show that imputation under the assumption of conditional independence for the anchor and partner variables leads to a strong bias toward zero in the estimated partial correlation between anchor and partner. Bridging studies similar to the original study in sampling and measurement can be used for the estimation of the partial correlation and lead to estimates with less bias after MI. However, the heterogeneity between studies may hinder the use of bridging studies. To obtain a better preliminary assessment of this problem, I take a short look at study heterogeneity regarding partial correlations between the three included surveys. As the handling of study heterogeneity in meta-analyses and (ex-post) survey harmonization is also a very important problem (DerSimonian and Laird 1986; Higgins 2003; Kontopantelis et al. 2013; Veroniki et al. 2016; Borenstein et al. 2017), this chapter also offers a contribution to this area of research.

Study 3 (Chapter 4): TippingSens: An R Shiny Application to Facilitate Sensitivity Analysis for Causal Inference Under Confounding Like the previous study, the study presented in Chapter 4 addresses the issue of systematically missing data. However, it deals with the case where an important confounder has not been observed for any respondent in a survey (or in any of the surveys, in the case of ex-post survey harmonization). Most strategies for causal inference based on quasi-experimental or observational data critically rely on the assumption of unconfoundedness. If this assumption is suspect, sensitivity analyses are an important tool to evaluate the impact of confounding on the analysis of interest. One of the earliest proposals for such a sensitivity analysis was suggested by Rosenbaum and Rubin (1983). However, whereas obtaining estimates

1. Introduction

for the causal effect under specific assumptions regarding an unobserved confounder is straightforward, conducting a full sensitivity analysis based on a range of parameter settings is unwieldy when, as in the case of Rosenbaum and Rubin (1983), simple forking tables are used.

To tackle the multiple parameter problem of Rosenbaum and Rubin's approach, I developed an interactive R Shiny (Chang et al. 2020) application called TippingSens, which visualizes the impact of various parameter settings on the estimated causal effect. Borrowing from the literature on tipping point analysis, this flexible app facilitates manipulating all parameters simultaneously.

After presenting an overview of possible sensitivity analyses, I demonstrate the usefulness of the app by conducting a sensitivity analysis for a quasi-experiment measuring the effect of vocational training programs on unemployed men, with heavy alcohol consumption as a possible unobserved confounder. A step-by-step introduction to the app and a vignette using medical data from Rosenbaum and Rubin (1983) are provided in the Appendix.

Study 4 (Chapter 5): Better Together? Regression Analysis of Complex Survey Data After Ex-post Survey Harmonization In Chapter 5, I present the fourth study conducted within the framework of this dissertation. It deals with missingness in the form of unit nonresponse. Specifically, I address the pooling of complex survey data with their accompanying survey weights in the context of (ex-post) survey harmonization in sociology, and I investigate how to conduct a regression analysis of pooled complex survey data after ex-post survey harmonization (Granda et al. 2010). The meta-analysis literature suggests two different approaches to the regression analysis of pooled raw data. The first approach entails combining estimated regression coefficients from the single data sets (so-called two-stage approach); the second entails estimating a regression on the combined data sets (so-called one-stage approach). Although there have been comparisons and evaluations of the two general methods of conducting meta-analyses

1. Introduction

with raw data (Burke et al. 2017), there have been no comparisons or evaluations for the case of survey-weighted data or regressions.

This chapter builds bridges between the fields of survey harmonization, survey statistics, and meta-analysis, making research results and approaches from one area useful to the other. Its main contributions are threefold. First, I present to the ex-post survey harmonization community two approaches from the meta-analysis literature – one-stage and two-stage meta-analysis for individual participant data (IPD). Second, I study the performance of these two approaches through using a Monte Carlo simulation. I show that the distribution of survey weights in the respective data set plays a role when these two meta-analytical approaches are used. Unless the coefficient of variance (i.e., the ratio of the standard deviation to the mean) for the survey weights is small, the assumption of known within-study variances for two-stage analysis is problematic and will result in biased point estimates. And finally, third, the difference between the two approaches is demonstrated exemplarily with a real-world example of same-sex couples and family satisfaction in Germany.

Conclusion This dissertation concludes with a discussion of the results, scientific contributions, and limitations of the studies presented in Chapters 2–5 and a broader outlook on the issue of study heterogeneity.

Bibliography

Abbott, A. (2007). Notes on Replication. *Sociological Methods & Research*, 36(2):210–219.

Allison, P. (2002). *Missing Data*. SAGE.

Bartlett, J. W., Seaman, S. R., White, I. R., and Carpenter, J. R. (2015). Multiple Imputation of Covariates by Fully Conditional Specification: Accommodating the Substantive Model. *Statistical Methods in Medical Research*, 24(4):462–487.

Blohm, M. and Koch, A. (2015). Führt eine höhere Ausschöpfung zu anderen Umfrageergebnissen? In Schupp, J. and Wolf, C., editors, *Nonresponse Bias*, pages 85–129. Springer.

Borenstein, M., Higgins, J. P. T., Hedges, L. V., and Rothstein, H. R. (2017). Basics of Meta-Analysis: I^2 Is Not an Absolute Measure of Heterogeneity. *Research Synthesis Methods*, 8(1):5–18.

Brüderl, J., Hank, K., Huinink, J., Nauck, B., Neyer, F. J., Walper, S., Alt, P., Borschel, E., Buhr, P., Castiglioni, L., Friedrich, S., Finn, C., Garrett, M., Hajek, K., Herzig, M., Huyer-May, B., Lenke, R., Müller, B., Peter, T., Schmiedeberg, C., Schütze, P., Schumann, N., Thönnissen, C., Wetzels, M., and Wilhelm, B. (2017). The German Family Panel (pairfam). Technical Report ZA5678 Data file Version 8.0.0, GESIS Data Archive, Cologne.

Bibliography

- Burgess, S., White, I. R., Resche-Rigon, M., and Wood, A. M. (2013). Combining Multiple Imputation and Meta-Analysis With Individual Participant Data. *Statistics in Medicine*, 32(26):4499–4514.
- Burke, D. L., Ensor, J., and Riley, R. D. (2017). Meta-Analysis Using Individual Participant Data: One-Stage and Two-Stage Approaches, and Why They May Differ. *Statistics in Medicine*, 36(5):855–875.
- Buuren, S. V., Brand, J. P., Groothuis-Oudshoorn, C. G., and Rubin, D. B. (2006). Fully Conditional Specification in Multivariate Imputation. *Journal of Statistical Computation and Simulation*, 76(12):1049–1064.
- Börsch-Supan, A. (2020). Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 5. Release Version: 7.1.0. SHARE-ERIC. Data Set.
- Börsch-Supan, A., Brandt, M., Hunkler, C., Kneip, T., Korbmacher, J., Malter, F., Schaan, B., Stuck, S., and Zuber, S. (2013). Data Resource Profile: The Survey of Health, Ageing and Retirement in Europe (SHARE). *International Journal of Epidemiology*, 42(4):992–1001.
- Carpenter, J. and Kenward, M. (2013). *Multiple Imputation and its Application*. Wiley, Hoboken.
- Chang, W., Cheng, J., Allaire, J., Xie, Y., and McPherson, J. (2020). *shiny: Web Application Framework for R*. R Package Version 1.5.0.
- Clayton, D., Spiegelhalter, D., Dunn, G., and Pickles, A. (1998). Analysis of Longitudinal Binary Data From Multiphase Sampling. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1):71–87.
- Curran, P. and Hussong, A. (2009). Integrative Data Analysis: The Simultaneous Analysis of Multiple Data Sets. *Psychological methods*, 14:81–100.

Bibliography

- Curtin, R., Presser, S., and Singer, E. (2005). Changes in Telephone Survey Nonresponse Over the Past Quarter Century. *Public Opinion Quarterly*, 69(1):87–98.
- De Leeuw, E. and Hox, J. (2018). International Nonresponse Trends Across Countries and Years: An Analysis of 36 Years of Labour Force Survey Data. *Survey Insights, Methods from the Field (SMIF)*.
- Debray, T. P. A., Moons, K. G. M., van Valkenhoef, G., Efthimiou, O., Hummel, N., Groenwold, R. H. H., Reitsma, J. B., and GetReal Methods Review Group (2015). Get Real in Individual Participant Data (IPD) Meta-Analysis: A Review of the Methodology. *Research Synthesis Methods*, 6(4):293–309.
- DerSimonian, R. and Laird, N. (1986). Meta-Analysis in Clinical Trials. *Controlled Clinical Trials*, 7(3):177–188.
- Deville, J.-C. and Särndal, C.-E. (1992). Calibration Estimators in Survey Sampling. *Journal of the American Statistical Association*, 87(418):376–382.
- Dubrow, J. K. and Tomescu-Dubrow, I. (2015). The Rise of Cross-National Survey Data Harmonization in the Social Sciences: Emergence of an Interdisciplinary Methodological Field. *Quality & Quantity*, 50(4):1449–1467.
- ethmig survey data (2020). The International Ethnic and Immigrant Minorities' Survey Data Network. <https://ethmigsurveydatahub.eu/>. [Online; accessed 20-September-2020].
- Freese, J. (2007). Replication Standards for Quantitative Social Science: Why Not Sociology? *Sociological Methods & Research*, 36(2):153–172.
- Friedrichs, J. and Nonnenmacher, A. (2010). *Welche Mechanismen erklären Kontexteffekte?*, pages 469–497. VS Verlag für Sozialwissenschaften, Wiesbaden.

Bibliography

- Gelman, A. and Vehtari, A. (2021). What Are the Most Important Statistical Ideas of the Past 50 Years? arXiv 2012.00174.
- Goebel, J., Grabka, M. M., Liebig, S., Kroh, M., Richter, D., Schröder, C., and Schupp, J. (2019). The German Socio-Economic Panel (SOEP). *Jahrbücher für Nationalökonomie und Statistik*, 239(2).
- Granda, P., Wolf, C., and Hadorn, R. (2010). *Harmonizing Survey Data*, chapter 17, pages 315–332. John Wiley & Sons.
- HaSpaD (2020). HaSpaD - Harmonizing and Synthesizing Partnership Histories From Different Research Data Infrastructures. <https://www.gesis.org/forschung/drittmittelprojekte/projektuebersicht-drittmittel/haspad-harmonisierung-und-synthese-von-paarbiografischen-daten>. [Online; accessed 20-September-2020].
- Higgins, J. P. T. (2003). Measuring Inconsistency in Meta-Analyses. *BMJ*, 327(7414):557–560.
- Hussong, A. M., Curran, P. J., and Bauer, D. J. (2013). Integrative Data Analysis in Clinical Psychology Research. *Annual Review of Clinical Psychology*, 9:61–89.
- Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, chapter 22, pages 496—510. Cambridge University Press.
- InGRID-2 (2020). Integrating Research Infrastructure for European Expertise on Inclusive Growth From Data to Policy. <http://www.inclusivegrowth.eu/>. [Online; accessed 20-September-2020].
- Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLoS Medicine*, 2(8):e124.

Bibliography

- IPUMS (2020). Integrated Public Use Microdata Series. <https://www.maelstrom-research.org/>. [Online; accessed 20-September-2020].
- Jolani, S. (2017). Hierarchical Imputation of Systematically and Sporadically Missing Data: An Approximate Bayesian Approach Using Chained Equations. *Biometrical Journal*, 60(2):333–351.
- Jolani, S., Debray, T., Koffijberg, H., van Buuren, S., and Moons, K. (2015). Imputation of Systematically Missing Predictors in an Individual Participant Data Meta-Analysis: A Generalized Approach Using MICE. *Statistics in Medicine*, 34(11):1841–1863.
- Kalmijn, M. and Liefbroer, A. C. (2010). Nonresponse of Secondary Respondents in Multi-Actor Surveys: Determinants, Consequences, and Possible Remedies. *Journal of Family Issues*, 32(6):735–766.
- Kalton, G. and Maligalig, D. S. (1991). A Comparison of Methods of Weighting Adjustment for Nonresponse. In *Proceedings of the 1991 Annual Research Conference*, volume 409428. US Bureau of the Census.
- Kenward, M. G. (1998). Selection Models for Repeated Measurements With Non-Random Dropout: An Illustration of Sensitivity. *Statistics in Medicine*, 17(23):2723–2732.
- Kontopantelis, E., Springate, D. A., and Reeves, D. (2013). A Re-Analysis of the Cochrane Library Data: The Dangers of Unobserved Heterogeneity in Meta-Analyses. *PLoS ONE*, 8(7).
- Kott, P. S. (2009). Calibration Weighting: Combining Probability Samples and Linear Prediction Models. In *Handbook of Statistics*, volume 29, pages 55–82. Elsevier.
- Kreuter, F., Olson, K., Wagner, J., Yan, T., Ezzati-Rice, T. M., Casas-Cordero, C., Lemay, M., Peytchev, A., Groves, R. M., and Raghunathan, T. E. (2010). Using

Bibliography

- Proxy Measures and Other Correlates of Survey Outcomes to Adjust for Non-Response: Examples from Multiple Surveys. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 173(2):389–407.
- Law, M. (2006). Reduce, Reuse, Recycle: Issues in the Secondary Use of Research Data. *IASSIST Quarterly*, 29(1):5–5.
- Little, R. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. Wiley.
- Little, R. J. and Vartivarian, S. (2003). On Weighting the Rates in Non-Response Weights. *Statistics in Medicine*, 22(9):1589–1599.
- Little, R. J. and Vartivarian, S. (2005). Does Weighting for Nonresponse Increase the Variance of Survey Means? *Survey Methodology*, 31(2):161.
- Murad, M. H., Asi, N., Alsawas, M., and Alahdab, F. (2016). New Evidence Pyramid. *BMJ Evidence-Based Medicine*, 21(4):125–127.
- ONBound (2020). ONBound - Old and New Boundaries: National Identities and Religion. <https://www.gesis.org/en/services/processing-and-analyzing-data/data-harmonization/onbound>. [Online; accessed 29-December-2020].
- Parker, J., Schenker, N., Ingram, D., Weed, J., Heck, K., and Madans, J. (2004). Bridging Between Two Standards for Collecting Information on Race and Ethnicity: An Application to Census 2000 and Vital Rates. *Public Health Reports*, 119(2):192–205.
- Paul, M. and Leibovici, L. (2014). Systematic Review or Meta-Analysis? Their Place in the Evidence Hierarchy. *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases*, 20:97–100.
- Resche-Rigon, M., White, I., Bartlett, J., Peters, S., and Thompson, S. (2013). Multiple

Bibliography

- Imputation for Handling Systematically Missing Confounders in Meta-Analysis of Individual Participant Data. *Statistics in Medicine*, 32(28):4890–4905.
- Resche-Rigon, M. and White, I. R. (2018). Multiple Imputation by Chained Equations for Systematically and Sporadically Missing Multilevel Data. *Statistical Methods in Medical Research*, 27(6):1634–1649.
- Rosenbaum, P. R. and Rubin, D. B. (1983). Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study With Binary Outcome. *Journal of the Royal Statistical Society. Series B (Methodological)*, 45(2):212–218.
- Rubin, D. B. (1976). Inference and Missing Data. *Biometrika*, 63(3):581–592.
- Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley.
- Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman and Hall.
- Schafer, J. L. and Graham, J. W. (2002). Missing Data: Our View of the State of the Art. *Psychological Methods*, 7(2):147.
- Schenker, N. and Parker, J. (2003). From Single-Race Reporting to Multiple-Race Reporting: Using Imputation Methods to Bridge the Transition. *Statistics in Medicine*, 22(9):1571–1587.
- SDR (2020). Survey Data Recycling. <https://www.asc.ohio-state.edu/dataharmonization/>. [Online; accessed 20-September-2020].
- Shaneyfelt, T. (2016). Pyramids Are Guides Not Rules: The Evolution of the Evidence Pyramid. *BMJ Evidence-Based Medicine*, 21(4):121–122.
- Singer, E. (2006). Introduction: Nonresponse Bias in Household Surveys. *Public Opinion Quarterly*, 70(5):637–645.

Bibliography

- Singh, R. K. (2020). Harmonizing Instruments with Equating. *Harmonization Newsletter on Survey Data Harmonization in the Social Sciences*, 6(1).
- Slomczynski, K. M. and Tomescu-Dubrow, I. (2018). *Basic Principles of Survey Data Recycling*, chapter 43, pages 937–962. John Wiley & Sons, Ltd.
- Stewart, G. B., Altman, D. G., Askie, L. M., Duley, L., Simmonds, M. C., and Stewart, L. A. (2012). Statistical Analysis of Individual Participant Data Meta-Analyses: A Comparison of Methods and Recommendations for Practice. *PLOS ONE*, 7(10):1–8.
- Stewart, L. A. and Tierney, J. F. (2002). To IPD or not to IPD?: Advantages and Disadvantages of Systematic Reviews Using Individual Patient Data. *Evaluation & the Health Professions*, 25(1):76–97.
- Valliant, R., Dever, J. A., and Kreuter, F. (2013). *Practical Tools for Designing and Weighting Survey Samples*. Springer International Publishing, Cham.
- Veroniki, A. A., Jackson, D., Viechtbauer, W., Bender, R., Bowden, J., Knapp, G., Kuss, O., Higgins, J. P., Langan, D., and Salanti, G. (2016). Methods to Estimate the Between-Study Variance and its Uncertainty in Meta-Analysis. *Research Synthesis Methods*, 7(1):55–79.
- Wagner, J. (2012). A Comparison of Alternative Indicators for the Risk of Nonresponse Bias. *Public Opinion Quarterly*, 76(3):555–575.
- Wagner, M. and Weiß, B. (2014). Meta-Analyse. In Baur, N. and Blasius, J., editors, *Handbuch Methoden der empirischen Sozialforschung*, pages 1117–1126. Springer Fachmedien Wiesbaden, Wiesbaden.
- Wang, X. V., Cole, B., Bonetti, M., and Gelber, R. D. (2016). Meta-STEPP: Subpopulation Treatment Effect Pattern Plot for Individual Patient Data Meta-analysis. *Statistics in Medicine*, 35(21):3704–3716.

Bibliography

Wolf, C., Schneider, S. L., Behr, D., and Joye, D. (2016). Harmonizing Survey Questions Between Cultures and Over Time. In *The SAGE Handbook of Survey Methodology*, pages 502–524. SAGE.

2. Multiple Imputation of Partially Observed Covariates in Discrete-Time Survival Analysis

2.1. Introduction

Many phenomena in the social and medical sciences can be characterized as events – that is, qualitative changes that occur at some point in time. Typical research questions focus on whether, when, and under what circumstances events occur. Examples of sociologically relevant events are divorce or a job offer after a period of unemployment. When analyzing such time-to-event (or survival time) data, one cannot rely on a simple linear regression model, as the time-to-event is missing (censored) for parts of the population. For example, some married people never experience divorce, and although everybody dies, data collection will almost certainly not continue until this point for all observations. Different analytical approaches have been developed to deal with the censoring problem. They include, for example, Cox regression for continuous-time survival analysis (Cox 1972). Cox (1972) also extended the proportional hazard model for discrete-time survival analysis, analyzing the conditional odds of an event occurring at a particular time point, given survival up to that point. For discrete-time survival analysis, data must first be

converted from the familiar person-oriented format (one row for each person/observational unit) to a person-period format (one row for each period in which a person was observed). One challenge that arises in the application of these survival models (and in other models) is that often one or more covariates have missing data. Simplistic approaches, such as listwise deletion (LD) and unconditional mean imputation, are still used in the social sciences (e.g. Böttcher 2006; Cooke 2006; Arranz Becker and Lois 2010; Manning et al. 2016; Cooper et al. 2018; Stoddard and Veliz 2019). However, these approaches may be highly inefficient or lead to severely biased variance estimates. Point estimates may also be biased after listwise deletion if missingness depends on the outcome (Hughes et al. 2019).

Another approach to handling missing data is multiple imputation (Rubin 1987, 1996; Schafer 1997; van Buuren et al. 1999). Multiple imputation leads to unbiased point and variance estimates if certain conditions concerning the missing data mechanism and the imputation model are met (Allison 2000). However, while there has been research on how best to impute missing covariate values for Cox regressions (van Buuren et al. 1999; Clark and Altman 2003; White and Royston 2009; Keogh and Morris 2018), the Cox cure model (Beesley et al. 2016) and the relative survival model (Nur et al. 2009), the imputation of covariates in discrete-time survival analysis is still understudied. For time-varying covariates, Murad et al. (2019) showed that multiple imputation approaches using information from the previous and current time points seem sufficient in most situations. However, in discrete-time survival analysis, not only time-varying covariates but also time-invariant covariates are used. *This article contributes to the literature by exploring how to specify a suitable imputation model for partially missing time-invariant covariates in discrete-time survival analysis.*

This is not an easy or straightforward task. Kenward and Carpenter (2007, 2007) shows that including outcome information in the imputation model for partially observed

covariates is crucial for unbiased estimates. However, with discrete-time survival models, the two outcome variables (event and time-to-event) are not fully observed due to censoring. Nor is it clear in which format the imputation procedure should be carried out: with data in person format or person-period format? And how should the relationship between the time-to-event variable and the covariates be modeled for imputation?

These are relevant issues, as discrete-time survival analysis is widely used in the social sciences, especially in family sociology (see, e.g., Barber 2001; Schoen et al. 2002; Cooke 2006; Nomaguchi 2006; Arranz Becker and Lois 2013), and also in the medical sciences (Murad et al. 2019).

The remainder of this article is organized as follows: In the next section, we introduce the formalization of the discrete-time survival analysis model (DTSAM) proposed by Singer and Willett (1993), we address person-oriented and person-period data set formats, and we briefly discuss the negative consequences of, and ways of dealing with missing covariate data. We then outline the method of multiple imputation and present several possible imputation approaches that differ in terms of the data format used and the specification of the imputation model. Following this, we conduct four simulations with discrete-survival data and varying degrees of unobserved heterogeneity. Using data from the German Family Panel (pairfam), we then provide an applied example with real-world data. The article concludes with a discussion of results and further steps.

2.2. Discrete-time survival analysis model

2.2.1. The model

In this section, we introduce the formalization of a discrete-time survival analysis model (DTSAM) proposed by Singer and Willett (1993).

The term discrete survival is used when the time-to-event can take only distinct values,

for example, one, two, three, or more years/semesters/weeks. Occasionally, discrete survival data are “truly discrete” (Kleinbaum and Klein 2012, 325); that is, the event can occur only at distinct values of time (e.g., fertility modeling, particularly the time from puberty to first childbirth). However, in most cases, discrete data are the result of interval-censoring: Events might happen in a continuous range of time, but they were observed only in grouped form instead of continuous-time data – for example, the year of divorce is recorded but not the month and day (Kleinbaum and Klein 2012, 318).

Following Singer and Willett (1993, 163), let T be a discrete random variable that indicates the time period j when the event occurs for a randomly selected individual from the population. For example, T could be the time until divorce in a person’s first marriage. Note that we focus here on non-repeatable events, and that event occurrence is thus inherently conditional. For instance, a person can experience the divorce of their first marriage only if he or she did not already experience it in any of the periods prior to j . We aim to describe T by a conditional probability density function. The conditional probability that an event will occur in each period given that it has not occurred earlier is called the discrete-time hazard, h_j .

Researchers are usually interested in whether the risk of event occurrence differs systematically between observations. For example, in a study of divorce risks, the risk might depend on age at the beginning of a marriage or on differences in social status between the spouses. For now, we examine *time-invariant predictors*. We have to distinguish between different individuals i , each with their predictor values X_{ki} for K ($k = 1, \dots, K$) predictors X_k .

We model the individual hazard h_{ij} as depending on periods P_j and predictors X_k through a logit link (Cox 1972; Allison 1982; Singer and Willett 1993).

The model to be estimated is that proposed by Singer and Willett (2003, 317):

$$\log_e \left(\frac{h_{ij}}{1 - h_{ij}} \right) = (\alpha_1 P_{1ij} + \alpha_2 P_{2ij} + \dots + \alpha_J P_{Jij}) + \quad (2.1)$$

$$(\beta_1 X_{1ij} + \beta_2 X_{2ij} + \dots + \beta_K X_{Kij}). \quad (2.2)$$

This model contains no single stand-alone intercept but rather a set of alpha parameters $[\alpha_1, \alpha_2, \dots, \alpha_J]$, each of which acts as an intercept for a specific time period. The β s indicate how much the logit-hazards shift with a unit shift in the parameters – for example, how much an additional year in age difference between partners shifts the logit of the hazard of divorce.

2.2.2. The data: Transformations and the person-period format

Having established our model, we take a closer look at our data and the data format needed for a DTSAM. Data are typically available in a person-oriented format, with one row (record) for each observational unit (i.e., person). Apart from the covariates, X_k , there are three possible outcome variables. First, there is the variable T , that is, the time-to-event. However, in most applications, we would almost certainly have a lot of missing data in T due to censoring. Usually, therefore, instead of the variable T , the variable Y – the last period in which a unit was observed – is used. This variable is observed regardless of whether the event occurred or censoring happened. We also add a dichotomous event indicator, E , set to one if the event occurred and to zero if the observation is censored.

To estimate a DTSAM (our substantive model), the person-oriented data set must be converted into a *person-period format* (Singer and Willett 1993, 172). In person-period format (see Table 2.1), there is a separate row for each period in which a unit was observed. Apart from the covariates, we usually have a variable indicating the observed

Table (2.1) Exemplary data set in person-period (PP) format

Obs	X_1	X_2	X_3	X_4	X_5	P	E	Y
1	9,3	3	female	rural	6	1	0	4
1	9,3	3	female	rural	6	2	0	4
1	9,3	3	female	rural	6	3	0	4
1	9,3	3	female	rural	6	4	1	4
2	4,1	2	male	urban	3	1	0	6
2	4,1	2	male	urban	3	2	0	6
2	4,1	2	male	urban	3	3	0	6
2	4,1	2	male	urban	3	4	0	6
2	4,1	2	male	urban	3	5	0	6
2	4,1	2	male	urban	3	6	0	6
3	2,1	4	male	urban	4	1	1	1
...
...
n
n

Note: Although the last period in which a unit was observed, Y , is not usually included, we need it for certain imputation approaches. However, it is used only for imputation, not for analysis. X_1 – X_5 are time-invariant variables.

period, P , and an event indicator, E . The event indicator is set to one only if the event occurred for this unit in this specific period. Although the last period in which a unit was observed, Y , is not usually included separately in person-period format, we will need it for certain imputation approaches. Y as an added variable in person-period format is exclusively used for imputation, not for analysis.

The dichotomous event indicator, E , is treated as a collection of independent values with a hypothesized logistic dependence on predictors (Singer and Willett 1993, 174). This model implicitly asserts that the variables exhaust all the sources of individual variation in the hazard rate. For example, the model implies that the variation in hazards is due *only* to differences in the independent variables and period, and that the model is correctly specified. The survival model literature describes these models as having no unobserved heterogeneity (Allison 1982, 82). The omission of a critical predictor of the outcome from the model is equivalent to mixing hazard profiles for the different populations defined by the discarded predictor values. The pooled hazard profile does

not have to look like any member of the general population. As an example, assume that all members of the population have a flat risk profile. Still, the height of the risk profile differs between members. Over time, members with a high risk drop out of the population. If we do not include the predictor that shifts the members' risk, the aggregate profile will show a risk profile that decreases with time (Singer and Willett 1993, 185). Building a model that exhausts *all sources of individual variation* is practically not feasible. We not *only* have to worry about possible confounders missing from our model, we also have to worry about all direct predictors of survival. However, even in the absence of important predictors, parameters can still be interpreted as an average across population hazard profiles (Xue and Brookmeyer 1996). For example, if we include obesity as a risk factor for diabetes in our model, and there is unobserved heterogeneity, we would estimate the log odds ratio for all people with obesity versus those without. If there were no unobserved heterogeneity, we could additionally interpret the regression coefficient as the log odds ratio for an individual before versus after developing obesity. Apart from complicating the interpretation of factors, Allison (1982, 83) also noted that in case of unobserved heterogeneity, "one would expect this dependence among the observations to lead to inefficient coefficient estimates and standard errors that are biased downward." As unobserved heterogeneity is not entirely avoidable, we will generate the data sets in our simulations with varying degrees of unobserved heterogeneity. However, before testing the different multiple imputation approaches, we take a closer look at the implications of missing data for discrete-time survival analysis.

2.2.3. Implications of missing data for discrete-time survival analysis

Surveys are often subject to missing data, and we have to decide how to treat partially observed covariates in discrete-time survival analysis. Note that, in this article, we are looking only at missing data in the covariates, not in the time-to-event variable. Multiple

imputation is not suited for missing data due to censoring in survival analysis (Allison 2010).

Before exploring different possible strategies for imputing partly missing covariates in discrete-time survival analysis, we take a more general look at the implications of missing data in regression analysis to demonstrate the need for a suitable multiple imputation strategy. One way to deal with missing data in regression analysis is to exclude incomplete observations. This is known as listwise deletion or complete case analysis. One consequence of this approach is a possibly considerable loss of efficiency, or information (Carpenter and Kenward 2013, 9) as only the observations with complete records are used. In other words, even if only one of many covariates is missing, this leads to the complete exclusion of the observation. This becomes especially problematic if the aim is to keep unobserved heterogeneity down and include not only possible confounders but also other essential predictors.

However, a loss of efficiency, or information, is only one side of the coin; the other is bias. Table 2.2 gives an overview of the conditions under which logistic regression coefficients will be biased after listwise deletion. Here, X_1 and X_2 are two independent variables, Y is the dependent binary outcome variable. For now, only the variables on which the completeness of a case depends (e.g., the outcome and X_1) are of relevance, not the type of missingness mechanism that is behind the nonresponse.

From Table 2.2, we can see that in the case of logistic regression – for example, the DTSAM – the coefficient estimate of an independent variable X_1 is biased only if the completeness of an observation depends

1. on the binary outcome and
2. and on the covariate X_1 itself.

Concerning bias, it does not make a difference for analysis results after listwise deletion whether X_1 is missing not at random (MNAR) dependent on X_1 and the outcome, or

Table (2.2) Bias in case of logistic regression using complete records. Adapted from Carpenter and Kenward (2013, 32)

Mechanism depends on	Biased point estimates?		
	Constant	Coeff. X_1	Coeff. X_2
Y	Yes	No	No
X_1	No	No	No
X_2	No	No	No
X_1, X_2	No	No	No
Y, X_1	Yes	Yes	No
Y, X_2	Yes	No	Yes
Y, X_1, X_2	Yes	Yes	Yes

whether X_2 is missing at random (MAR) dependent on X_1 and the outcome (Carpenter and Kenward 2013, 32).

Therefore, if it must be assumed that the estimated coefficient of our variable X_1 will be biased after listwise deletion, other approaches to handling missing data are needed. Nevertheless, even if we are confident that our analysis will not be biased after listwise deletion, we will lose information from incomplete cases. Thus, listwise deletion cannot be an adequate solution in most cases, and we will look into multiple imputation as a possible remedy for these unwanted effects of missing data.

2.3. Multiple imputation

2.3.1. Multiple imputation in general

One of the most popular approaches to tackle missing data is multiple imputation (Little and Rubin 2002). MI allows for the analysis of incomplete data sets by substituting missing values. Substitution is performed by imputing values of a variable based on other variables, mostly those of the analysis model. In the analyses, these imputed values are not treated the same as observed values, as this would lead to biased variance estimates. Therefore, several values instead of just one are imputed for each missing value in order

to avoid treating imputed values as observed. Each data set is then analyzed separately, and estimates and variances are combined across imputations using rules developed by Rubin (1987).

To impute missing values, we need models of their distribution. There are two main approaches, joint modeling (JM) and fully conditional specification (FCS), also known as multivariate imputation by chained equations or MICE (van Buuren 2007).¹ Joint modeling MI (Schafer 1997) draws missing values simultaneously for all incomplete variables using a multivariate distribution. However, specifying such a joint model is often challenging, for example, for categorical variables. By contrast, FCS divides the problem into a series of univariate problems (van Buuren 2007). FCS involves specifying a series of univariate models for the conditional distribution of each partially observed variable, given all the other variables (White et al. 2011). It is more flexible than the joint modeling approach because adequate regression models can be selected for every variable (e.g., linear regression for continuous partially observed variables, logistic regression for binary partially observed variables).

When specifying the imputation model, it is crucial to account for the *substantive model* of interest – which is often also called the *analytical model*. The associations to be examined in the substantive model must also be represented in the imputation model. Otherwise, bias toward zero will be the likely consequence (Fay 1992). The imputation and substantive models should be *compatible*. Compatible means that there exists a joint model with conditionals corresponding to the imputation and substantive models (see Bartlett et al. 2014 and for the related term of congeniality see Meng 1994). In addition to JM and FCS, a variation of FCS that allows to easier specify a compatible imputation model was developed by Bartlett et al. (2014). It is called *substantive model compatible fully conditional specification* (SMC-FCS). SMC-FCS is used when it is hard to find a compatible standard FCS imputation model because the substantive model, i.e.,

¹FCS is known under a multitude of names, e.g., also as imputation by chained equations or “ice”.

the model the researcher is interested in, is either non-linear (e.g., a Cox regression) or contains non-linear (e.g., squared or interaction) terms. SMC-FCS can also be used for models without non-linear terms. As with FCS, separate models are specified for the partially observed variables. What differentiates SMC-FCS from regular FCS is that the conditional distribution of a variable (given the other variables) is combined with the specified substantive model to define an imputation model. This combination ensures that the missing data are drawn from models *compatible* with the specified substantive model.

2.3.2. Handling missing covariates values in case of a DTSAM as a substantive model

Representing the associations in the substantive model of interest is not straightforward in the case of a DTSAM and its complicated outcome structure. Generally, we have two outcome variables: the event indicator, E , and the last period in which a unit was observed, Y (either because the unit was subsequently censored or the event occurred during that period). Moreover, it is not clear whether to impute in a person-oriented (P) or person-period (PP) format. Imputing while the data set is still in a person-oriented format would lead to imputed values that are identical for all observed periods. However, it is not possible to specify a compatible imputation model in that case because the substantive analysis model – a DTSAM – is estimated with the data set in person-period format. Imputing with the data set in person-period format would, however, lead to values for time-invariant variables that potentially differ between persons.

As there is no clear-cut FCS solution to this problem, we will explore several imputation FCS approaches/approximations that differ in terms of (1) the data format used, (2) the general imputation approach used, and (3) the imputation model specification. The data format in which we impute has two possible formats: person-oriented and person-period

format. We examine both the FCS and the SMC-FCS approaches. Our imputation models differ in terms of whether we include variables for different periods or (censored) survival times or whether we also treat the (uncensored) time-to-event, T , as a partially observed covariate (Beesley et al. 2016) and impute conditional on this partially imputed variable.

FCS specifications with data in person-oriented format

We begin with the imputation approaches in a person-oriented format (P), that is, before converting the data to a person-period (PP) format. We present several approaches, some of which are taken from the literature on MI with continuous survival (cure) data (Beesley et al. 2016). They differ mainly in terms of the implemented conditioning on the (censored) time-to-event. For all imputation models presented, we also condition on the other covariates. Imputation is performed using the R (R Core Team 2019) package `mice` (van Buuren 2007) with single level normal imputation.

Let X_k be one of K incomplete continuous time-invariant random variables ($k = 1, \dots, K$) and let X be $X = (X_1, \dots, X_K)$. Let $X_{-k} = (X_1, \dots, X_{k-1}, X_{k+1}, X_K)$. Let Z_m be one of M complete variables ($m = 1, \dots, M$) and let $Z = (Z_1, \dots, Z_M)$.

In the following, we will mainly discuss how to include the information included in the event indicator, E , as well as the completely observed last-observed-period variable, Y , or the uncensored but incompletely observed time-to-event variable, T .

The first approach (FCS P $Y + E$) uses dummy variables Y_j (j periods, $j = 1, 2, \dots, J$) for each possible period j . Let Y be $Y = (Y_1, \dots, Y_J)$. The aim is to allow for enough flexibility in the relationship between Y and the other covariates.

The imputation model is:

$$X_k = [Y, E, X_{-k}, Z]\beta + e, e \sim N(0, I\sigma^2) \quad (2.3)$$

with persons as observational units.

This specification with dummy variables for each possible time point is flexible, but it is possible only if the number of discrete-time points is not too large, the hazard is not expected to be near zero in some periods, and the risk sets are sufficiently large for each time point. Other specifications of the relationship between (censored) time-to-event are possible in these cases. They include, for example, linear, quadratic, cubic, or higher-order polynomials. It is also possible to use the logarithm of time (Klein et al. 2013) or step functions for grouped periods. For our simulation, however, we do not simulate more than 15 possible time points. Thus, we use dummy variables to keep the specification as general as possible.

Another approach in the person-oriented format (FCS P $\log(T)$) is to treat the time-to-event, T , as a partially observed covariate (Beesley et al. 2016), and thus to impute it in the same way as other partially observed covariates. We impute the partially observed covariates conditional on the other covariates and the partly imputed time-to-event, T , and not on the last period a unit was observed, Y , and the event indicator, E .

The imputation model is thus

$$X_k = [\log(T), E, X_{-k}, Z]\beta + e, e \sim N(0, I\sigma^2) \quad (2.4)$$

with persons as observational units.

FCS specifications with data in person-period format

It is also possible to impute after converting the data to a person-period (PP) format. Note that persons will possibly receive varying imputed values for time-invariant covariates. As in the approaches in the person-oriented (P) format, we always impute conditional on all other covariates. Imputation is again done with `mice` in R.

For our first approach (FCS PP $P + E$) in person-period format, we impute conditional

2. Multiple Imputation of Partially Observed Covariates in Discrete-Time Survival Analysis

on the event indicator, E , and dummy variables P_j for every period j ($j = 1, \dots, J$). Let P be $P = (P_1, \dots, P_J)$.

The imputation model is thus

$$X_k = [P, E, X_{-k}, Z]\beta + e, e \sim N(0, I\sigma^2), \quad (2.5)$$

with periods within persons as single units.

However, we expect that we will lose important information, especially for the rows belonging to the first few periods. If we do not include the last period in which a unit was observed, Y , we do not directly condition on the censored survival time. Therefore, we also try imputing conditional on the last period in which a unit was observed instead of the current period. We use dummy variables Y_j (FCS PP $Y + E$).

The imputation model is thus

$$X_k = [Y, E, X_{-k}, Z]\beta + e, e \sim N(0, I\sigma^2) \quad (2.6)$$

with periods within persons as observational units.

SMC-FCS specifications with data in person-period format

After presenting several possible FCS approximations, we now examine the performance of SMC-FCS in this setting. SMC-FCS has shown excellent results in different simulation settings (Bartlett et al. 2014; Beesley et al. 2016). We explore how this approach fares with a DTSAM as a substantive model. In contrast to the other imputation approaches, we now have to include a substantive model.

When we include the substantive model (see Equation 2.2) using the R package `smcfcs` (Bartlett and Keogh 2019) and impute in person-period format (SMC-FCS PP $P + E$), we effectively condition our imputations on the current period, P , of the row aside from the

other covariates and the event indicator, E . Therefore, the imputation model includes all variables that are also part of the substantive model of interest, and all the data are in the same format. However, the imputation model does not depict the clustered structure of the observation, and it does *not* directly include the last period in which a unit was observed, Y .

We also review another approximation (SMC-FCS PP $Y + E$). We regress the event indicator, E , on the covariates, but we also include the last period a unit was observed, Y . Although we recognize that this is a departure from the original SMC-FCS approach concerning the data format, we think it is worthwhile to test this approximation. It also allows modeling part of the clustering of periods within persons, possibly relevant in the case of unobserved heterogeneity, which we model in three of the four simulations in our simulation study (see next section).

2.4. Simulation study

2.4.1. Data-generating mechanisms

We now examine the performance of the imputation approaches in a series of four simulations. We first present the details (data-generating mechanisms, the introduction of missingness, performance measures) of the simulations and then discuss the results.

For all four simulations, we create five continuous covariates, X_k ($k = 1, \dots, 5$), drawn from a multivariate normal distribution with means of 0 and variances of 1. The covariates are moderately correlated with each other ($r = 0.1$).

Concerning the generation of the survival times, we use the DTSAM model (see Equation 2.2), which will be estimated after the introduction of missingness and the subsequent imputation. This yields (truly) discrete survival data that fulfill the assumption of proportional odds for the DTSAM, the substantive model of interest. If we do not add

a frailty term to the generation of survival times, the generated data also meet the central DTSAM assumption of no unobserved heterogeneity. However, as several authors have noted, the assumption of no unobserved heterogeneity is highly unrealistic (Allison 1982, 83). Therefore, no unobserved heterogeneity is assumed only in one of the four simulations. In the other three simulations, we assume successively larger amounts of unobserved heterogeneity. In all three cases, the frailty term is normally distributed, with a mean of 0 and variance that increases between simulations (0.25, 1, and 4). We use $\alpha = (-5.00, -4.72, -4.44, \dots, -0.8)$ and $\beta = (0.8, 2.2, -0.5, 0.3, -1.4)$ as parameter vectors (see Equation 2.2). We simulate 15 possible time points; after the last time point, all observations are censored. We also randomly censor 10 percent of all observations. For each simulation scenario, we generate 1,000 simulated data sets to prevent the Monte Carlo error from masking differences between methods, and we draw 1,000 samples with 2,000 persons each. Data set length in person-period format will vary.

2.4.2. Missingness

To introduce missing data in each of the simulations, we set 30 percent of the observations for each covariate to missing with probabilities depending on the other covariates and the censored survival time. Therefore, the completeness of an observation row depends both on the outcome and all the covariates, which leads to biased coefficient estimates after listwise deletion (LD).²

After introducing different possible imputation strategies and creating data sets with missing observations, we are now able to examine the performance of the different imputation approaches.

²We avoid perfect predictors and the accompanying computational problems (White et al. 2011, 394), namely, that the complete cases include only failures or non-failures for a specific time point. We avoid these problems by including 30 fail-safe observations (two for each possible time point, one with event indicator one and one with event indicator zero) in all analyzed data sets used in the simulations (full data sets, data sets with missing data, data sets with imputed values).

2.4.3. Methods compared and performance measures

We perform MI of partially observed covariates in discrete-time survival analysis using the imputation specifications described earlier in this article. For each simulation and method, we create five imputed data sets. We then compute the mean coefficient, relative bias, mean squared error (MSE), confidence interval (CI) length, and coverage for estimated DTSAM parameters across 1,000 Monte Carlo repetitions for each imputation model specification.

Performance is often evaluated only for regression point estimates and variance coefficients. However, logistic regression estimates have come under scrutiny because they do not behave like linear regression estimates. Logistic regression estimates are influenced by unobserved heterogeneity – that is, omitted variables (Mood 2009, 67). To allow comparisons between models with different covariate specifications, researchers use average marginal effects (AME) to interpret substantive results. An AME is the average effect of an independent variable on the predicted probability (Mood 2009, 75) – in case of a DTSAM, the hazard. Due to the popularity of AMEs, we also provide a comparison of the AME means for all approaches.

2.4.4. Simulation results

Figure 2.1 displays the bias for the estimates of β_1 for no unobserved heterogeneity and increasing unobserved heterogeneity with frailty term variance $\sigma^2 = 0.25$, $\sigma^2 = 1$ and $\sigma^2 = 4$, respectively.

We concentrate first on the imputation methods under no unobserved heterogeneity on the left. The assumptions of no unobserved heterogeneity and proportional odds are thus fulfilled for the analysis model with full data. As listwise deletion (LD) overestimates the true coefficient, and the CI length and MSE are about double that of the full data, there is room for improvement in terms of bias and efficiency. Turning to the different

imputation methods and their performance, we notice profound differences. We register the highest bias for the imputation approach FCS P $\log(T)$, where we first imputed the missing time-to-event, T , and used this variable (apart from the covariates) to impute our covariates. As in the case of the Cox model, this approach is completely inadequate (Beesley et al. 2016, 4711). The second and third-to-worst approaches are the two FCS approaches in person-period format (5. FCS PP $P + E$ and 6. FCS PP $Y + E$) with the current period and the censored survival time, respectively, included in the imputation models.

The FCS approach in person-oriented format using the information from the (censored) survival time and the event indicator (3. FCS P $Y + E$) is a little better but still not satisfying. This leaves us with the two SMC-FCS approaches (7. SMC-FCS PP $Y + E$ and 8. FCS PP $P + E$). The compatible model (with the period information) performs better than the incompatible one with the censored time-to-event (T).

The same is true if we add unobserved heterogeneity to the data-generating process (frailty term variance $\sigma^2 = 0.25$ or $\sigma^2 = 1$, $\sigma^2 = 4$), even though the compatible model ignores the clustered structure of the periods within persons. As the assumption of no unobserved heterogeneity is thus not fulfilled – a very common violation – we explore the performance of imputation approaches under this condition. With increasing heterogeneity, the logistic regression coefficients are drawn toward zero by unobserved heterogeneity, that is, omitted variables (Mood 2009). Nevertheless, the ranks of the different imputation approaches in relative bias stay the same. The SMC-FCS approaches in person-period format fare very well, whereas the other approaches do not perform sufficiently in terms of relative bias. We can confirm the strong performance of the SMC-FCS approaches (7 and 8) concerning MSE and coverage if we look at other performance measures. Whereas some of the performance measures for FCS approaches perform as well as the compatible

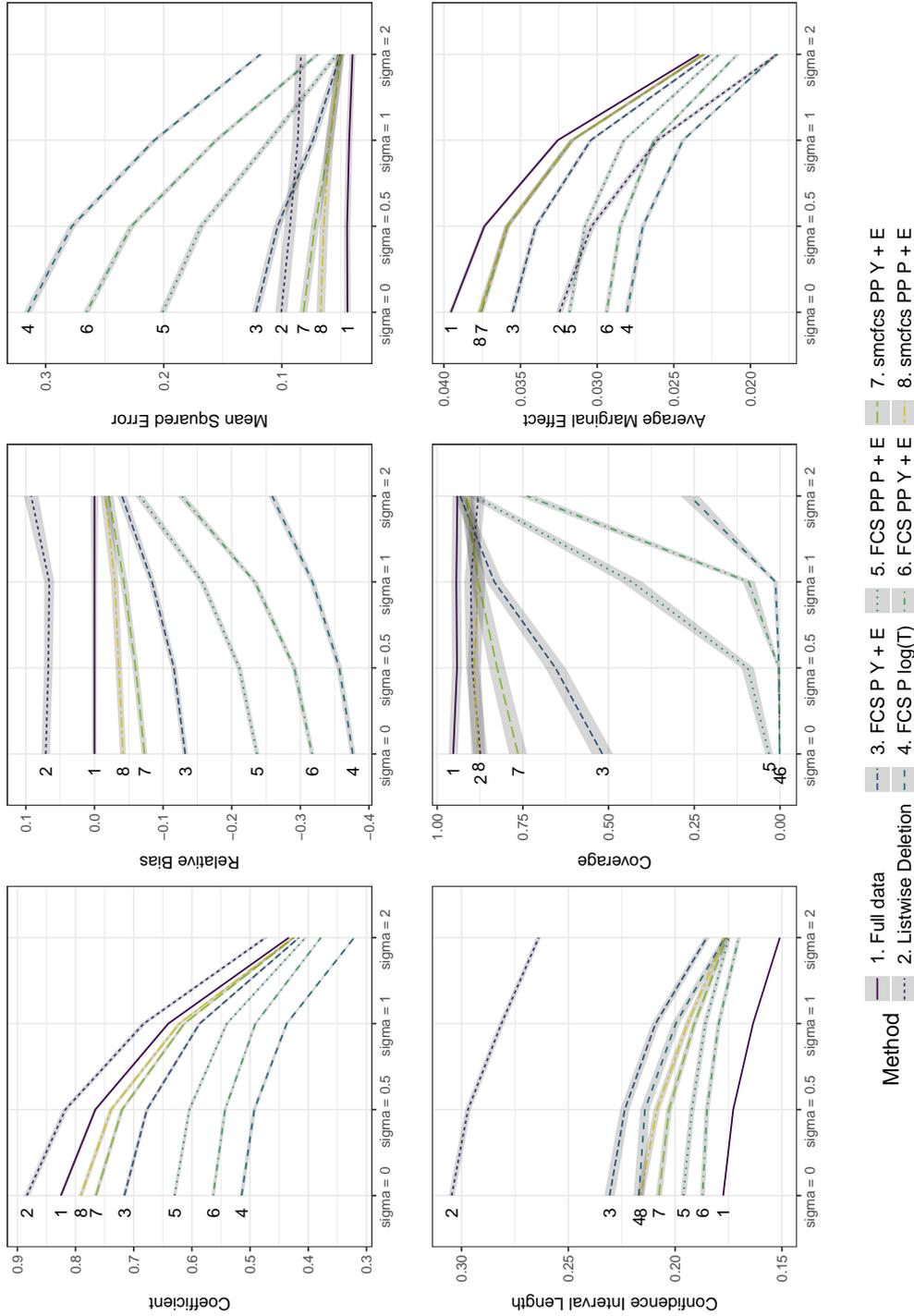


Figure (2.1) Performance measures for coefficient β_1 (for X_1). Het. = unobserved heterogeneity; PP = person-period format; P = person-oriented format; FCS = fully conditional specification; SMC-FCS = substantive-model compatible fully conditional specification. Grey areas 95% confidence interval for performance measures. 1,000 Monte Carlo (MC) repetitions per simulation. Results are also available in table format in the Appendix Section A.

SMC-FCS approach in the case of high heterogeneity, they do not perform better than the SMC-FCS approaches.

As AMEs are often used in the comparison of different DTSAM models (e.g., Wagner et al. 2019, 84), we also include them in our performance evaluation. The first thing one notices is that there is no longer any recognizable difference between the two SMC-FCS approaches (7 and 8). We can conclude that the difference in coefficient performance is due to fact that the SMC-FCS imputation that included the last period in which a unit was observed (7), introduced too much unobserved heterogeneity into the imputed data set. This, in turn, leads to coefficient estimates that are biased toward zero but leaves the AMEs approximately unbiased.

From all that we have seen so far, the compatible SMC-FCS model outperforms or performs at least as well as all other imputation models and is therefore recommended. The currently most common approach, listwise deletion (LD), leads to high efficiency losses and possibly high bias, and should therefore be avoided. The amount of heterogeneity that can be modeled in a DTSAM will vary strongly with available variables, but the compatible SMC-FCS performs well under all examined conditions.

Apart from genuinely discrete survival data, DTSAMs with a logistic link are often used to analyze interval-censored survival data. Therefore, we base another scenario on the continuous Weibull distribution for survival times and rounded variables in order to create interval-censored survival times (Lee and Go 1997) Results are presented in table form in Table A.1 in the Appendix.³ The most important take-away from this additional simulation is that the conjugate SMC-FCS imputation model (8. SMC-FCS PP $P + E$) again shows excellent properties in terms of (relative) bias, MSE, CI length, and coverage. The FCS approaches in person-period format, however, do not perform well (high bias and coverage near zero).

³The created data fulfill the criteria of proportional hazards. In the case of a DTSAM, our assumption is, however, that the odds are proportional (it is also called the proportional odds model). If the hazard probabilities are small (<0.2), however, this represents no problem since $h_{ij} \approx h_{ij}/(1 - h_{ij})$.

2.5. An applied example with the German Family Panel pairfam

To provide an applied example under real-world data conditions, we now conduct an example analysis from the field of relationship and family research using data from the German Family Panel project called pairfam (Brüderl et al. 2017).

The 2008-launched pairfam panel (“Panel Analysis of Intimate Relationships and Family Dynamics”) is a longitudinal study for research on relationships and families in Germany. The data is collected from a nationwide random sample of the three birth cohorts 1971-73, 1981-83, 1991-93, and their partners, parents, and children. For our real-world example, we used data from the data set *biopart*, which includes prospective and retrospective information on the anchor’s relationships, including relationships, cohabitation, and marriage history. We used the data set 8.0.0 (Brüderl et al. 2017), which includes updated information from the survey waves 1-9 (for more details on the panel see Huinink et al. 2011).

We use a simple substantive model from the field of relationship stability research. Note that our goal here is not to estimate any causal models but rather to explore how our imputation approaches fare with a real data set. Let us assume that we are interested in the relationship between the probability that a couple i splits up and several time-invariant independent variables.

We first include six indicators $j, j = \{1, 2, \dots, 6\}$, for grouped periods, each representing five years of a relationship and the last one all years after 25 years of a relationship. These variables include the time point at which the relationship started (*begin* in years since 1900), the age in years when the anchor person began the relationship (*age*), the difference in ages of the two relationship partners (*difference age*), an indicator for whether the partners are married (*married*) and an indicator for whether the parents

separated during the anchor’s childhood (*parents’ separation*).

$$\begin{aligned} \text{separated} \sim & \\ & \sum_{j=1}^6 \alpha_j P_j + \beta_1 \text{begin} + \beta_2 \text{age} + \beta_3 \text{difference age} \\ & + \beta_4 \text{married} + \beta_5 \text{parents' separation} \end{aligned} \quad (2.7)$$

We reduce the data set to the fully observed first-reported relationships of all anchors. This leaves us with a sample of 2,173 relationships. Transforming the data set to the relationship-year format leads to 26,554 rows (i.e., observed periods). For our data set with missing data, we deleted 30 percent of the observations for three of the variables, namely, *age*, *begin*, and *parents’ separation*. The values are missing at random (MAR) – that is, missingness depends on the censored time to survival, whether the relationship failed, whether the partners are married to each other, and the two respective other variables with missing data. We then impute the missing values using the same approaches we already tested in the simulations.

Examining the results in Figure 2.2, we notice that after listwise deletion, the coefficient estimate for *begin* has a higher variance. We also see that after imputing with our various approaches, the imputation approach FCS P $\log(T)$ (4) again performs differently and worse than the other imputation approaches; the coefficient of *begin* is way off; the 95% confidence interval does not even come close to the interval of the full data. The FCS approach in person-period format with period dummies (5. FCS PP $P + E$) also seems to perform especially poorly: confidence intervals are very wide. The SMC-FCS approach with a compatible imputation model (8. SMC-FCS PP $P + E$) performs adequately in this real-world setting, as was also the case in the simulation.

2. Multiple Imputation of Partially Observed Covariates in Discrete-Time Survival Analysis

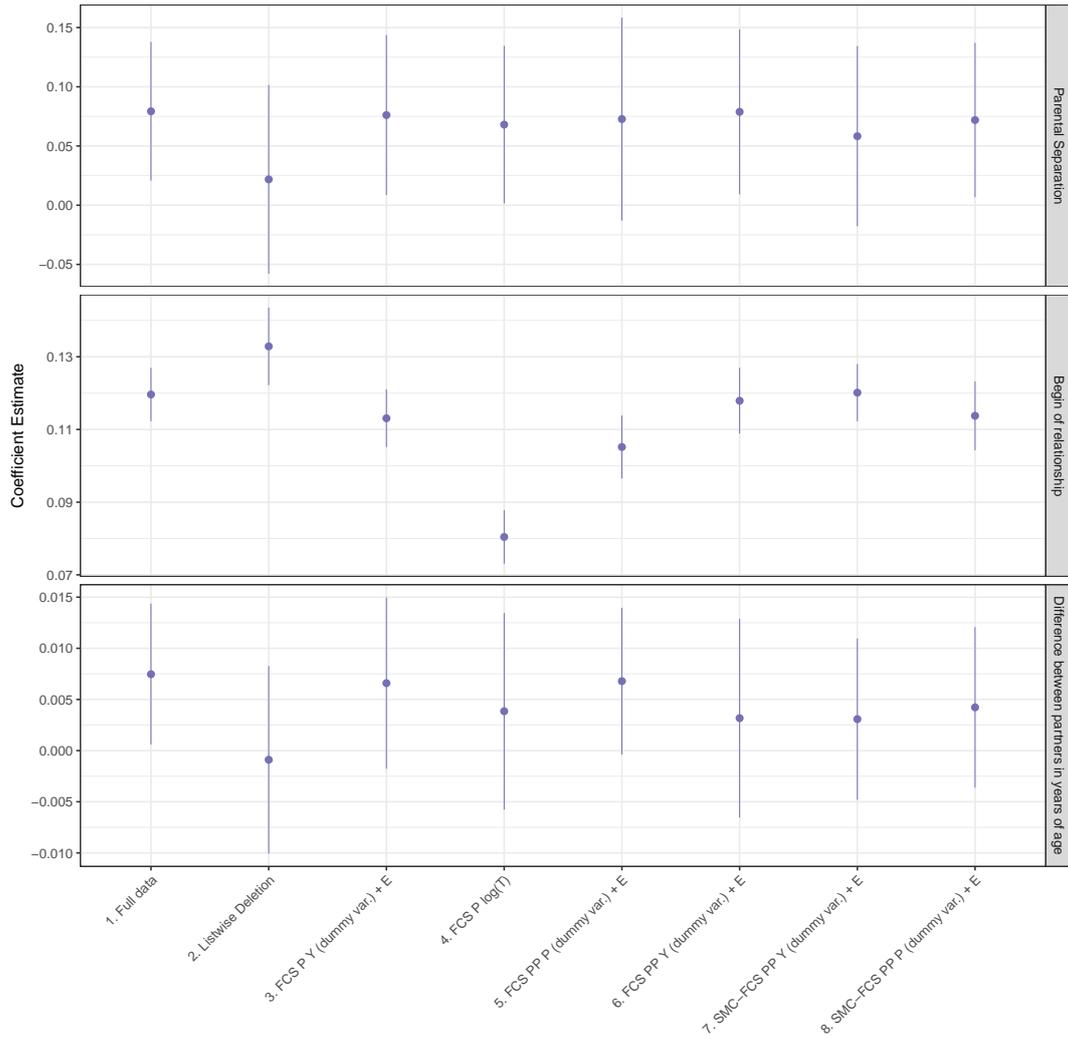


Figure (2.2) Selected estimated coefficients for a DTSAM of the separation of the first fully observed reported relationships. Pairfam data set 8.0.0 with missing data introduced by the authors. For the specification of the DTSAM see Equation 2.7.

In sum, the findings resemble those found in the simulations. They show (1) that some imputation approaches – for example, treating the time-to-event variable as partly unobserved – are not advisable; (2) that imputing in person-period format with FCS cannot be recommended; and (3) that SMC-FCS with a compatible imputation model performs adequately.

2.6. Discussion

Results

Like many other types of data analysis, the analysis of discrete-time survival data is often challenged by missing data in one or more covariates. Negative consequences of such missing data include efficiency losses and bias. A popular approach to circumventing these consequences is MI. However, in MI, it is crucial to include outcome information in the model for imputing partially observed covariates. Unfortunately, this is not straightforward in the case of discrete-time survival data because (1) we usually have a partially observed (left- or right-censored or both) outcome; (2) we do not have just one outcome variable, but two (i.e., event and time-to-event); and (3) we have to decide whether to impute while the data set is still in the person-oriented format or after conversion to person-period format, especially if we are looking at time-invariant variables.

In this article, we have tested different approaches for imputing missing covariates in the DTSAM setting. For this purpose, we performed four simulations that differed in the amount of unobserved heterogeneity. Some of the investigated methods are from the literature on the imputation of time-constant variables for the Cox model (van Buuren et al. 1999; White and Royston 2009; Clark and Altman 2003; Keogh and Morris 2018; Beesley et al. 2016). We also took a closer look at SMC-FCS (Bartlett et al. 2014) and its performance. We also provided an applied example using pairfam data (Brüderl et al. 2017).

Our findings lead us to agree with Beesley et al. (2016) that treating censored survival times as partially unobserved and imputing other covariates depending on the multiply imputed missing survival times yields unsatisfying results in all cases. Whereas Beesley et al. (2016) observed this for the Cox (cure) model, we have confirmed it for discrete-time survival analyses.

Furthermore, the performance of imputation methods in person-period format with FCS is disappointing. Apart from the inherent incoherence between the imputed values for different times in the same person, coverage and relative bias are unsatisfying. The SMC-FCS approach using a compatible imputation model performs best in our simulations with and without unobserved heterogeneity and is therefore strongly recommended.

Limitations and future research

Note that the approach that fared best in our simulation study (i.e., the approach using SMC-FCS with a compatible imputation model) is also suitable in principle for the imputation of time-varying covariates. This is because imputation is done in person-period format, and it is, therefore, possible to impute different values for different observed periods of the same person. However, research to confirm adequate performance in the case of time-varying covariates has yet to be undertaken.

Another open question is how to deal with time-varying effects of time-invariant and time-varying covariates. Keogh and Morris (2018) have shown that a variation of the SMC-FCS approach also performs sufficiently well in the case of time-varying covariates. Again, an evaluation for discrete-time survival models has not yet been conducted.

Conclusions

We recommend using the SMC-FCS approach proposed by Bartlett et al. (2014) with a compatible imputation model. This approach performed at least as well as – and usually better than – the other (SMC-)FCS approaches we explored in this article. This is true in the case in which the data set fulfilled the assumptions of a DTSAM. It is also true in the case of varying degrees of unobserved heterogeneity of and interval-rounded survival data with proportional hazards instead of proportional odds. The implementation of SMC-FCS in Stata (Bartlett 2015) and in R (Bartlett and Keogh 2019) and the specification of the substantive (analytical) model are easy to do. The SMC-FCS approach has already

2. Multiple Imputation of Partially Observed Covariates in Discrete-Time Survival Analysis

shown excellent performance in the case of included quadratic covariates and interaction effects in linear and logistic regression as well as for the imputation of covariates in Cox (cure) regressions (Bartlett et al. 2014; Bartlett and Taylor 2016; Beesley et al. 2016; Keogh and Morris 2018). In conclusion, we also recommend this approach for the imputation of time-invariant covariates in discrete-time survival analysis.

Bibliography

- Allison, P. (1982). Discrete-Time Methods for the Analysis of Event Histories. *Sociological Methodology*, 13:61–98.
- Allison, P. (2000). Multiple Imputation for Missing Data: A Cautionary Tale. *Sociological Methods & Research*, 28(3):301–309.
- Allison, P. D. (2010). Event History and Survival Analysis. In Hancock, G. R., Stapleton, L. M., and Mueller, R. O., editors, *The Reviewer's Guide to Quantitative Methods in the Social Sciences*, chapter 7, pages 413–424. Routledge, New York.
- Arranz Becker, O. and Lois, D. (2010). Unterschiede im Heiratsverhalten westdeutscher, ostdeutscher und mobiler Frauen: Zur Bedeutung von Transformationsfolgen und soziokulturellen Orientierungen. *Soziale Welt*, 61(1):5–26.
- Arranz Becker, O. and Lois, D. (2013). Competing Pleasures? The Impact of Leisure Time Use on the Transition to Parenthood. *Journal of Family Issues*, 34(5):661–688.
- Barber, J. (2001). Ideational Influences on the Transition to Parenthood: Attitudes toward Childbearing and Competing Alternatives. *Social Psychology Quarterly*, 64:101–127.
- Bartlett, J. (2015). Multiple Imputation of Covariates by Substantive-Model Compatible Fully Conditional Specification. *Stata Journal*, 15(2):437–456.

Bibliography

- Bartlett, J. and Keogh, R. (2019). *smcfcs: Multiple Imputation of Covariates by Substantive Model Compatible Fully Conditional Specification*. R Package Version 1.4.0.
- Bartlett, J. W., Seaman, S. R., White, I. R., and Carpenter, J. R. (2014). Multiple Imputation of Covariates by Fully Conditional Specification: Accommodating the Substantive Model. *Statistical Methods in Medical Research*, 24(4):462–487.
- Bartlett, J. W. and Taylor, J. M. G. (2016). Missing Covariates in Competing Risks Analysis. *Biostatistics*, 17(4):751–763.
- Beesley, L., Bartlett, J., Wolf, G., and Taylor, J. (2016). Multiple Imputation of Missing Covariates for the Cox Proportional Hazards Cure Model. *Statistics in Medicine*, 35(26):4701–4717.
- Böttcher, K. (2006). Scheidung in Ost- und Westdeutschland. *KZfSS Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 58(4):592–616.
- Brüderl, J., Hank, K., Huinink, J., Nauck, B., Neyer, F. J., Walper, S., Alt, P., Borschel, E., Buhr, P., Castiglioni, L., Friedrich, S., Finn, C., Garrett, M., Hajek, K., Herzig, M., Huyer-May, B., Lenke, R., Müller, B., Peter, T., Schmiedeberg, C., Schütze, P., Schumann, N., Thönnissen, C., Wetzels, M., and Wilhelm, B. (2017). The German Family Panel (pairfam). Technical Report ZA5678 Data file Version 8.0.0, GESIS Data Archive, Cologne.
- Carpenter, J. and Kenward, M. (2013). *Multiple Imputation and its Application*. Wiley, Hoboken.
- Clark, T. and Altman, D. (2003). Developing a Prognostic Model in the Presence of Missing Data: An Ovarian Cancer Case Study. *Journal of Clinical Epidemiology*, 56(1):28 – 37.

Bibliography

- Cooke, L. P. (2006). Doing Gender in Context: Household Bargaining and Risk of Divorce in Germany and the United States. *American Journal of Sociology*, 112(2):442–472.
- Cooper, M., Loukas, A., Case, K. R., Marti, C. N., and Perry, C. L. (2018). A Longitudinal Study of Risk Perceptions and E-cigarette Initiation Among College Students: Interactions With Smoking Status. *Drug and Alcohol Dependence*, 186:257 – 263.
- Cox, D. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220.
- Fay, R. E. (1992). When are Inferences From Multiple Imputation Valid? In *Proceedings of the Survey Research Methods Section of the American Statistical Association*, volume 81.
- Hughes, R., Heron, J., Sterne, J., and Tilling, K. (2019). Accounting for Missing Data in Statistical Analyses: Multiple Imputation Is Not Always the Answer. *International Journal of Epidemiology*, 48(4):1294–1304.
- Huinink, J., Brüderl, J., Nauck, B., Walper, S., Castiglioni, L., and Feldhaus, M. (2011). Panel Analysis of Intimate Relationships and Family Dynamics (pairfam): Conceptual Framework and Design. *Zeitschrift für Familienforschung : ZfF*, 23(1):77–101.
- Kenward, M. G. and Carpenter, J. (2007). Multiple Imputation: Current Perspectives. *Statistical Methods in Medical Research*, 16:199–218.
- Keogh, R. and Morris, T. (2018). Multiple Imputation in Cox Regression when there are Time-Varying Effects of Covariates. *Statistics in Medicine*, 37(25):3661–3678.
- Klein, T., Kopp, J., and Rapp, I. (2013). Metaanalyse mit Originaldaten. Ein Vorschlag zur Forschungssynthese in der Soziologie. *Zeitschrift für Soziologie*, 42(3):222–238.
- Kleinbaum, D. G. and Klein, M. (2012). *Survival Analysis. A Self-Learning Text*. Springer New York, 3 edition.

Bibliography

- Lee, E. and Go, O. (1997). Survival Analysis in Public Health Research. *Annual Review of Public Health*, 18(1):105–134.
- Little, R. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. Wiley.
- Manning, W. D., Brown, S. L., and Stykes, J. B. (2016). Same-Sex and Different-Sex Cohabiting Couple Relationship Stability. *Demography*, 53(4):937–953.
- Meng, X.-L. (1994). Multiple-Imputation Inferences with Uncongenial Sources of Input. *Statistical Science*, 9(4):538–558.
- Mood, C. (2009). Logistic Regression: Why We Cannot Do What We Think We Can Do, and What We Can Do About It. *European Sociological Review*, 26(1):67–82.
- Murad, H., Dankner, R., Berlin, A., Olmer, L., and Freedman, L. S. (2019). Imputing Missing Time-Dependent Covariate Values for the Discrete Time Cox Model. *Statistical Methods in Medical Research*.
- Nomaguchi, K. M. (2006). Time of One’s Own. Employment, Leisure, and Delayed Transition to Motherhood in Japan. *Journal of Family Issues*, 27:1668–1700.
- Nur, U., Shack, L., Rachet, B., Carpenter, J., and Coleman, M. (2009). Modelling Relative Survival in the Presence of Incomplete Data: A Tutorial. *International Journal of Epidemiology*, 39(1):118–128.
- R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, Hoboken.
- Rubin, D. B. (1996). Multiple Imputation after 18+ Years. *Journal of the American Statistical Association*, 91(434):473–489.

Bibliography

- Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman and Hall.
- Schoen, R., Astone, N. M., Rothert, K., Standish, N. J., and Kim, Y. J. (2002). Women's Employment, Marital Happiness, and Divorce. *Social Forces*, 81(2):643–662.
- Singer, J. D. and Willett, J. B. (1993). It's About Time: Using Discrete-Time Survival Analysis to Study Duration and the Timing of Events. *Journal of Educational Statistics*, 18(2):155–195.
- Singer, J. D. and Willett, J. B. (2003). *Applied Longitudinal Data Analysis*. Oxford University Press.
- Stoddard, S. A. and Veliz, P. (2019). Summer School, School Disengagement, and Substance Use During Adolescence. *American Journal of Preventive Medicine*, 57(1):11 – 15.
- van Buuren, S. (2007). Multiple Imputation of Discrete and Continuous Data by Fully Conditional Specification. *Statistical Methods in Medical Research*, 16(3):219–242.
- van Buuren, S., Boshuizen, H. C., and Knook, D. L. (1999). Multiple Imputation of Missing Blood Pressure Covariates in Survival Analysis. *Statistics in Medicine*, 18(6):681–694.
- Wagner, M., Mulder, C. H., Weiss, B., and Krapf, S. (2019). The Transition From Living Apart Together to a Coresidential Partnership. *Advances in Life Course Research*, 39:77 – 86.
- White, I. R. and Royston, P. (2009). Imputing Missing Covariate Values for the Cox Model. *Statistics in Medicine*, 28(15):1982–1998.
- White, I. R., Royston, P., and Wood, A. M. (2011). Multiple Imputation Using Chained Equations: Issues and Guidance for Practice. *Statistics in Medicine*, 30(4):377–399.

Bibliography

Xue, X. and Brookmeyer, R. (1996). Bivariate Frailty Model for the Analysis of Multivariate Survival Time. *Lifetime Data Analysis*, 2(3):277–289.

3. Systematically Missing Partner Variables and Multiple Imputation Strategies: A Case Study With German Relationship Data

3.1. Introduction

Multi-actor studies collect information on persons who maintain a significant relationship or connection with each other. For the analyses of couples or marriages, the anchor respondent's partner can be included as a so-called secondary respondent in a multi-actor survey (Kalmijn and Liefbroer 2010; Dykstra et al. 2012; Kalmijn et al. 2018). However, due to monetary and time restrictions in data collection, specific variables are commonly recorded only for anchor respondents and not for their partners. Questionnaires for secondary respondents may not include all items collected for anchors (Dykstra et al. 2012; Brüderl et al. 2017). Therefore, researchers who need to use the same variables from anchors and partners for their analyses often face the problem of *systematically missing data* (Resche-Rigon et al. 2013). In our case, the systematically missing data are data that are missing for all secondary respondents.

3. Systematically Missing Partner Variables

The problem of systematically missing data is not confined to the analysis of data on primary and secondary respondents like anchors and their partners. Systematically missing data can arise in various settings, for example, if there is a change in measurement in repeated surveys (e.g., in official statistics like censuses). The old measurement is missing after the change, and the new measurement is missing before the change (Parker et al. 2004; Schenker and Raghunathan 2007). Systematically missing data can also be present for analyses of pooled and harmonized surveys. In ex-post survey harmonization, a variable can be considered systematically missing if the study in question does not include the specific variable, but the other studies do. A popular solution for both cases is using multiple imputation (MI) techniques to impute the missing variables (Schenker and Parker 2003; Resche-Rigon and White 2016).

We will look into whether we can transfer the strategies employed in these related fields to the case of missing partner variables. This paper explores *if and how multiple imputation can be used when information on partner respondents is systematically missing for a partner variable*. Since we cannot estimate the correlation between the systematically missing partner variable and the corresponding anchor variable with the original data set, we have to fall back on assumptions. We can draw two possible assumptions from the literature in related fields. The first one is to assume that the partial correlation between the systematically missing partner variable and its corresponding anchor variable given other available variables is zero. This assumption is also called the assumption of conditional independence. Another possibility is to estimate the correlation in another study, i.e., a so-called bridging study (Parker et al. 2004). This approach requires the assumption that these estimates from the bridging study are transferable to the original study with the systematically missing partner variable. We will evaluate both strategies to give insight into which assumption is more suitable for the imputation of systematically missing partner variables. We want to note that due to the fact that these approaches

3. Systematically Missing Partner Variables

are based on strong assumptions, they may not be suitable for main analyses but rather for supplementary and sensitivity analyses. Sensitivity analyses allow us to study how various sources of uncertainty contribute to the model's overall uncertainty.

The paper is structured as follows: First, we give a brief overview of multiple imputation (MI) principles for missing data. Second, we will review previous research on systematically missing data, focusing on approaches that aim to resolve this problem through *multiple imputation*. This encompasses the following areas: (*ex-post*) *survey harmonization* of multiple surveys with a variable missing in one or multiple surveys completely, changes in measurements for extended time series in *official statistics*, and *split questionnaire designs*. We briefly introduce each area and summarize the critical differences in the pattern of systematically missing data. Third, building on this literature, we outline MI approaches for systematically missing data for partner respondents. We illustrate the approaches through a simulation with data from the German Socio-Economic Panel (Goebel et al. 2019), pairfam – The German Family Panel (Brüderl et al. 2017), and the German sub-study of SHARE, the Survey of Health, Ageing and Retirement in Europe (Börsch-Supan et al. 2013; Börsch-Supan 2020). We use the two latter ones in our simulation as bridging studies. We chose these two as bridging studies for the broad overlap in observed variables, and chose panel waves that were close in time to the selected GSOEP wave. Still, we notice substantial study heterogeneity between the studies, especially regarding the included age cohorts. Thus, we also take a broader look at study heterogeneity between the three surveys to get a better preliminary evaluation of this problem. We close with a discussion of limitations and the next steps.

3.2. Literature overview: Systematically missing data in related fields

We will start with an overview of different cases of systematically missing data. A visual representation is given in Figure 3.1, a summary in tabular form in Table 3.1.

The term *systematically missing data* was coined for meta-analysis with raw data, also called individual person data (IPD) meta-analysis (Resche-Rigon et al. 2013). In an IPD meta-analysis, raw data from several studies are combined and then jointly analyzed. In the social sciences, research projects combining, pooling, and harmonizing data are usually not called IPD meta-analyses but instead described as ex-post survey harmonization projects (Granda et al. 2010). Survey harmonization projects have become rather popular in recent years; some current projects include CLOSER (2020), which harmonizes data from longitudinal studies in the UK, HaSpaD (2020) combining German relationship data from different surveys, or the Survey Data Recycling (SDR 2020) project focusing on cross-national research on social capital and political participation. If an analysis were to be conducted only with one data set with a systematically missing variable, it would not be possible to include the systematically missing variable. For example, in Figure 3.1 a, variable 1 is not observed for studies 5 and 6 but it is in all other studies. They can serve as “bridges” to impute the data in Study 5 and 6. It is possible to impute systematically missing variables through multi-level models while at the same time modeling the heterogeneity between studies (Resche-Rigon et al. 2013; Jolani et al. 2015; Jolani 2017).

Systematically missing data can also occur in official statistics if measurements change over time (Parker et al. 2004; Schenker and Raghunathan 2007), i.e., between waves of a panel or for a time series (see Figure 3.1 b). The change in measurements then hinders comparisons over time. However, both measurements can be included in a survey for a

3. Systematically Missing Partner Variables

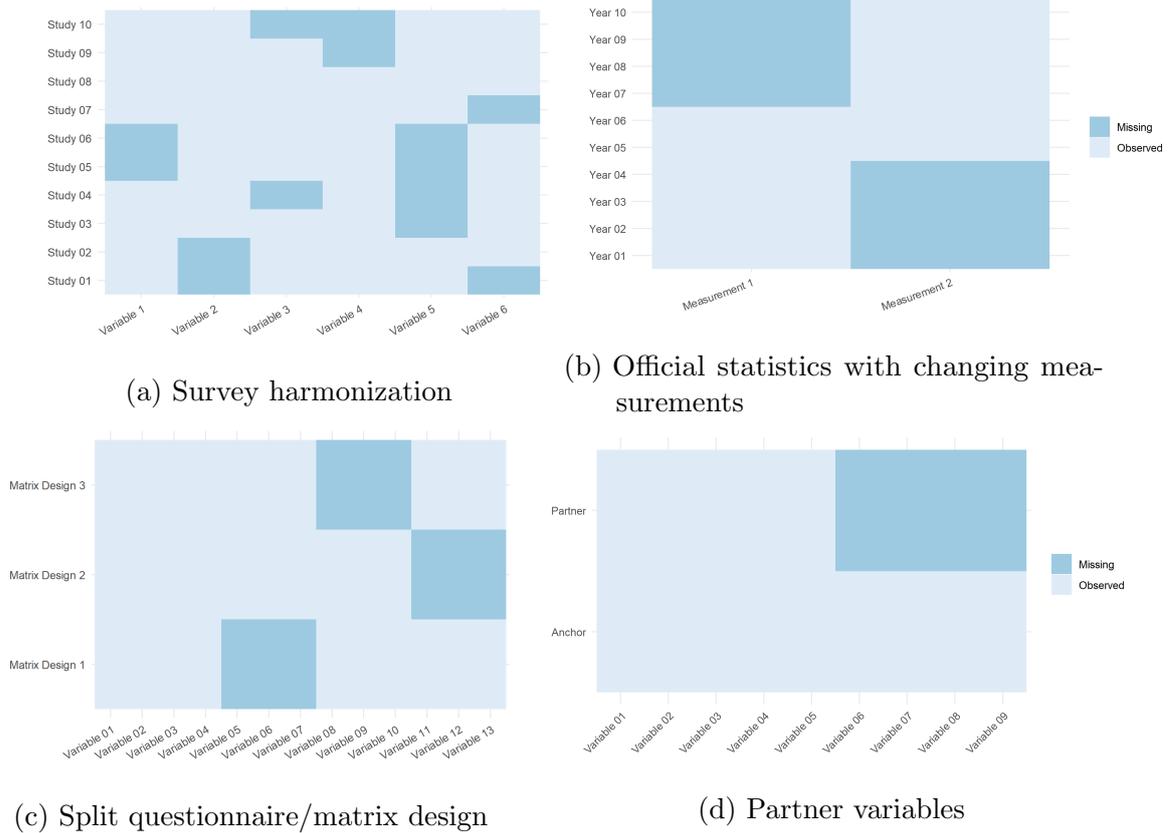


Figure (3.1) Systematically missing data in case of:

limited period. This time period then serves as *bridging study*. The relationship between both measurements can then be estimated and used for multiple imputation, allowing the creation of a complete timeline for either of the measures.

A similar situation arises in split questionnaires, also called matrix design (see Figure 3.1 c), in which the questionnaire is split into several modules. Respondents are only asked varying subsets of modules (see Figure 3.1 c). After data collection, missing components can be imputed. Each split serves as a “bridging study” for other splits in the same questionnaire. Split questionnaire design is used to reduce respondents’ burden and survey costs (Rässler 2003). Bridging studies and split questionnaire designs have in common that the researchers can typically control which variables are observed together (see Raghunathan 2006 and Adigüzel and Wedel 2008 for more information regarding the implementation of optimal matrix designs).

3. Systematically Missing Partner Variables

Table (3.1) Comparison of different problems related to systematically missing data

	Miss- ingness control	Number of studies/subgroups	Heterogeneity between studies	Overlap between systematically missing variables
Survey harmonization	No	Usually high (many studies)	Potentially high	Depends on included studies
Changes in measurement official statis- tics	Yes	Low (few changes in measurement usually)	Low, but may increase in time	Yes, if bridging study available
Split questionnaire design	Yes	Low to high	Very low	Yes, design is controlled by researcher
Anchor and partner variables	No	Low (Two subgroups)	Low, potentially high in regard to bridging study	No, only through external study

The comparison of the different areas, in which systematically missing data appear, reveals differences and similarities (see Figure 3.1 and Table 3.1). For example, the respondents' grouping varies by year of survey in the case of bridging studies, by study in the case of survey harmonization, and by matrix subset in split questionnaires. Another important difference is the level of similarity between the study or the studies with systematically missing data and the bridging study. Ideally, the external bridging study should be as similar as possible to the original study with respect to the underlying population, sampling, interview mode, and measurements. In the context of missing partner variables, the bridging study and its data structure are used to extrapolate the relationship between the anchor and the partner variable, similar to systematically missing data for different official statistics measurements. In official statistics and split questionnaires, changes in study designs present optimal imputation situations compared with survey harmonization since there is control over measurements and sampling. And yet, Parker et al. (2004) and

3. Systematically Missing Partner Variables

Schenker and Parker (2003) warn against blindly extrapolating relationships found in a bridging study to official statistics studies years down the road since they can change over time.

Comparing systematically missing information from partner respondents with other cases of systematically missing data, we first notice that the group for which the data is missing is that of the partners (see Figure 3.1 d). A subset in which the variables are observed for both anchor and partners does not exist. Therefore, we need to make either additional assumptions concerning the (conditional) dependence between observed and unobserved variables or rely on an external data set, which can serve as a bridging study for multiple imputation.

3.3. Motivational example: Life satisfaction of partners in Germany

We will build our simulation on a simple example: the dependency of anchors' and partners' life satisfaction. Similarities between partners can exist for many reasons, like shared experiences, the influence of partners over one another, or homophily (Kenny and la Voie 1985; Ledermann and Kenny 2012). Homophily (also called assortative mating) is people's tendency to seek out or be attracted to those who are similar to themselves (Byrne 1971). Possible reasons for dissimilarities between partners include compensation effects between partners, social competition, or task sharing. (Dis-) similarities between partners have been studied too, such as heterogeneous couples with regard to social-economic status (Edwards 1969; Kalmijn 1998; Skopek et al. 2011). For such analyses, partner variables are often necessary as central predictors or critical confounding variables. Regarding our simulation on life satisfaction, previous research has shown that persons' life satisfaction in a relationship is also affected by their partner's characteristics (Dyrenforth

3. Systematically Missing Partner Variables

et al. 2010; Gustavson et al. 2016). A specific example of such an influence would be the regression coefficient of partner's life satisfaction with anchor's life satisfaction as an outcome variable controlled by other variables such as socio-demographic characteristics and the Big Five personality traits of anchors and partners. While these covariates are available for anchors and partners, partner's life satisfaction is systematically missing in our example. Therefore, we test and evaluate different imputation strategies in the case of systematically missing partner variables on this motivational research topic. We will work with data from the German Socio-Economic Panel (Goebel et al. 2019), and, as additional bridging studies, the German pairfam panel (Brüderl et al. 2017) as well as the German sub-study of the Survey of Health, Ageing and Retirement in Europe (Börsch-Supan et al. 2013; Börsch-Supan 2020). For more details on the data sets see Section 3.5.1.

In the following, we will give an overview of multiple imputation and our strategies.

3.4. Imputation strategies for systematically missing partner variables

3.4.1. Multiple imputation

One of the most popular approaches to tackling missing data is multiple imputation (Little and Rubin 2002). Multiple imputation allows for the analysis of incomplete data sets through substituting missing values. We replace them by “imputing” values of a variable conditioned on other variables; in general, conditioned at least on the variables of the substantial analysis model (Carpenter and Kenward 2013, 72). This procedure is repeated several times, creating several data sets. Each data set is analyzed separately, and estimates are combined across imputations using rules developed by Rubin (1987). Using a single imputation approach instead of multiple imputations would lead to biased

3. Systematically Missing Partner Variables

variance estimates.

To impute missing values, we need models specifying the distribution of the missing values. There are two main approaches, the *joint modeling* (Schafer 1997) approach and the *fully conditional specification* (van Buuren et al. 2006). Joint modeling (JM) means drawing missing values simultaneously for all incomplete variables using a multivariate distribution. Fully conditional specification (FCS) is known under a multitude of names, e.g., as multiple imputation by chained equations, short *mice* or *ice*. In contrast to JM, FCS “splits” the problem into a series of univariate problems (van Buuren 2007). FCS involves specifying each partially observed variable’s conditional distribution through a series of univariate models, given all the other variables (White et al. 2011). Imputation under FCS is then done by iterating over the conditionally specified imputation models. It is more flexible than the JM approach since we can select adequate regression models for every variable (e.g., linear regression for continuous partially observed variables, logistic regression for binary partially observed variables). An important weakness of this approach is that the specified conditional densities can be incompatible. Thus, we possibly do not know the joint distribution to which the imputation algorithm converges. However, in practice, the approach’s actual performance has very often been good (van Buuren 2007).

Due to its popularity and flexibility, we will use FCS to test the different approaches. Since the aim of the article is to explore whether the use of *bridging studies is a possible remedy for systematically missing partner variables in general*, we will not compare different imputation packages. Instead, we will restrict ourselves to one of the most popular ones (*mice* in R). However, the approaches tested can also be carried out with other FCS imputation packages or transferred to joint modeling approaches. We move on to the different imputation approaches for systematically missing partner variables.

3.4.2. Multiple imputation for systematically missing partner variables – Assumptions and bridging studies

In this article, we evaluate different multiple imputation approaches, which we will now present. We first include the full SOEP data set (marked as *01. Full data set* in the results and figures) before deleting the partner variable for comparison and reference. For a detailed description of the data set, see the following Section 3.5.1.

For the imputation approaches, we will start with the least complex ones and move our way up to the possible inclusion of bridging studies in the imputation procedures. A simple way to impute the missing data is to impute in long format, i.e., with the information on anchors and partners in separate rows (Figure 3.2 a). However, this imputation approach (*02. MI in long format*) ignores possible dependencies between anchors and partners in a relationship. This will result in correlation and regression point estimates for the anchor's life satisfaction and the life satisfaction of the partner that are strongly biased towards zero (see Section 3.6).

Imputing in *wide format* with one row for each relationship, i.e., the format required for the target analysis (Figure 3.2 b), is not possible with standard FCS procedures. Since we do not have any observations in the systematically missing partner variable, the univariate imputation model cannot be estimated.

A possibility would be to first transform the data into the so-called pairwise format (Figure 3.2 c). The pairwise format is well known in family research since the popular actor-partner interdependence models (APIM) require this data structure (Kenny and la Voie 1985; Kenny et al. 2006; Ledermann and Kenny 2012). APIMs measure bidirectional effects in interpersonal relationships. To transform the data set from the wide format to the pairwise format, we create a copy of the original data set with anchor and partner allocation exchanged (see Figure 3.2 c).¹ After restructuring, “current life satisfaction of

¹The pairwise format that is used in APIMs is also sometimes called double entry format (Ledermann

3. Systematically Missing Partner Variables

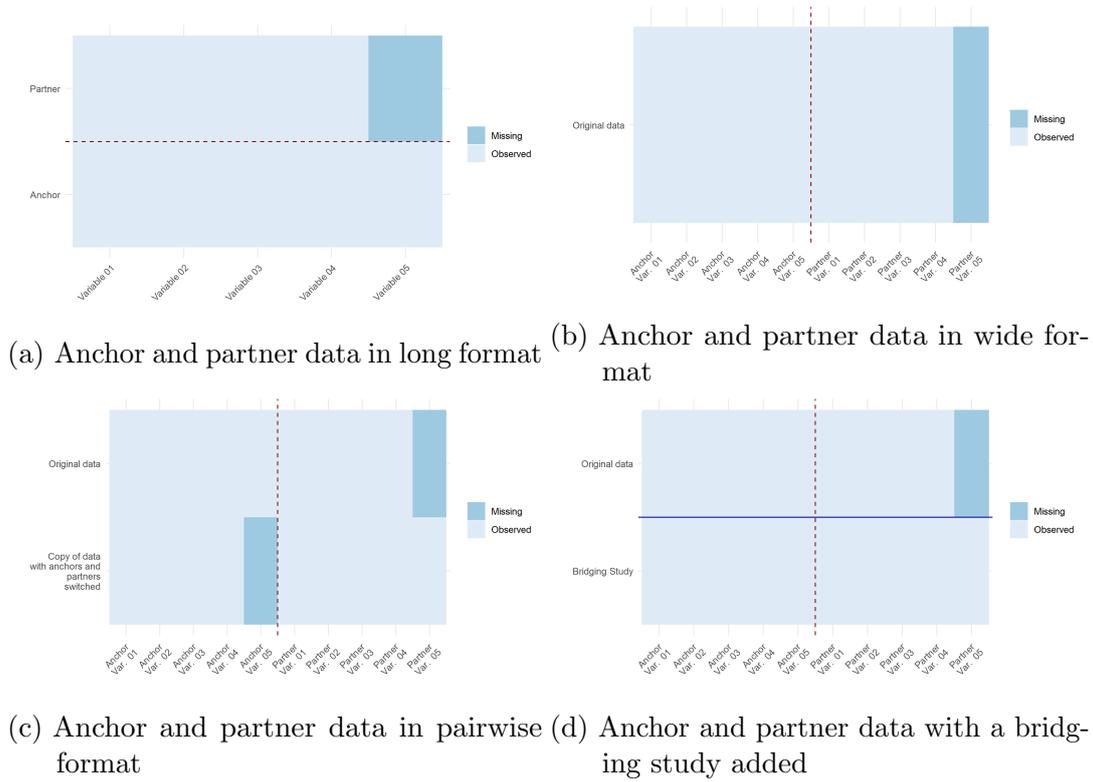


Figure (3.2) Data format for different imputation approaches. The red line separates anchor and partners, the blue line separates the additional bridging study from the original data set.

partner” is filled, and “current life satisfaction of anchor” is empty for the copied data set. This allows us to impute the missing partner life satisfaction conditional on all other anchor *and* partner variables with a standard FCS procedure. However, while imputing, we still implicitly assume the partial correlation between anchor’s and partner’s life satisfaction given all other anchor and partner variables to be zero since both variables are never observed together (see Figure 3.2 c).

Both imputation approaches that we have covered so far, MI in long (2.) or pairwise format (3.) are thus built on the assumption of conditional independence. We assume that the partial correlation between the systematically missing partner variable and its corresponding anchor variable given other variables is zero. While the two imputation approaches are relatively easy to implement, we will see whether this strong assumption

and Kenny 2015) and, while widely popular in dyadic analyses, is unusual in most other fields.

3. Systematically Missing Partner Variables

holds in real-world settings.

Suppose we do not wish to assume conditional independence but instead suspect remaining dependence between the two variables. In that case, we need to make reasonable assumptions about the amount of dependence. External information as a “bridging study” allows us to make such assumptions on the (conditional) dependence between the anchor and the (systematically missing) partner variable. There are two ways to use the information on anchor and partner respondents from bridging studies.

One way is to add the bridging study to the original data set after the ex-post survey harmonization (see Figure 3.2 d). This allows the estimation of the relationship between anchor and partner variables from the additional data set and, therefore, also reflects the chosen assumptions (van Buuren 2017). The additional data will only be used for imputation and deleted before the analysis, except in the case that the researcher wants to use both data sets together as in ex-post survey harmonization/IPD meta-analysis. A second possibility for using the imputation from bridging studies is to estimate a regression of the variable systematically missing on the other variables (especially the corresponding anchor variable) and use estimated parameters from the bridging study in the imputation model.

We will use different artificially created and real-world bridging studies to exemplify possible pitfalls and consequences of their usage. We first present a benchmark, i.e., an ideal bridging study. We only include such a bridging study (*04. MI with an ideal bridging study*) to demonstrate a benchmark that other more realistic bridging studies will have to measure up to. Our ideal bridging study consists of a new set of 1,000 drawn observations from the GSOEP panel with replacement. Measurements, target populations, and sampling are the same for the original target data set with systematically missing data. However, real-world bridging studies will often deviate from the original data set, e.g., in the variables included or the measurements used or other characteristics

3. Systematically Missing Partner Variables

of a study, such as the mode, the year in which the survey was conducted, or the sampling and population. Ideally, the differences or changes are as small as possible (Schenker and Parker 2003). But for systematically missing partner variables, it is very unlikely that we have an ideal bridging study (with the same variables observed, measurement, target population, etc.).

Naturally, we are not only interested in the possible information gains under ideal conditions but also under real-world conditions. We will explore this topic both in the simulation and in a more general exploration of study heterogeneity in Section 3.7.

We employ two different real-world external data sets as *bridging studies* for our exemplary simulation: *pairfam* as well as the *Survey of Health, Ageing, and Retirement in Europe (SHARE)*. Even if only used as a bridging study and not for a “complete” survey harmonization project, careful data preparation and harmonization is still necessary (see Table A.7). As a reminder, these added bridging studies are used for imputation only and not for analysis; consequently, they are deleted before the analysis step.

For both bridging studies, we evaluate two approaches:

1. Appending the additional bridging study to the original data set and then imputing the systematically missing data (*05. MI with SHARE added*² and *07. MI with pairfam added*)
2. Separately estimating the parameters for the imputation model of the systematically missing variable and then using the obtained estimates to impute the values for the systematically missing variable (*06. MI with SHARE parameters* and *08. MI with pairfam parameters*). We follow the same steps as in regular imputation by the normal model as defined by Rubin (1987), except that we use the data from the bridging study to estimate all parameters for the imputation model of the

²The Big Five personality traits are systematically missing for SHARE in the included wave. For *05. MI with SHARE added*, we only included the socio-demographic variables of anchors and partners as predictors for the imputation of partners’ current life satisfaction.

3. Systematically Missing Partner Variables

systematically missing variable. Apart from that, we follow the same steps as in the regular `mice.impute.norm`-function (see Appendix A for the steps).³

3.5. Data and simulation

3.5.1. Data sets: GSOEP, SHARE and pairfam

We will exemplify the consequences of different possible imputation approaches with data from the German Socio-Economic Panel (Goebel et al. 2019) and, as additional bridging studies, the German pairfam panel (Brüderl et al. 2017) as well as the German sub-study of the Survey of Health, Ageing and Retirement in Europe (Börsch-Supan et al. 2013; Börsch-Supan 2020). All three studies are multi-actor surveys and include self-reported information from the partner respondents; they collect similar data and provide information on cohabiting couples in Germany around the year 2013. While there is substantial overlap regarding several aspects of the study design – which is why we chose pairfam and SHARE as bridging studies – the three studies have significant differences. In the simulation, we will examine how these similarities and differences allow or prevent their use as a bridging study.

The German Socio-Economic Panel is a longitudinal household survey of about 11,000 private households in Germany. We will use data from the wave “bd” (2013). We restricted the data set to persons cohabiting with their partners. The study then encompasses 5584 relationships. Variables include information on household composition, employment,

³Since the later steps are not straightforward to implement, we will explain the implementation of both approaches in R with `mice`. The R package `mice` does not allow us to specify point coefficients for the imputation models directly as optional arguments to `mice.impute.norm` (Bayesian linear regression/the normal model) or the function for imputations by predictive mean matching `mice.impute.pmm`. However, it is possible to specify an imputation model with the built-in approach for “passive imputation” usually used for transformed, combined, or recoded versions of variables. This function can draw from the conditional predictive distribution of the missing partner variables with the coefficients drawn from the coefficients posterior via the sampling distribution of the bridging study parameters.

3. *Systematically Missing Partner Variables*

occupation, health, and satisfaction indicators. Socio-demographics, psychological factors, and life satisfaction were observed for both anchors and their partners. To evaluate the different imputation approaches, we delete the information on the partners' life satisfaction, thus intentionally creating systematically missing partner information. The data set's richness allows us to examine different imputation approaches by comparing their performance to the "gold standard" (the original data set).

Bridging studies employed in this article are the SHARE and pairfam surveys. SHARE (Börsch-Supan et al. 2013) surveys about 140,000 individuals aged 50 or older, covering 27 European countries and Israel. With the younger cohorts missing from SHARE, we already notice significant differences in comparison to GSOEP. We used data from the fifth wave of the survey, as it was conducted the same year as the GSOEP, wave "bd" (2013). We selected a subsample as a bridging study to increase the similarity between the samples: we selected only persons in Germany living together with their partners. The resulting SHARE bridging study then encompasses 1920 relationships. We only carry out the multiple imputation procedure with the reduced variable set of socio-demographic variables since the Big Five personality traits were not available for the fifth SHARE wave.

The second study, the pairfam panel (Brüderl et al. 2017), offers data on both anchor respondents and their partners (as well as other family members). We use data from the second wave of the pairfam panel, conducted around the same time (2011) as the selected GSOEP and SHARE waves: 2,192 cohabiting partnerships for which both anchor and partner responded to the survey. By its own description, it is a "multi-disciplinary, longitudinal study for researching partnership and family dynamics in Germany." We use it since it is a data set with rich information on both anchors and partners. A peculiarity is its target population and sample: it consists of a nationwide random sample of about 12,000 persons from the three birth cohorts 1971-73, 1981-83, 1991-93, and their partners,

3. Systematically Missing Partner Variables

parents, and children.

Before we move on to the different multiple imputation approaches, we take a look at the differences between the data sets. The measurement of variables is not identical between surveys. There are, for example, differences between pairfam and GSOEP. In the GSOEP, the Big Five personality traits were surveyed through three items per Big Five trait and on a 7-point scale instead of a 5-point scale. We harmonized the GSOEP Big Five variables by shrinking them linearly to a 5-point scale (Singh 2020). For the mapping of all variables used as predictors to a common scale, see Table A.7 in the Appendix. Apart from the Big Five personality factors, we harmonized variables for anchor's and partner's sex, age, health status in the past four weeks, indicators for being employed and being in education, as well as the number of biological children. The variable life satisfaction (systematically missing for partners) was measured in all surveys on an 11-point scale with the same alignment and the same question text.

3.5.2. Simulation conditions

We now briefly cover the simulation parameters and the settings.

For each simulation run, we randomly select 1,000 observations with replacement from the GSOEP cohabiting relationships from the survey year 2013. Life satisfaction for the partner respondent is deleted before testing all MI approaches. We use all available data for the bridging studies, i.e., data from 1,920 relationships from SHARE and 2,192 relationships from pairfam. We use 1,000 repetitions to compare all approaches.

We will start by examining the estimates of the correlation coefficient between anchor's and partner's life satisfaction to get an impression of the amount of dependence that can be recovered with simple MI approaches based on the assumption of conditional independence. As we will see, this assumption is problematic for the estimation of the correlation coefficient.

3. Systematically Missing Partner Variables

The next section will then further demonstrate the inappropriateness of the conditional independence assumption. Our second target analysis is a regression of anchor’s life satisfaction on partner’s life satisfaction as well as socio-demographics and the Big Five personality traits for the anchors *and* partners. In the target analysis, the information on anchor and partners is not used in long format (where all anchors and partners are in separate rows, see Figure 3.2 a), but in wide format, i.e., there is one row for each relationship with anchor and partner information in separate variables (see Figure 3.2 b). For all simulation runs and methods tested, every missing value is imputed five times ($m = 5$), and the number of iterations is set to 10. We checked convergence plots for every method and found no problems. We then use standard analysis for each of the $m = 5$ imputed data sets, enabling us to calculate the imputed data estimate and its estimated variance using Rubin’s rules (Rubin 1976). We use the package `mice` (Version 3.5.0) for `R` (Version 3.5.1).

3.6. Results

We now give an overview of the results of the simulation. We will first present the results for the estimates of the correlation between anchor’s and partner’s life satisfaction and, subsequently, the estimates of the regression coefficients.⁴ We examine both point estimates and standard errors. Point estimates are presented in violinplots (Hintze and Nelson 1998), a visualization technique developed from boxplot and kernel density plots showing the distribution of data across several levels of a categorical variable. This allows us to compare these distributions easily. Standard errors are presented as boxplots.

⁴We also included simulation results for the mean and standard deviation of partner’s life satisfaction in the Appendix.

3.6.1. Correlation results

We first look at the estimated *raw* correlation coefficients between the anchor respondent and the partner respondent's current life satisfaction. We compare the results after MI in long format with the results from the full data sets (before the artificial deletion of the partner's life satisfaction values). We notice that after MI in long format (2.) the estimated point coefficients for the correlation between the anchor's and partner's current life satisfaction are strongly biased towards zero (see Figure 3.3). The results are biased since with MI in long format (see the data format in Figure 3.2 a), the dependency between anchors and partners is not properly taken into account.

The bias towards zero for the estimated correlation coefficient is also prevalent after MI in pairwise format (3.). While information from both anchors *and* partners are now included in the imputation model with the data in pairwise format, the implicit assumption of conditional independence still leads to biased estimates and is not appropriate. This problem has also been noticed in other contexts related to systematically missing data, such as record linkage (Thibaudeau 1993; Winkler 1989). It is recommended to incorporate more suitable assumptions about the conditional dependence between the two variables than conditional independence (Bosch and Gaffert 2017).

To get an idea of what can be achieved with a bridging study, we include an ideal bridging study with high observation numbers and no differences in measurements, variables, or sampling. The point estimates are unbiased after MI with a bridging study consisting of another sample of 1,000 observations from the GSOEP (4.). Standard errors are higher than for the full data set due to the loss of information from missing data (Carpenter and Kenward 2013, 54).

We now move on to the real-world bridging studies (5. to 8.). While we still underestimate the regression coefficient after MI when using the information in either pairfam or SHARE, the estimates are much less biased than under the assumption of conditional independence.

3. Systematically Missing Partner Variables

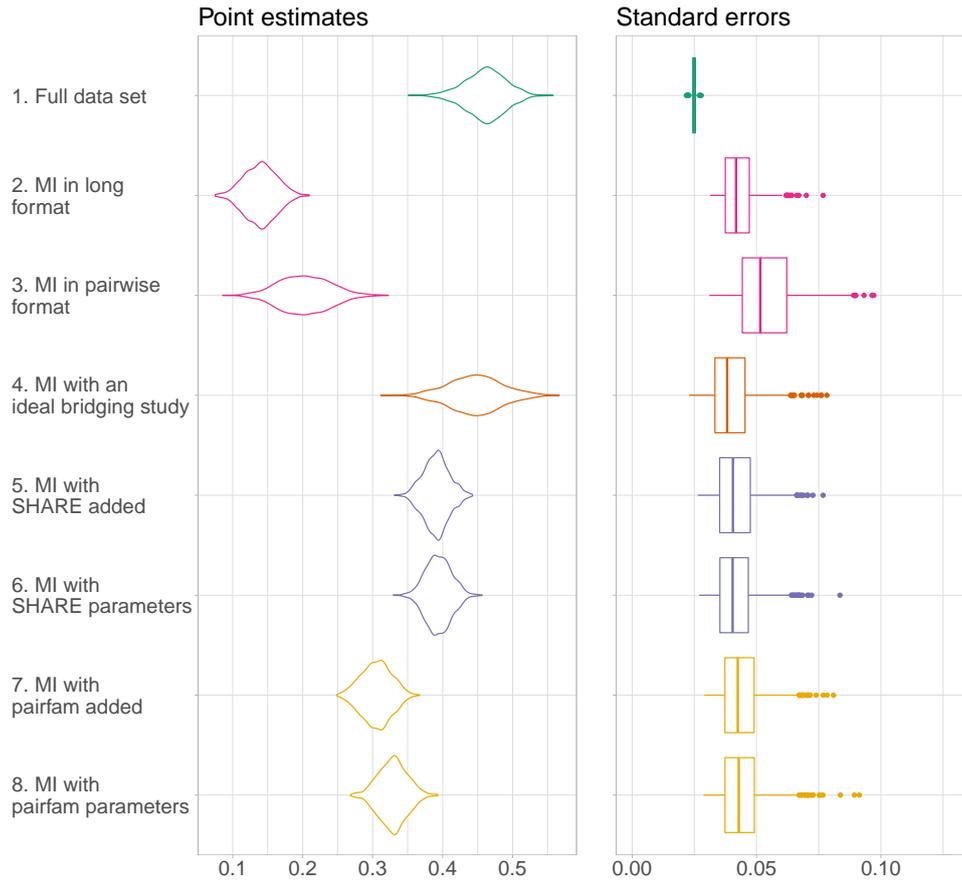


Figure (3.3) Left: Violinplot (Hintze and Nelson 1998) of estimated correlation coefficients between anchor’s current life satisfaction and partner’s current life satisfaction. Right: Boxplot of standard errors of the correlation between anchor’s satisfaction with current life and partner’s satisfaction with current life. Estimates are from 500 simulation repetitions for each method.

While bridging studies can provide the information required in the imputation model, researchers should be aware that they exchange the assumption of conditional independence for another assumption: that relationships observed in one study are transferable to another, i.e., that heterogeneity between studies is not too large. One assumption is not necessarily more appropriate than the other and does not always lead to a lower bias. In Section 3.7, we take a closer look at study heterogeneity and how much estimates vary between studies in practice.

The differences between appending the bridging studies pairfam and SHARE to the

3. Systematically Missing Partner Variables

original data set (5. and 7.) and using the sampling distribution of the estimates from the bridging studies for the imputation model (6. and 8.) are minor here. Practical considerations may be more important in making the difference for one option over another. For example, if only aggregate data, i.e., estimated regression coefficients, are available as information, it can be easier to use the parameters. Suppose the bridging study and the original data set are to be analyzed in an IPD meta-analysis. In that case, there is little reason to first estimate the parameters for the imputation model separately, and both data sets can be used together for the imputation.

3.6.2. Regression coefficient results

In this section, we focus on the *partial* correlation between the anchor's current life satisfaction and the partner's current life satisfaction. We conduct two separate simulations:

- Simulation 1: The variables in the analysis model are the same as in the imputation model (both socio-demographic and Big Five personality trait variables apart from partner's life satisfaction are included as predictors). The dependent variable is the life satisfaction of the partner. Results are reported in Figure 3.4.
- Simulation 2: The analysis model only includes socio-demographic variables. The imputation model still consists of the same variables as before (both the socio-demographic and Big Five personality traits variables). The imputation model is now *richer* than the analysis model. Auxiliary variables are used for imputation for two reasons: If the auxiliary variables are good predictors of missing values, they will help recover missing information. Plus, if they are also good predictors of the missingness, they may even correct bias (Carpenter and Kenward 2013). Results are reported in Figure 3.5.

3. Systematically Missing Partner Variables

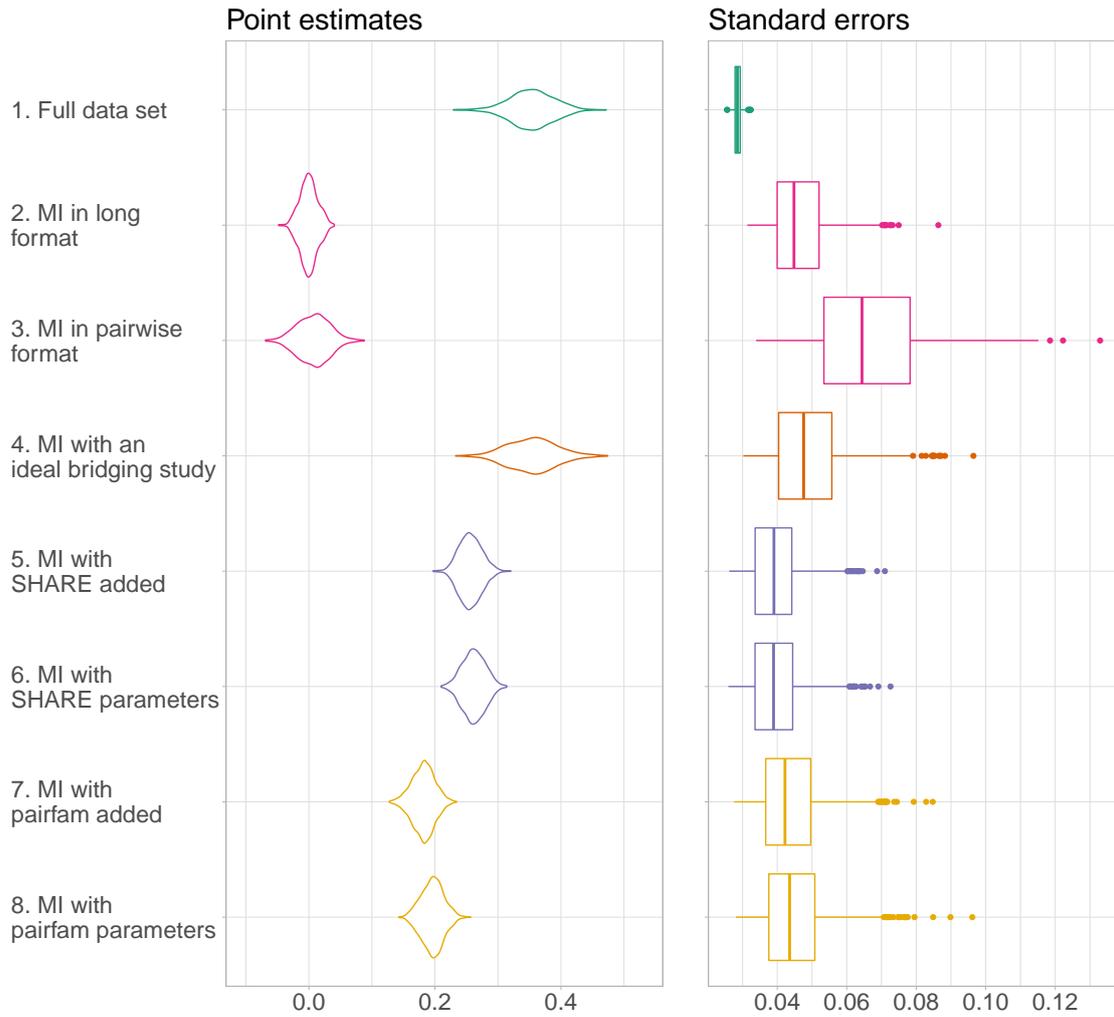


Figure (3.4) Simulation 1 (dependent variable of the target analysis: anchor’s life satisfaction, independent variables: socio-demographic variables and Big Five personality traits for anchors and partners plus partner’s life satisfaction). Left: Violinplot of estimated regression coefficients of partner’s satisfaction with current life. Right: Boxplot of standard errors. Estimates are from 500 simulation repetitions for each method.

When looking at the results from the first simulation, where the set of variables included in the imputation and analysis is identical, we immediately notice that the density of the estimated point coefficients after both MI in long format (2.) and MI in pairwise format (3.) are centered around zero and, therefore, heavily biased (Figure 3.4). When we impute the data in long format, we neglect the fact that anchor and partner’s variables are correlated with each other. Similarly, when we assume conditional independence

3. Systematically Missing Partner Variables

between an anchor and partner variable pair during imputation with the data in pairwise format, we will receive analysis results that show this independence after imputation.

Regarding bridging studies, the ideal bridging study performs well (4. in Figure 3.4). SHARE as a bridging study, either with the data set added (5. and 7.) or with the use of parameters (6. and 8.), leads to estimated regression coefficients close to those of the full data set. The use of pairfam data (either as added data set or using the parameters estimated from that data set) leads to bias towards zero for the point estimate, but less so than after MI in long format and MI in pairwise format.

In our case, there is thus no big difference between adding a bridging study and using parameters that were estimated separately from the bridging study before imputing. One has to be very careful about the generalizability of these results. Results could differ more substantially if the overlap in variables between the bridging study and the original study is smaller, or if relationships between variables differ more strongly between studies and there is also a high level of sporadically missing data in the original study. It is advisable to test both approaches to get an idea of the stability of results. We also recommend to look at several studies or a meta-analysis than just one study to get a better understanding of the heterogeneity of effects between studies.

Summarizing the results of the first simulation, we find that the assumption of conditional independence leads to non-optimal results. As we suspected, the assumption of conditional independence is not appropriate. Using bridging studies (or their parameters) leads to better results in regard to lower bias. However, the two bridging studies perform differently. These differences could result from the strong differences in target population between the original data set and the respective bridging study. We will take a closer look at what can be expected in terms of the magnitude of study heterogeneity in Section 3.7.

3. Systematically Missing Partner Variables

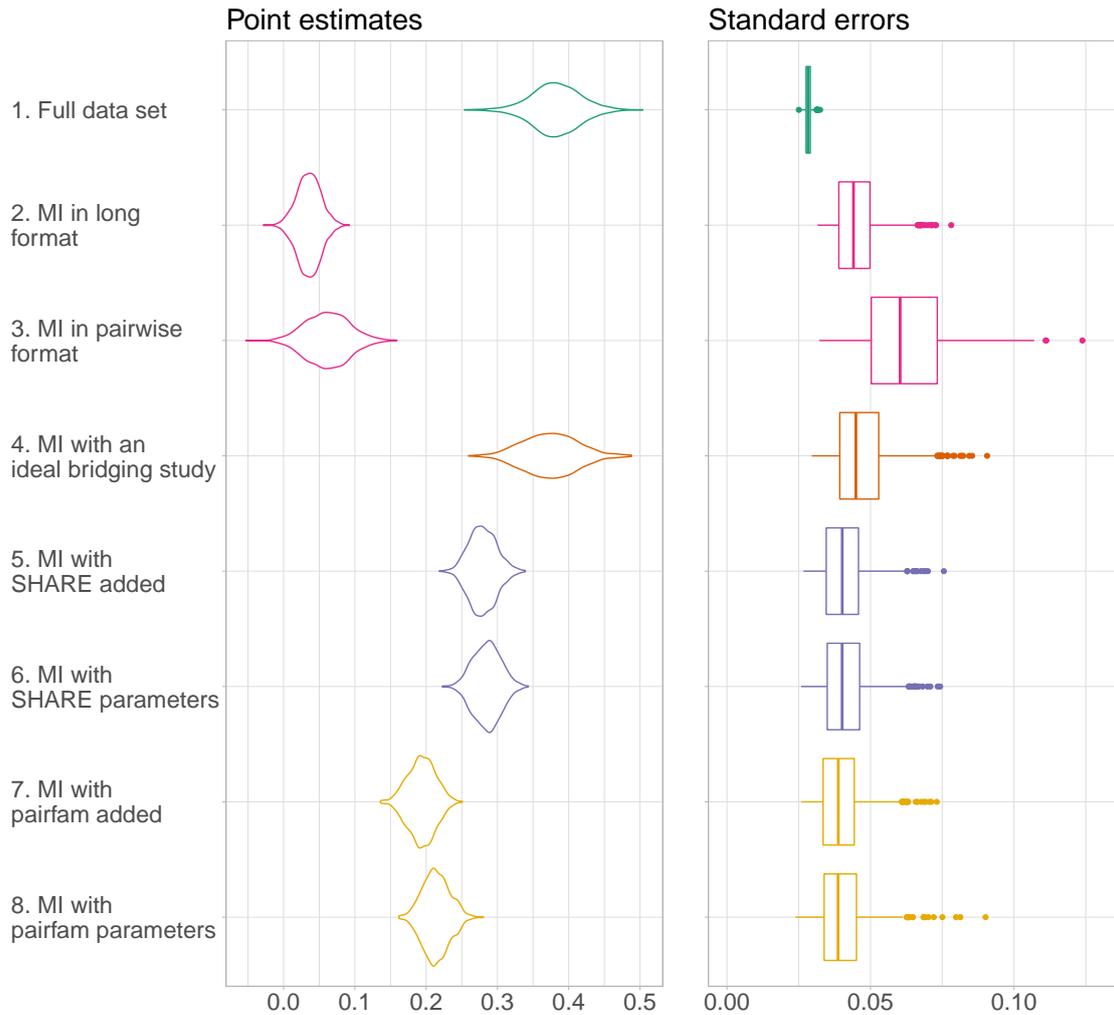


Figure (3.5) Simulation 2 (dependent variable: anchor’s satisfaction with current life, independent variables: only socio-demographic variables, Big Five excluded). Left: Violinplot of estimated regression coefficients of partner’s satisfaction with current life. Right: Boxplot of standard errors. Estimates from 500 simulation repetitions for each method.

Before turning to study heterogeneity, we report the second simulation results, in which we did not include the Big Five personality traits in the *analysis model*. We still include them as auxiliary variables (Enders 2017) in the imputation model when possible (so for all approaches except for SHARE since they were not available in the respective wave). Our imputation model for 02.–04. and 07./08. is thus *richer* than the analysis model. The results for the second simulation strongly resemble the results of the first simulation

3. Systematically Missing Partner Variables

(Figure 3.4 and Figure 3.5 for the second). The point estimates after MI in long format and MI in pairwise format are again biased downwards in the second simulation. They are, however, not centered around zero anymore. This is because the analysis model does not include all variables from the imputation model. These additional/auxiliary variables are correlated with anchor's and partner's current life satisfaction, and as a result, the regression coefficient of partner's life satisfaction is not zero. Similar to the first simulation, MI with pairfam and now also SHARE leads to the underestimation of the point coefficients. Again, the bias is not as substantial as after MI in pairwise format or after MI in long format.

From our previous simulation and analyses, we can draw several conclusions. The assumption of conditional independence is not suited for the imputation of partner variables in our case since it leads to a substantial underestimation of the correlation between anchor and partner variables.⁵

Ideal bridging studies have a sufficiently high number of observations and are similar to the original study. Usually, regression coefficients are assumed to be more robust than means, shares, or variables. After using a bridging study for MI, results have nevertheless to be treated carefully and with the possibility in mind that the studies could be too heterogeneous. Our last section will look at a bigger set of estimated regression coefficients between anchor and partner variables and how they differ between the surveys. This will give a better insight into how much heterogeneity is to be expected in practice.

⁵We want to note that estimates like the mean and the standard error of the partner variable can be imputed without an additional bridging study in our case. Since there is no systematic difference between anchors and partners in our case, both mean and standard error do not differ systematically between anchors and partners in the full data set and after MI in pairwise format (see Appendix Figure A.1).

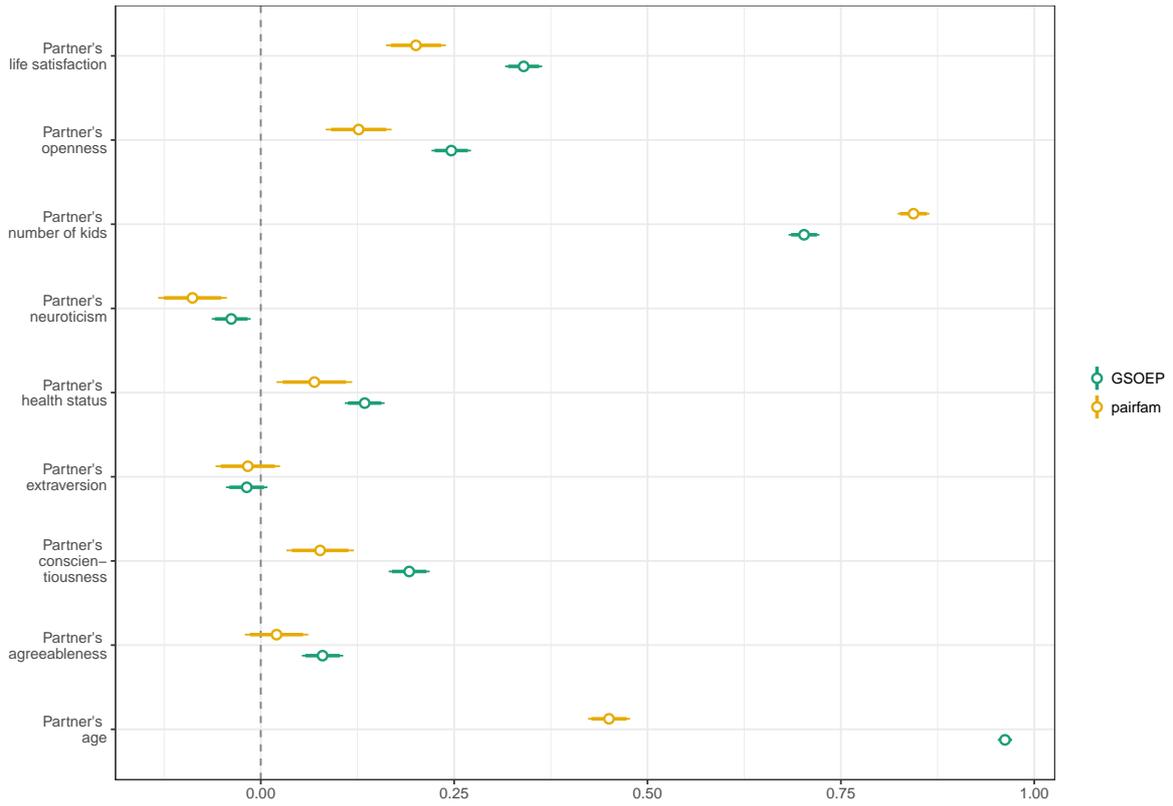
3.7. Study heterogeneity

We have seen that study heterogeneity can hinder the use of bridging studies. But how much study heterogeneity can we expect for other research questions of data sets? This question cannot be answered once and for all but will depend on the variables selected, the target population, sampling, modes, and other study characteristics. However, there are some possibilities to assess study heterogeneity regarding the regression coefficients. For example, meta-analyses or systematic reviews can be the first source to check whether study heterogeneity is likely to affect correlations or regression estimates in the respective research field.

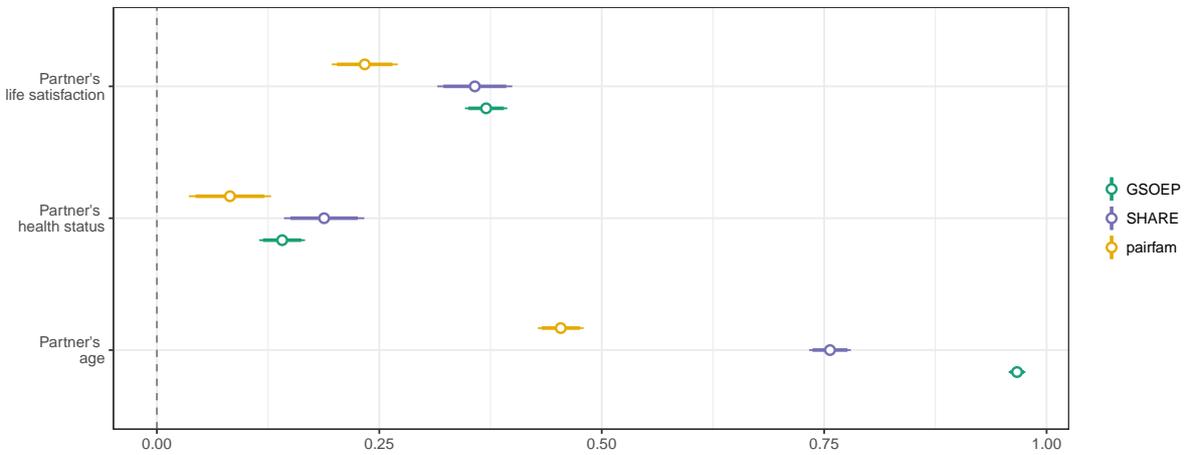
To better understand the possible effects of study heterogeneity for the use of bridging studies in MI, a series of regression estimates for the relationship between the different anchor and partner variables in the pairfam, GSOEP, and SHARE data sets are shown in Figure 3.6. We restricted all data sets to anchors cohabiting with their partner and for whom data on partners is available.

For every anchor variable as the outcome variable, i.e., for every estimate displayed, a separate model was estimated. The estimated regression coefficient for every partner variable (always with the corresponding anchor variable as the dependent variable) is displayed. For Figure 3.6 a) all other socio-demographic *and Big Five personality traits* variables for both anchor and partner are always included as predictors, i.e., are conditioned on. For Figure 3.6 b) all other socio-demographic variables for both anchor and partner are included as predictors. While regression estimates do vary between surveys (and often significantly), they are, in general, of the same direction and of similar magnitude between surveys. These results demonstrate the possibility of the (careful) use of bridging studies, at least in supplementary and sensitivity studies.

3. Systematically Missing Partner Variables



(a)



(b)

Figure (3.6) Point estimates of partner variables. Shown is always the regression estimate for the partner variable corresponding to the current outcome (anchor) variable. Bars show the 95%-confidence interval.

(a) Socio-demographic and *Big Five* variables for both anchor and partner are included as predictors.

(b) Socio-demographic variables for both anchor and partner are included as predictors.

3. Systematically Missing Partner Variables

However, a notable exception is the partner's age, where the estimated regression coefficients differ substantially between the surveys. Remarkably, the heterogeneity is largest for these characteristics since the three studies differ in their sampling, especially regarding the age cohorts (GSOEP surveys the general population aged 16 and older, SHARE only includes anchors aged 50 or older, and pairfam surveys only anchors from the birth cohorts 1971-73, 1981-83, 1991-93). Thus, the assumption of transferability is not always unproblematic and not to be made without care, even if most estimates of dependency are quite stable in our case.

3.8. Discussion

This chapter explored if and how multiple imputation can be used when information on partner respondents is systematically missing for some variables. We used a motivational example from partnership research to exemplify the potentials as well as the pitfalls of multiple imputation and bridging studies.

In our motivational example, we used the German Socio-Economic Panel (GSOEP), in which the variables of interest, partners' and anchors' life satisfaction, were initially observed. By deleting all information on partners' life satisfaction, we artificially created a situation of systematically missing data in partner variables. We used different *multiple imputation* approaches to impute the missing information for partners' life satisfaction, with and without the use of *bridging studies*. We tested two different imputation approaches without additional bridging studies. The first option was to impute while the data set is in long format, i.e., providing the information on anchors and partners in separate rows. Missing values are then imputed depending on the partner's observed value but *not* depending on the values of the respective anchor respondent in the relationship. This led to serious underestimation of the interdependence of anchor and partner respondents' life satisfaction. The second option was to impute with the data in

3. Systematically Missing Partner Variables

pairwise format (observation units are relationships and anchor and partner's information are combined in one row), thus assuming conditional independence between partners' and anchors' life satisfaction, given all other observed anchor *and* partner variables. However, the results showed that this assumption is still unrealistic in our case.

Based on methods in related fields like survey harmonization, measurement changes in official statistics, and split questionnaire design, we explore the use of bridging studies to impute systematically missing variables (Schenker and Parker 2003; Schenker and Raghunathan 2007; Resche-Rigon and White 2016).

We started with an "ideal" bridging study to demonstrate the best-case scenario, another subsample randomly drawn from the GSOEP data that we used for the simulation. To simulate a more realistic situation, we additionally chose different studies to serve as bridging studies – SHARE (Börsch-Supan et al. 2013; Börsch-Supan 2020) and pairfam (Brüderl et al. 2017).

We noticed that the results using bridging studies are better in terms of bias than the two approaches based on the conditional independence assumptions. However, we also see that using bridging studies to impute systematically missing variables is also based on the (strong) assumption, that study heterogeneity is not too great to hinder the transfer of relationships between studies. The use of bridging studies may be more appropriate to explore what-if scenarios, for example, as supplementary or sensitivity analysis.

Our study also highlights that there are two ways of using the information from bridging studies for imputation. One is to append the bridging study (only) to the data set with the missing data for imputation, and the other is to separately estimate the parameters of the imputation model for the systematically missing variable and use them in the imputation.

The differences between the two ways to employ bridging studies (added data set vs. the use of parameters estimated from the bridging study) were minor in our example.

3. Systematically Missing Partner Variables

However, we want to mention two crucial differences between the two approaches and not deny that they could have bigger consequences in other circumstances. If we impute with the data set added to the original data set, the imputation models for the *sporadically* missing variables will by default also be estimated with information from the combined data set. Thus, the bridging study will also influence the imputation of the sporadically missing variables and then, in turn, also the imputation of the systematically missing variable. First estimating the imputation model parameters and using them as part of the passive imputation routine in `mice` avoids this problem. However, users should also note the downside of the passive imputation approach; the parameters of the imputation model of the systematically missing variable are *not* updated every iteration; they remain fixed. An area for future research would be to use the information as (informative) prior in the imputation; this way, it would be possible to update the parameters. However, it is not possible to specify priors for the imputation model in `mice` at the moment.

Another question to be further investigated is that of studies' heterogeneity and the robustness of research results over different studies and measurements. We saw in Section 3.7 that correlations between variables could vary between studies. This problem has become much more visible in recent years with increased research synthesis efforts and the replication crisis. These efforts are connected to our article in two ways. First, meta-analyses or literature reviews can inform us how stable the correlation between two variables is between studies. Second, if available, we can also use the meta-analytical results to impute the systematically missing information. In that sense, we point towards the need for more vigorous research synthesis efforts to consolidate knowledge about isolated research questions, and make them fruitful for other related research areas.

Bibliography

Adigüzel, F. and Wedel, M. (2008). Split Questionnaire Design for Massive Surveys. *Journal of Marketing Research*, 45(5):608–617.

Bosch, V. and Gaffert, P. (2017). Multiple Imputation in Data Fusion: Making Better Assumptions than Conditional Independence. Joint Statistical Meeting 2017.

Brüderl, J., Hank, K., Huinink, J., Nauck, B., Neyer, F. J., Walper, S., Alt, P., Borschel, E., Buhr, P., Castiglioni, L., Friedrich, S., Finn, C., Garrett, M., Hajek, K., Herzig, M., Huyer-May, B., Lenke, R., Müller, B., Peter, T., Schmiedeberg, C., Schütze, P., Schumann, N., Thönnissen, C., Wetzel, M., and Wilhelm, B. (2017). The German Family Panel (pairfam). Technical Report ZA5678 Data File Version 8.0.0, GESIS Data Archive, Cologne.

Byrne, D. E. (1971). *The Attraction Paradigm*, volume 462. Academic Press.

Börsch-Supan, A. (2020). Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 5. Release Version: 7.1.0. SHARE-ERIC.

Börsch-Supan, A., Brandt, M., Hunkler, C., Kneip, T., Korbmacher, J., Malter, F., Schaan, B., Stuck, S., and Zuber, S. (2013). Data Resource Profile: The Survey of Health, Ageing and Retirement in Europe (SHARE). *International Journal of Epidemiology*, 42(4):992–1001.

Bibliography

- Carpenter, J. and Kenward, M. (2013). *Multiple Imputation and its Application*. Wiley.
- CLOSER (2020). The Home of Longitudinal Research. <https://www.closer.ac.uk/>. [Online; accessed 20-September-2020].
- Dykstra, P., Kalmijn, M., Knijn, T., Komter, A., Liefbroer, A., and Mulder, C. (2012). Codebook of the Netherlands Kinship Panel Study: A Multi-Actor, Multi-Method Panel Study on Solidarity in Family Relationships, Wave 2, Version 2.0.
- Dyrenforth, P. S., Kashy, D. A., Donnellan, M. B., and Lucas, R. E. (2010). Predicting Relationship and Life Satisfaction from Personality in Nationally Representative Samples From Three Countries: The Relative Importance of Actor, Partner, and Similarity Effects. *Journal of Personality and Social Psychology*, 99(4):690.
- Edwards, J. N. (1969). Familial Behavior as Social Exchange. *Journal of Marriage and Family*, 31(3):518–526.
- Enders, C. K. (2017). Multiple Imputation as a Flexible Tool for Missing Data Handling in Clinical Research. *Behaviour Research and Therapy*, 98:4 – 18. Best Practice Guidelines for Modern Statistical Methods in Applied Clinical Research.
- Goebel, J., Grabka, M. M., Liebig, S., Kroh, M., Richter, D., Schröder, C., and Schupp, J. (2019). The German Socio-Economic Panel (SOEP). *Jahrbücher für Nationalökonomie und Statistik*, 239(2).
- Granda, P., Wolf, C., and Hadorn, R. (2010). *Harmonizing Survey Data*, chapter 17, pages 315–332. John Wiley & Sons.
- Gustavson, K., Røysamb, E., Borren, I., Torvik, F. A., and Karevold, E. (2016). Life Satisfaction in Close Relationships: Findings From a Longitudinal Study. *Journal of Happiness Studies*, 17(3):1293–1311.

Bibliography

- HaSpaD (2020). HaSpaD - Harmonizing and Synthesizing Partnership Histories from Different Research Data Infrastructures. <https://www.gesis.org/forschung/drittmittelprojekte/projektuebersicht-drittmittel/haspad-harmonisierung-und-synthese-von-paarbiografischen-daten>. [Online; accessed 20-September-2020].
- Hintze, J. L. and Nelson, R. D. (1998). Violin Plots: A Box Plot-Density Trace Synergism. *The American Statistician*, 52(2):181–184.
- Jolani, S. (2017). Hierarchical Imputation of Systematically and Sporadically Missing Data: An Approximate Bayesian Approach Using Chained Equations. *Biometrical Journal*, 60(2):333–351.
- Jolani, S., Debray, T., Koffijberg, H., van Buuren, S., and Moons, K. (2015). Imputation of Systematically Missing Predictors in an Individual Participant Data Meta-analysis: A Generalized Approach Using MICE. *Statistics in Medicine*, 34(11):1841–1863.
- Kalmijn, M. (1998). Inter-marriage and Homogamy: Causes, Patterns, Trends. *Annual Review of Sociology*, 24(1):395–421.
- Kalmijn, M., Ivanova, K., van Gaalen, R., de Leeuw, S. G., van Houdt, K., van Spijker, F., and Hornstra, M. (2018). A Multi-Actor Study of Adult Children and Their Parents in Complex Families: Design and Content of the OKiN Survey. *European Sociological Review*, 34(4):452–470.
- Kalmijn, M. and Liefbroer, A. C. (2010). Nonresponse of Secondary Respondents in Multi-Actor Surveys: Determinants, Consequences, and Possible Remedies. *Journal of Family Issues*, 32(6):735–766.
- Kenny, D. A., Kashy, D. A., and Cook, W. L. (2006). *Dyadic Data Analysis*. Guilford Press.

Bibliography

- Kenny, D. A. and la Voie, L. (1985). Separating Individual and Group Effects. *Journal of Personality and Social Psychology*, 48(2):339–348.
- Ledermann, T. and Kenny, D. A. (2012). The Common Fate Model for Dyadic Data: Variations of a Theoretically Important but Underutilized Model. *Journal of Family Psychology*, 26(1):140.
- Ledermann, T. and Kenny, D. A. (2015). A Toolbox With Programs to Restructure and Describe Dyadic Data. *Journal of Social and Personal Relationships*, 32(8):997–1011.
- Little, R. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. Wiley.
- Parker, J., Schenker, N., Ingram, D., Weed, J., Heck, K., and Madans, J. (2004). Bridging Between Two Standards for Collecting Information on Race and Ethnicity: An Application to Census 2000 and Vital Rates. *Public Health Reports*, 119(2):192–205.
- Raghunathan, T. E. (2006). Combining Information From Multiple Surveys for Assessing Health Disparities. *Allgemeines Statistisches Archiv*, 90(4):515–526.
- Resche-Rigon, M., White, I., Bartlett, J., Peters, S., and Thompson, S. (2013). Multiple Imputation for Handling Systematically Missing Confounders in Meta-Analysis of Individual Participant Data. *Statistics in Medicine*, 32(28):4890–4905.
- Resche-Rigon, M. and White, I. R. (2016). Multiple Imputation by Chained Equations for Systematically and Sporadically Missing Multilevel Data. *Statistical Methods in Medical Research*, 27.
- Rubin, D. B. (1976). Inference and Missing Data. *Biometrika*, 63(3):581–592.
- Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley.
- Rässler, S. (2003). A Non-Iterative Bayesian Approach to Statistical Matching. *Statistica Neerlandica*, 57(1):58–74.

Bibliography

- Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman and Hall, London.
- Schenker, N. and Parker, J. (2003). From Single-Race Reporting to Multiple-Race Reporting: Using Imputation Methods to Bridge the Transition. *Statistics in Medicine*, 22(9):1571–1587.
- Schenker, N. and Raghunathan, T. (2007). Combining Information From Multiple Surveys to Enhance Estimation of Measures of Health. *Statistics in Medicine*, 26(8):1802–1811.
- SDR (2020). Survey Data Recycling Project. <https://www.asc.ohio-state.edu/dataharmonization/>. [Online; accessed 20-September-2020].
- Singh, R. K. (2020). Harmonizing Instruments with Equating. *Harmonization Newsletter on Survey Data Harmonization in the Social Sciences*, 6(1).
- Skopek, J., Schulz, F., and Blossfeld, H. P. (2011). Who Contacts Whom? Educational Homophily in Online Mate Selection. *European Sociological Review*, 27(2):180–195.
- Thibaudeau, Y. (1993). The Discrimination Power of Dependency Structures in Record Linkage. *Survey Methodology*, 19(1):31–38.
- van Buuren, S. (2007). Multiple Imputation of Discrete and Continuous Data by Fully Conditional Specification. *Statistical Methods in Medical Research*, 16(3):219–242.
- van Buuren, S. (2017). mice Issue 32: Get at the Final Model Used in the MICE Iterations? <https://github.com/stefvanbuuren/mice/issues/32>.
- van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, C. G. M., and Rubin, D. B. (2006). Fully Conditional Specification in Multivariate Imputation. *Journal of Statistical Computation and Simulation*, 76(12):1048–1064.

Bibliography

- White, I. R., Royston, P., and Wood, A. M. (2011). Multiple Imputation Using Chained Equations: Issues and Guidance for Practice. *Statistics in Medicine*, 30(4):377–399.
- Winkler, W. E. (1989). Methods for Adjusting for Lack of Independence in an Application of the Fellegi-Sunter Model of Record Linkage. *Survey Methodology*, 15(1):101–117.

4. TippingSens: An R Shiny Application to Facilitate Sensitivity Analysis for Causal Inference Under Confounding

4.1. Introduction

Questions of interest in the applied sciences are often questions of causality rather than description or association. The gold standard for estimating causal effects are randomized experiments. With experiments, researchers have full control over the treatment assignment process, substantially facilitating the estimation of causal effects. However, while in the medical sciences, experimental designs are widely used; they are often not feasible or unethical in epidemiological, sociological, or economic research. Instead, empirical research is usually based on pre-existing, observational data. When attempting to make causal claims from observational data, researchers have to make strong assumptions. Arguably the most controversial of these assumptions is the requirement that the assignment to the treatment is unconfounded, i.e., given all the information available in the collected data, the probability of receiving the treatment does not depend

4. *TippingSens R Shiny Application*

on the (potential) outcomes of interest (selection on observables). The unconfoundedness assumption implies that adjusting for differences in observed variables removes biases in causal estimates based on a direct comparison between treated and control groups (Imbens 2003, 126).

The unconfoundedness assumption is not always realistic, and it can never be tested based on the observed data. A possible way to proceed if the unconfoundedness assumption is suspect is to conduct sensitivity analyses under various assumptions regarding the confounding mechanism.

While intuitive in theory, conducting sensitivity analyses in practice is often difficult. Flexible approaches based on the idea of partial information (Manski 1990), which try to incorporate the uncertainty about the assignment mechanism directly, are usually not helpful in practice since the level of uncertainty about the actual causal effect will often become so large that no meaningful conclusions can be drawn based on the collected data. Thus, specific assumptions need to be postulated about the properties of the confounding variable or the assignment process. Turning these assumptions into parameters can be challenging in practice. In this chapter, we will focus on a sensitivity analysis approach proposed by Rosenbaum and Rubin (1983), which achieves this difficult task for the particular case of a binary confounding variable.

Despite limiting the range of scenarios in which the sensitivity analysis can be used, binary confounding variables seem plausible in many applications. For example, when analyzing health care records, it might be realistic to assume that the treatment assignment depends on specific health conditions such as obesity or certain blood test results, which are not available in the database for confidentiality reasons. Since treatment guidelines are often based on threshold rules (if the blood pressure is lower than x , prescribe dose y , else prescribe dose z), the confounding variable can be considered binary. Another example, which we use to illustrate the functionality of the app in Section 4.5, are confounding

4. *TippingSens R Shiny Application*

variables in the context of assignment to labor market programs. Assignment decisions will be affected by factors such as alcohol abuse or poor personal hygiene, which are observable for the case manager deciding about the treatment assignment but are not recorded in the database used when evaluating the effectiveness of various labor market programs.

Even in the simple case of a single binary confounding variable, Rosenbaum-Rubin sensitivity analyses can be cumbersome since four different parameters need to be specified (see Section 4.4.1 for details), and the impact on the analysis of interest needs to be evaluated for varying parameter combinations. Rosenbaum and Rubin (1983) work with forking tables to present the results of their sensitivity analysis. However, tables are not suited to explore a broader range of parameter sets since the resulting tables will quickly become very large and complicated to read. For example, evaluating five different values for each of the four parameters would already result in a table with $5^4 = 625$ table cells.

We postulate that this is one of the main reasons why other strategies, such as the sensitivity analysis proposed by Rosenbaum (1995) or Manski's partial information approach, are often preferred in practice. However, their main advantage – less assumptions about the confounding variables – is also their main problem: without additional assumptions, the plausible range for the causal effect often becomes so wide that no practical recommendations can be given based on the findings. If (partial) information about the confounders is available, the Rosenbaum-Rubin approach offers a flexible tool to integrate this information, helping to narrow the range of plausible values for the causal effect. Strategies to address the high dimensional table problem of the Rosenbaum-Rubin approach have been suggested in the literature. For example, Imbens (2003) suggests looking at the relationships in the observed data and picking the most extreme values as informed guesses for the unobserved parameters. The underlying assumption is that

4. *TippingSens* R Shiny Application

if the study was carefully designed, it seems unlikely that confounders exist that show higher correlations with the outcome or the treatment indicator than any of the observed correlations. Still, researchers are usually not interested in only one set of (extreme) assumptions but try to evaluate how robust their findings are under a whole range of plausible assumptions. After all, if the correct parameters were known, no sensitivity analysis would be required.

To address this problem, we developed an interactive R Shiny app called *TippingSens* that simplifies conducting Rosenbaum-Rubin sensitivity analyses for a large number of parameter sets. Our app enables researchers to gain quick insights regarding the robustness of their findings and to publish comprehensible visualizations of their sensitivity analysis. We tackle the four-dimensional parameter space through the interactive component of our app. The *TippingSens* app is freely available on the R Shiny server <https://tippingsens.shinyapps.io/TSEApp/>.

The remainder of the chapter is organized as follows: In Section 4.2, we give a brief introduction to causal inference with observational data and the problems resulting from self-selection into treatment. Different approaches to sensitivity analysis which are used to examine the robustness of the unconfoundedness assumption are presented in Section 4.3. Section 4.4 focuses on the technical details of the sensitivity approach developed by Rosenbaum and Rubin (1983) and introduces our new visualization tool *TippingSens*. Section 4.5 offers a practical illustration of the *TippingSens* app examining the robustness of an analysis by Bernhard (2016), which estimates treatment effects of vocational training programs on unemployed men in Germany. The chapter concludes with a discussion of the limitations of the approach and some suggestions for future research. A step-by-step guide how to invoke the *TippingSens* app with data from Rosenbaum and Rubin (1983) is provided in the Appendix A.3.

4.2. Rubin’s Causal Model and the assumption of unconfoundedness

4.2.1. The Rubin Causal Model

Throughout this article, we will discuss causal inference based on the potential outcomes framework developed by Rubin (1974).¹ Let Y_i be the outcome of interest for unit i , $i = 1, \dots, N$. Let W_i be an indicator which treatment unit i received. In many applications W_i is binary, i.e., $W_i = 1$ if unit i receives the treatment, and $W_i = 0$ otherwise. Under this assumption, the potential outcomes framework defines $Y_i(1)$ as the potential outcome for unit i , if unit i received the treatment, and $Y_i(0)$ defines the potential outcome if the unit did not receive the treatment. The individual (additive) treatment effect τ_i can then be computed as $\tau_i = Y_i(1) - Y_i(0)$. However, only one of the two potential outcomes will be observed. The other one will be the ex-post counterfactual, which can never be observed. This is the well-known *fundamental problem of causal inference* (Holland 1986). Thus, the individual treatment effect can never be observed directly. To be able to draw causal conclusions at least at an aggregate level, further assumptions are required. Besides the stable unit value treatment assumption (SUTVA), which requires that the outcome of unit i does not depend on whether unit j receives the treatment and that there is no unobserved variability in the treatment, a common assumption in the context of causal inference is the assumption of a regular assignment mechanism. An assignment mechanism is called regular if the assignment is individualistic, probabilistic, and unconfounded. The first two components require that the probability of assignment to the different treatments for unit i only depends on its characteristics and not on the characteristics of the other units. Furthermore, the probability of receiving any of the treatments must be strictly positive for all units. The

¹For a different perspective on this topic see Pearl et al. (2016).

unconfoundedness assumption states that the probability of assignment does not depend on the outcomes of interest given the observed characteristics available in the data. For example, in the case of a quasi-experimental medical drug test, the treatment assignment must not depend on the (expected) survival rates after conditioning on the observed covariates.

Since for randomized experiments, the assignment mechanism is under the control of the researcher, the assumption of a regular assignment mechanism is typically valid for carefully designed experiments. Furthermore, since units generally are randomly assigned to the different treatment groups independent of their characteristics, differences before the treatment are only by chance, and the treatment is called *exogenous*. This further simplifies the analysis, since average treatment effects can be estimated by directly comparing the average outcomes in the different treatment groups.

4.2.2. Quasi-experiments and the assumption of unconfoundedness

In quasi-experiments, assignment to the treatment is no longer under the control of the researcher. Without random assignment, subjects in different groups might not be comparable at the baseline, that is, before treatment. This is because the composition of different treatment groups could result from a selective process. Quasi-experimental evaluations typically impose comparability at the baseline by homogenizing treatment groups on observed characteristics, for example, through propensity score matching.

However, even though various options exist to adjust for differences at the baseline, most adjustment methods assume that the assignment mechanism is regular. While the assignment will often be individualistic and probabilistic, the unconfoundedness assumption is more controversial. This assumption is not verifiable based on the observed data, and differences in unobserved covariates between treated and controls may remain. Bias may arise if the outcome and the unobserved characteristics are correlated.

4. *TippingSens R Shiny Application*

To address this problem, researchers can evaluate the sensitivity of the results regarding the unconfoundedness assumption. Through simulations, it is possible to explore which properties the unmeasured covariate(s) need to have to substantially change the results and conclusions of the study.

Sensitivity analyses are strongly related to the study of treatment effect robustness after dropping one or more of the observed covariates (Heckman 1989; Smith and Todd 2001; Lechner and Wunsch 2013). Nevertheless, Imbens (2003, 126) stresses one of the main differences between sensitivity analysis and the study of treatment effect robustness:

The attraction of the sensitivity analysis is that it is more directly relevant: one is not interested in what would have happened in the absence of covariates observed, but in biases that are the result of not observing all relevant covariates.

4.3. Sensitivity analysis in the context of causal inference

The general idea of a sensitivity analysis was first proposed by Cornfield et al. (1959) to defend the plausibility of a causal effect of cigarette smoking on lung cancer. The authors demonstrated that the lack of such a relationship was only possible through the existence of an unmeasured confounder with an unrealistically high association with lung cancer and smoking habits. Building on these ideas, several strategies for evaluating the sensitivity of the confoundedness assumption have been proposed in the literature. These strategies can be loosely grouped into three categories based on the assumptions they require.

Perhaps the most radical approach is to drop the assumption of exogeneity/unconfoundedness completely, specifying a range of plausible values for the estimated treatment effect, which

4. *TippingSens R Shiny Application*

accounts for the additional uncertainty regarding the unknown assignment mechanism (Manski 1990). While this approach is attractive from a theoretical perspective as it requires no untestable assumptions, it also strictly limits the information that can be obtained from observational data. In practice, the uncertainty in the estimated causal effect is typically large, i.e., the uncertainty bounds that define the interval in which the true causal effect might fall are very wide. Thus, the findings often cannot provide useful information, for example, to guide decisions in an evidence-based policy setting.

Since the so-called Manski bounds are often not very informative, the other two general approaches specify limited departures from the unconfoundedness assumption instead of dropping the assumption of exogeneity/unconfoundedness completely. The approaches differ regarding the parameters that need to be specified by the user. The first approach only requires specifying the association between the unobserved confounder(s) and the treatment assignment. The second approach is limited to one confounder and additionally specifies the association between this confounder and the outcome.

Two early proponents of these two approaches are Rosenbaum (1995) (R95) and Rosenbaum and Rubin (1983) (RR83). R95 defines the association between the unobserved confounders and the treatment assignment indirectly by setting a threshold parameter, which specifies the maximum difference between the estimated treatment propensity based on the observable data and the true propensity. The researcher still takes a Manski-style approach regarding the associations between the hidden confounder and the potential outcomes. The RR83 approach additionally requires to explicitly specify the relationship between the confounder and the potential outcomes and assumes that confounding is limited to a single binary variable. Furthermore, both approaches implicitly assume that the relationships between the unobserved confounder(s) and the treatment does not vary as a function of the observed data (RR83 requires a similar assumption regarding the confounder and the outcome in the two treatment groups).

4. *TippingSens R Shiny Application*

Other approaches for sensitivity analysis proposed in the literature can mostly be classified as belonging to one of these three types. For example, Manski and Pepper (2000, 2009) extend the approach of Manski (1990), and applications are discussed in Blundell et al. (2007), Kang (2011), and Hof (2014) among others. Gastwirth et al. (1998), Rosenbaum (2010), and Rosenbaum (2018) discuss several extensions of R95 for specific settings. Practical applications of the strategy are discussed in Rosenbaum (1999), Rosenbaum (2003), Rosenbaum (2007), Rosenbaum (2010), Kitahata et al. (2009), Zubizarreta et al. (2013), and Rosenbaum (2018), and software implementations are available in Stata (Gangl 2004; Becker and Caliendo 2007) and R (Keele 2010; Rosenbaum 2014). Extensions of the ideas of RR83 are presented for example in Harding (2003), Greenland (1996), and VanderWeele and Arah (2011). Applications are discussed in Imbens (2003) and Ichino et al. (2008).

Comparing the RR83 and R95 approaches, a major advantage of R95 is that less assumptions regarding the confounding variable are required. However, not surprisingly, the uncertainty bounds for the treatment effect will be wider, limiting the conclusions that can be drawn. If additional information is available, which allows limiting the range of plausible values for the correlation between the outcome and the potential confounder, RR83 allows tightening the uncertainty bounds. Since extreme cases, such as perfect positive or negative correlation, can typically be ruled out, RR83 can often be helpful in practical settings.

A downside of RR83 is the required assumption that the confounder is univariate and binary. Still, as discussed in the introduction and as illustrated in the application in Section 4.5, this assumption can be plausible in many circumstances (see also Liu et al. 2013, who point out that the binary confounder can be seen as a combination of unobserved confounders). As Imbens and Rubin (2015) illustrate, under this assumption, the approaches of Manski (1990) and R95 can be seen as special cases of RR83, fixing

4. *TippingSens R Shiny Application*

some of the parameters at extreme values. Thus, under the assumption of a binary confounder, RR83 offers more flexibility to evaluate the impacts on the estimated causal effect under various assumptions regarding the relationship between the confounder and the outcome.

As pointed out above, we believe that the main reason for the popularity of the approaches akin to R95 is their simplicity. Since only one parameter needs to be specified, visualizing the impact of various assumptions about this parameter is straightforward. RR83-type approaches require monitoring several parameters simultaneously. While calculating the uncertainty bounds is still straightforward when fixing the parameters at specific (extreme) values as done in Imbens (2003) and Liu et al. (2013), evaluating the impacts under various assumptions can be cumbersome. However, this will be a common scenario in practice. For example, in the labor market context it seems prudent to assume that alcohol abuse has a negative impact on both, the probability of receiving the treatment, i.e., the probability of being assigned to a labor market program, and on the outcome, i.e., on the probability of finding a job. It is exactly in such a situation, where RR83 offers an advantage over R95. Even if the exact relationship between alcohol abuse and the outcome might be unknown, the knowledge that the correlation will most likely be negative can be used in the RR83 approach to narrow the uncertainty regarding the causal effect of the labor market program compared to R95.

However, monitoring the impact over a whole range of parameter settings can be difficult with RR83. Simplifying the sensitivity analysis for such a scenario through useful visualization tools was the main motivation for developing the TippingSens app. Before we describe the app in more detail and illustrate its features in an application, we briefly review the details of the RR83 approach in the next section.

4.4. The Rosenbaum-Rubin sensitivity analysis and the TippingSens App

4.4.1. Technical details

The Rosenbaum-Rubin sensitivity analysis assumes that the unconfoundedness assumption holds given an additional unobserved binary covariate U_i :

$$W_i \perp\!\!\!\perp (Y_i(0), Y_i(1)) | X_i; U_i,$$

where X_i denotes the vector of observed covariates. To evaluate the impact of this unobserved covariate on the analysis of interest, we can postulate parametric models for the marginal distribution of the unobserved covariate U , the conditional distribution of the treatment W given U and X , and the conditional distributions of the two potential outcomes $Y(w)$, $w \in \{0, 1\}$, given U and X . In the following, we drop the observed variables X to avoid clutter. The formulae provided below can easily be extended to include observed variables (see, e.g., Imbens 2003, 127) or more complex models.

Given that U_i is binary, we specify

$$q = Pr(U_i = 1) = 1 - Pr(U_i = 0).$$

The probability of receiving the treatment given U_i is modeled using a logistic regression

$$Pr(W_i = 1 | U_i = u) = \frac{\exp(\gamma_0 + \gamma_1 \cdot u)}{1 + \exp(\gamma_0 + \gamma_1 \cdot u)}.$$

Similarly, a logistic relationship is assumed between the binary outcome and U_i in both

4. *TippingSens* R Shiny Application

treatment groups:

$$Pr(Y_i(1) = 1|U_i = u) = \frac{\exp(\alpha_0 + \alpha_1 \cdot u)}{1 + \exp(\alpha_0 + \alpha_1 \cdot u)}$$

and

$$Pr(Y_i(0) = 1|U_i = u) = \frac{\exp(\beta_0 + \beta_1 \cdot u)}{1 + \exp(\beta_0 + \beta_1 \cdot u)}.$$

The parameter q as well as the parameters γ_1 , β_1 , and α_1 are sensitivity parameters. It is not the goal to estimate these parameters from the data set. Instead the researcher can postulate plausible ranges for them, gained from the literature or previous analyses. Conditional on the specified values for $(q, \gamma_1, \alpha_1, \beta_1)$, the remaining parameters $(\gamma_0, \alpha_0, \beta_0)$ can be estimated through maximum likelihood. Once all parameters are estimated, standard statistics such as the average treatment effect τ can be calculated accounting for the unobserved confounder U (see for example Imbens and Rubin 2015, 502 for details). Imbens (2003, 128) notes that the sensitivity parameters α_1 , β_1 , and γ_1 are difficult to interpret directly as they refer to log odds ratios. Exponentiating simplifies the interpretation, e.g., when $e^{\alpha_1} = 2$ the odds for $Y = 1$ in the treatment group double under $u = 1$ compared to $u = 0$.

4.4.2. The *TippingSens* app as a visualization tool

The sensitivity analysis approach developed by Rosenbaum and Rubin (1983) is an excellent way to examine specific violations of the unconfoundedness assumption. However, in most practical settings, the true values for the different parameters are unknown, and researchers are interested in evaluating the robustness of their findings under a whole range of plausible assumptions regarding the associations between the unobserved variable and the outcome or the treatment assignment. The number of simulated possible treatment effects rises quickly with the number of chosen values for the four different

4. *TippingSens R Shiny Application*

parameters and examination of results through tables becomes very cumbersome.

The interactive R shiny app called *TippingSens*, which we introduce in this chapter greatly simplifies conducting sensitivity analysis in this situation. It visualizes the impacts of various assumptions regarding the unknown parameters on the estimated average treatment effect. Thus, the app can be a handy tool to evaluate under which conditions the analytic conclusions would change. Snapshots of the visualization can be downloaded and integrated into any research output to offer easily comprehensible robustness checks of the research findings.

With the *TippingSens* app, it is possible to specify ranges for all parameters. The app requires a two-column matrix containing the binary treatment indicator value and the binary outcome of interest value for all units. Note that the app assumes that the data are already balanced regarding the observed covariates, that is, any steps for achieving balance, such as matching or trimming, need to be conducted before using the app. Once the data have been loaded, one can freely choose which two of the four parameters should be treated as fixed and specify the values for these parameters. Drop-down menus and sliders allow changing the settings interactively. The drop-down menus specify which parameters should be treated as fixed, and which parameters should be displayed on which axis. The sliders are a convenient tool for adjusting the values for the fixed parameters as well as for specifying the range of values considered for the free parameters. We acknowledge that it would also be possible to fix only one parameter and present the results in a three-dimensional plot. However, we feel that this would sacrifice the intuitiveness of the visualization. The design of the output grid was inspired by the tipping point analysis of Liublinska and Rubin (2014) for missing data sensitivity. The Appendix also contains a step-by-step illustration on how to invoke the app.

We will demonstrate the use of the *TippingSens* App in the next section with an example. The app is available at <https://tippingsens.shinyapps.io/TSApp/>, and the code be-

hence the app can be accessed at <https://github.com/CaroHaensch/TippingSensApp>.

4.5. A practical example: Sensitivity analysis for a quasi-experimental evaluation of a German vocational training program for the unemployed

To illustrate the usefulness of the TippingSens app, we use a quasi-experimental evaluation of vocational training for unemployed in Germany conducted by Bernhard (2016). Another example based on the data used in Rosenbaum and Rubin (1983) is given in the Appendix. Quasi-experimental labor market program evaluations like the one in Bernhard (2016) typically include socio-demographics, information on labor market histories, and regional characteristics to control for the selectivity regarding the assignment to the treatment (see also Lechner 1999 or Heckman et al. 1998). But they typically cannot incorporate personality traits, skills, preferences, attitudes, or social networks since this information is not available. However, these variables can still influence the job prospects (Bayer et al. 2008; Heckman et al. 2006; Mueller and Plug 2006; Pannenberg 2010). They may also be a key driver of selection into training (Heckman et al. 1997). Estimates of causal effects using only the observed data can be biased in this situation since the unobserved variables are correlated with both, the outcome of interest as well as the treatment assignment. Thus, sensitivity analyses should be conducted to evaluate how strong these correlations need to be to change the research findings.

4.5.1. Study details

The study of Bernhard (2016) uses the Integrated Employment Biographies (IEB) of the Institute for Employment Research in Germany (Dorner et al. 2010). The IEB is a large administrative database integrating five different sources of information

4. *TippingSens R Shiny Application*

collected by the Federal Employment Agency in Germany through different administrative procedures: the Employment History, the Benefit Recipient History, the Participants-in-Measures History, the Unemployment Benefit II Recipient History, and the Jobseeker History. It contains socio-demographic characteristics and individual daily information on employment, unemployment, benefit receipt, and participation in programs of active labor market policy for the universe of German employees and unemployed.

Based on these data, the treatment group was defined as the total inflow of unemployed welfare recipients into vocational training within a three-months-period in 2005. The control group consisted of a 20% random sample of unemployed welfare recipients a day before this three-months-period. The controls did not enter vocational training within this three-months-period, but they could start vocational training afterward to avoid conditioning on future events (Fredriksson and Johansson 2008).

The following observable information was used to model the assignment into treatment: individual socio-demographic characteristics (age, migration background, disability, qualification), characteristics of the household (single/partner, children), individual labor market history over the last five years (e.g., duration of employment, characteristics of the previous job such as wage, full-/part-time position, time in unemployment since last employment), labor market history of the partner, and local labor market characteristics (Rüb and Werner 2007). The data were stratified by several socio-demographic characteristics such as gender, age group, migration, and background. Separate propensity score models for the treatment assignment were estimated for each stratum. Within each stratum, the final model was obtained using a stepwise selection procedure. Caliper matching based on the estimated propensity scores was used to get the final data set (see Bernhard 2016 for further details on the matching procedure).

For more than eight years after the (hypothetic) start of the training, labor market outcomes of participants, and the matched control group were compared on three

4. *TippingSens R Shiny Application*

dimensions: share of welfare recipients, share of employees, and average monthly real wage. As common with matching approaches for causal inference, the study estimated the average treatment effect on the treated (ATET) and not the average treatment effect (ATE). Focusing on the ATET was useful in this context since it allowed to evaluate the effects on those for whom the program was intended (Heckman et al. 1997). However, the switch from ATE to ATET affects the interpretation of the sensitivity parameters as we will discuss below.

The overall results of Bernhard (2016) closely resemble previous findings not only for German data but also in an international context: The beginning of vocational training is an investment phase. During training, the search intensity for new jobs decreases, and employment prospects and wages of training participants are lower in comparison to non-participants. This fact is known as Ashenfelter's dip (Ashenfelter 1978). However, after a few months, positive effects of training on employment prospects, wages, and further welfare receipts can be observed, and these effects persist for up to eight and a half years after the training started. These results are in line with other quasi-experimental evaluations that find positive impacts of training on employment outcomes (Card et al. 2010).

In our illustrative application of the TippingSens app, we will focus on men in West Germany. Our outcome of interest will be the employment status (employed vs. unemployed) over time. For the subgroup of men in West Germany, this outcome follows the overall trend described above. Assuming no confounding, unemployed West German men have a nine percentage points higher chance to be employed two years after starting a vocational training compared to West German men that did not participate in the training program.

4.5.2. Sensitivity analysis with the TippingSens application

Specifying the required parameters and information for the TippingSens app is straightforward. Assuming the matching procedures described above resulted in a well-balanced data set regarding the observed characteristics, we only need two vectors from the matched data: the binary treatment indicator and the binary outcome indicator (see the Appendix for details regarding how to set up the app). In our case, the percentage of men in Western Germany that have found a job two years after starting vocational training is 0.46 for the treatment group (those who participated in vocational training) and 0.37 for the control group (those who did not participate in vocational training).

We conduct a Rosenbaum-Rubin sensitivity analysis because we are concerned about bias from unobservables. When evaluating vocational programs, such unobserved variables could be health conditions like alcoholism, obesity, or other factors. These variables are not recorded in the administrative data, but since the placement officer at the labor market agency might observe some of them, they might not only correlate with future job perspectives but might also affect the probability of attending vocational training. We will concentrate on alcoholism as an example here.

We need to think about four sensitivity parameters, i.e., we have to think about the association between alcoholism and participation in the training program (γ), between alcoholism and employment after two years in the treatment group (α) and in the control (β) group, and we need an estimate for the prevalence of heavy alcohol drinking (q).

As mentioned previously, we do not estimate or extract the sensitivity parameters from the data. Instead, we use other sources of information to narrow down the range of plausible values for the parameters. A literature review reveals that a majority of studies suggest that heavy alcohol consumption has negative effects on employment probabilities in Western industrialized states (Popovici and French 2016; Mullahy and Sindelar 1996). A reanalysis by Terza (2002) controlling for endogeneity found a strong negative effect

4. *TippingSens R Shiny Application*

on the probability of being employed in the US. MacDonald and Shields (2004) used data from the Health Survey of England and found heavy alcohol drinking negatively associated with the likelihood of employment, Johansson et al. (2007) obtained similar results for Finland. Devaux and Sassi (2015) estimate that heavy drinking has a strong negative effect on employment in men (white-collar men, OR: 0.54 [0.29; 1.00]). We will take these findings as a reference and use it for the sensitivity parameters α_1 and β_1 . We assume that both parameters have an upper bound of zero since based on the literature it seems implausible that alcoholism will have a positive effect on job perspectives irrespective of whether a unit belongs to the treatment or the control group. Next, we need to define a plausible range for the parameter q . If we were interested in the ATE, the parameter would represent the prevalence of alcoholism among men in West Germany. However, given that the study focuses on the ATET, the interpretation of q changes. It is now the prevalence of alcoholism among individuals with observable characteristics similar to those of the treatment group. Given that the treatment group might be a selective subset of the population (for example, all of them are unemployed at the start of the training program), it is more difficult to narrow down the range of plausible values for q . To specify an appropriate range for q , we use a literature review of unemployment and substance use by Henkel (2011). He puts the percentage at 14% for unemployed German men, but estimates vary across studies and measurements. We will examine a range of possible values from 2.0% to 18%.

Defining a suitable range for γ_1 is most difficult since there are no studies – at least to our knowledge – that investigate how alcohol abuse influences the probability of being assigned to a labor market program. We will thus evaluate a wide range of values for γ_1 . To visualize the results from the Rosenbaum-Rubin sensitivity analysis, the app requires to fix two of the parameters at specific values, while plausible ranges can be specified for the remaining parameters. It is up to the user to decide which of the four parameters to

4. *TippingSens R Shiny Application*

fix.

We start our evaluation with the setting that appears to be most plausible concerning the literature. We set $q = 0.14$ and $\alpha_1 = \log(0.54) = -0.62$. We note that by fixing α_1 at -0.62 we implicitly assume that the odds ratio of getting a job for heavy drinkers for the treated is the same as for the general population of unemployed German men. We will evaluate the impacts of loosening this assumption below. Given that we expect negative effects of alcoholism on employment perspectives, and that it seems prudent to assume that alcoholism also has negative effects on the probability of receiving the treatment, we can fix the upper bound for the range of plausible values for the other parameters at zero. Defining meaningful lower bounds is more difficult. In our illustrative application, we set the lower bound for γ_1 to -3 , implying that we believe that the odds ratio of being assigned to the training program will not be less than 0.05 for individuals with a drinking problem relative to individuals without drinking problems.

For the parameter β_1 , we set the lower bound to -1.31 . We choose this value, as we assume that the odds ratio of getting a job for individuals with a drinking problem in the control group will never be less than half the odds ratio for those participating in the training program (solving $\exp^{\beta_1} / \exp^{\alpha_1} \geq 0.5$ for β_1 gives $\beta_1 \geq -1.31$). The output of the app based on our data and the parameter settings are depicted in Figure 4.1. Lighter colored tiles imply smaller treatment effects. When assuming that the odds ratio of employment for heavy drinkers is about 0.75 ($\beta_1 = -0.29$) in the control group, and that the odds ratio for treatment assignment given alcoholism is about 0.72 ($\gamma_1 = -0.33$), while the prevalence rate of alcoholism among individuals with observable characteristics similar to those in the treatment group is 0.14, the treatment effect of vocational training drops from 0.09 in the analysis based on the unconfoundedness assumption to 0.086 (black box in Figure 4.1). The stronger the negative effect of alcoholism on the outcome for the matched controls and on the treatment assignment, the more important becomes

4. TippingSens R Shiny Application

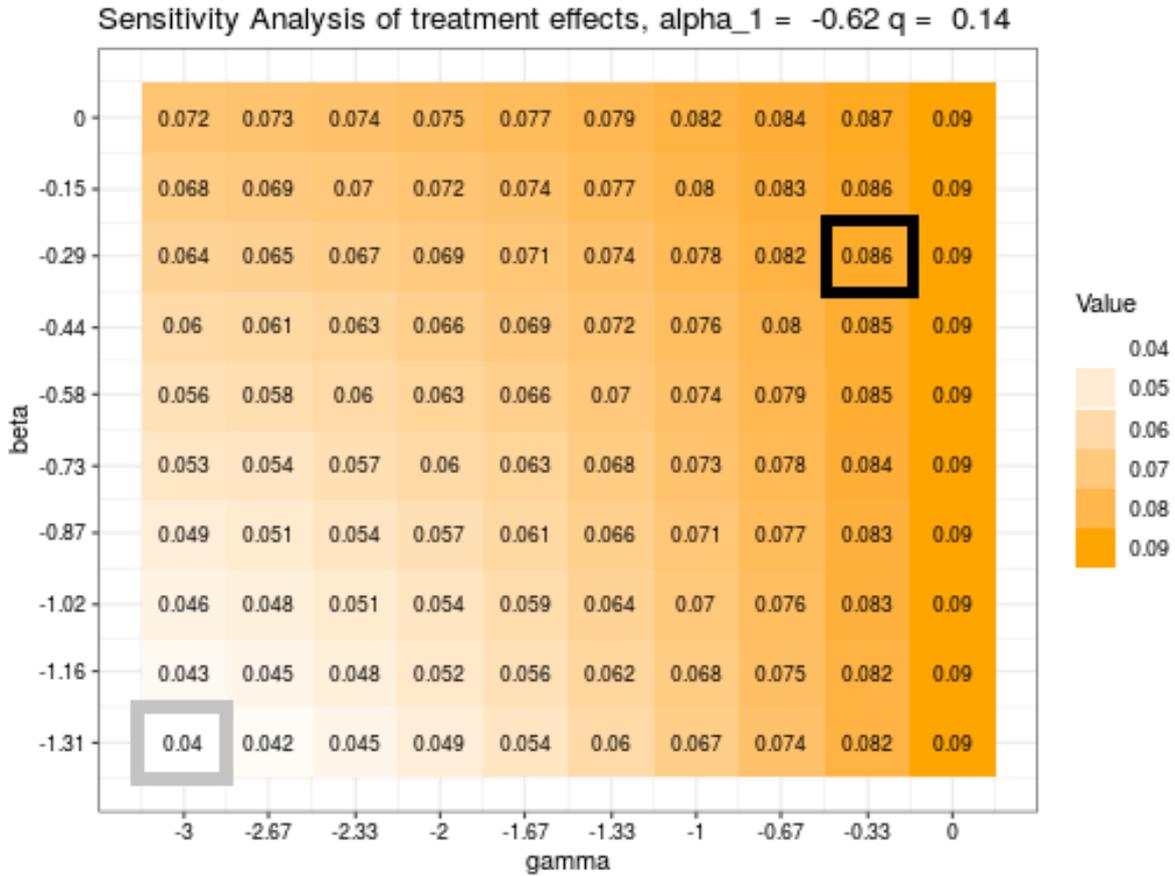


Figure (4.1) Sensitivity analysis for the evaluation of vocational training by Bernhard (2016). The gray box contains the treatment effect assuming the following values for the sensitivity parameters: $\alpha_1 = -0.62$, $\beta_1 = -1.31$, $\gamma_1 = -3$, $q = 0.14$. The black box contains the treatment effect assuming the following values for the sensitivity parameters: $\alpha_1 = -0.62$, $\beta_1 = -0.29$, $\gamma_1 = -0.33$, $q = 0.14$.

the unobserved covariate in explaining differences between treatment and control group leading to smaller estimated treatment effects. We note that even under the most extreme setting considered in Figure 4.1 ($\gamma = -3$, $\beta = -1.31$) the estimated treatment effect remains positive (gray box in Figure 4.1). Thus, the results of the analysis based on the unconfoundedness assumption seems to be quite stable.

With the TippingSens plot, we can also examine the estimated treatment effects in other settings. We can, for example, take a closer look at the effects of different percentages of

4. TippingSens R Shiny Application

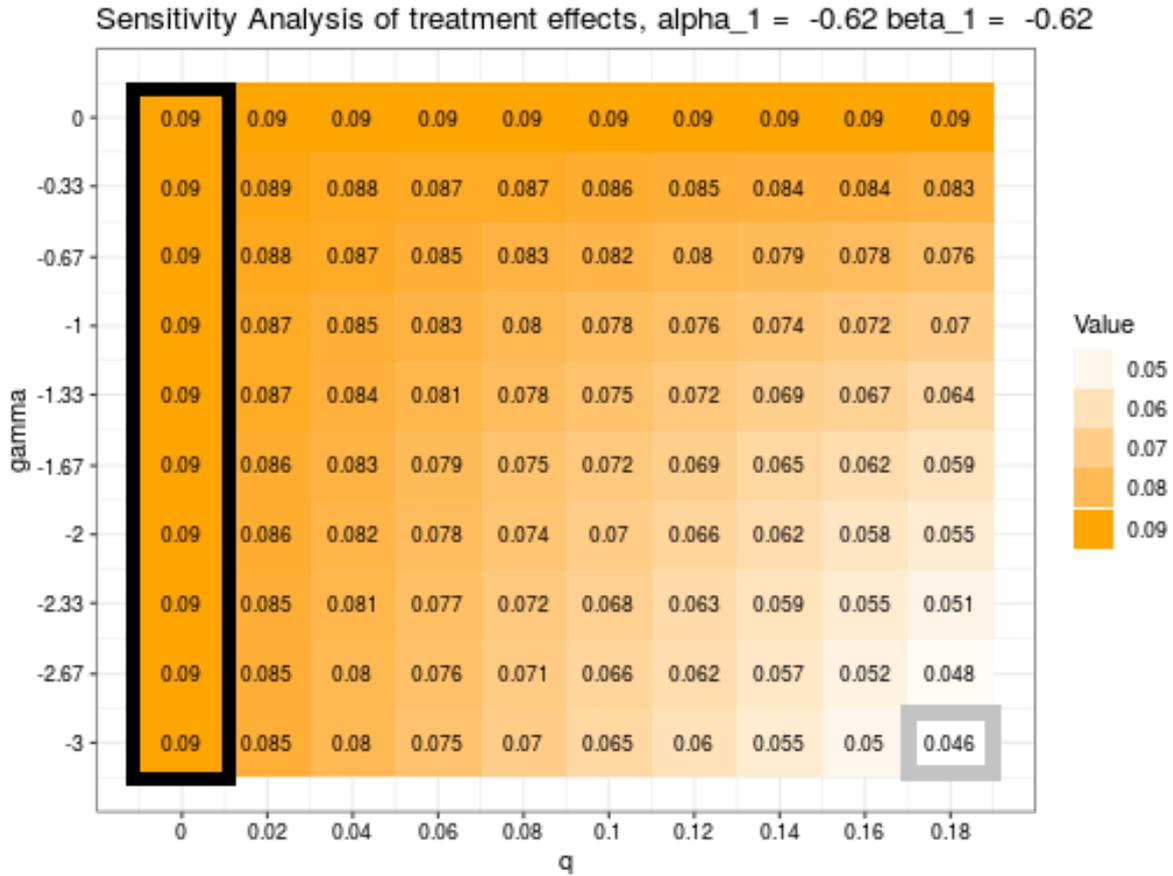


Figure (4.2) Sensitivity analysis for the evaluation of vocational training by Bernhard (2016). The black box contains the treatment effect assuming the following values for the sensitivity parameters: $\alpha_1 = \beta_1 = -0.62$, $\gamma_1 \in [-3, 0]$, $q = 0$. The gray box contains the treatment effect assuming the following values for the sensitivity parameters: $\alpha_1 = \beta_1 = -0.62$, $\gamma_1 = -3$, $q = 0.18$.

individuals with a drinking problem in our subpopulation of interest. We switch axes in the app and receive a plot with γ on the vertical axis and q on the horizontal axis. We fix α_1 and β_1 at -0.62 implicitly assuming that the odds ratio of employment given alcohol abuse is the same in both treatment groups. The results are depicted in Figure 4.2. If we assume there are no men with drinking problems, the unobserved covariate should have no effect. This is confirmed in the first column of Figure 4.2 (where $q = 0$), showing a constant estimated treatment effect of 0.09, which is the estimated treatment effect based on the unconfoundedness assumption (black box in Figure 4.2). Moving

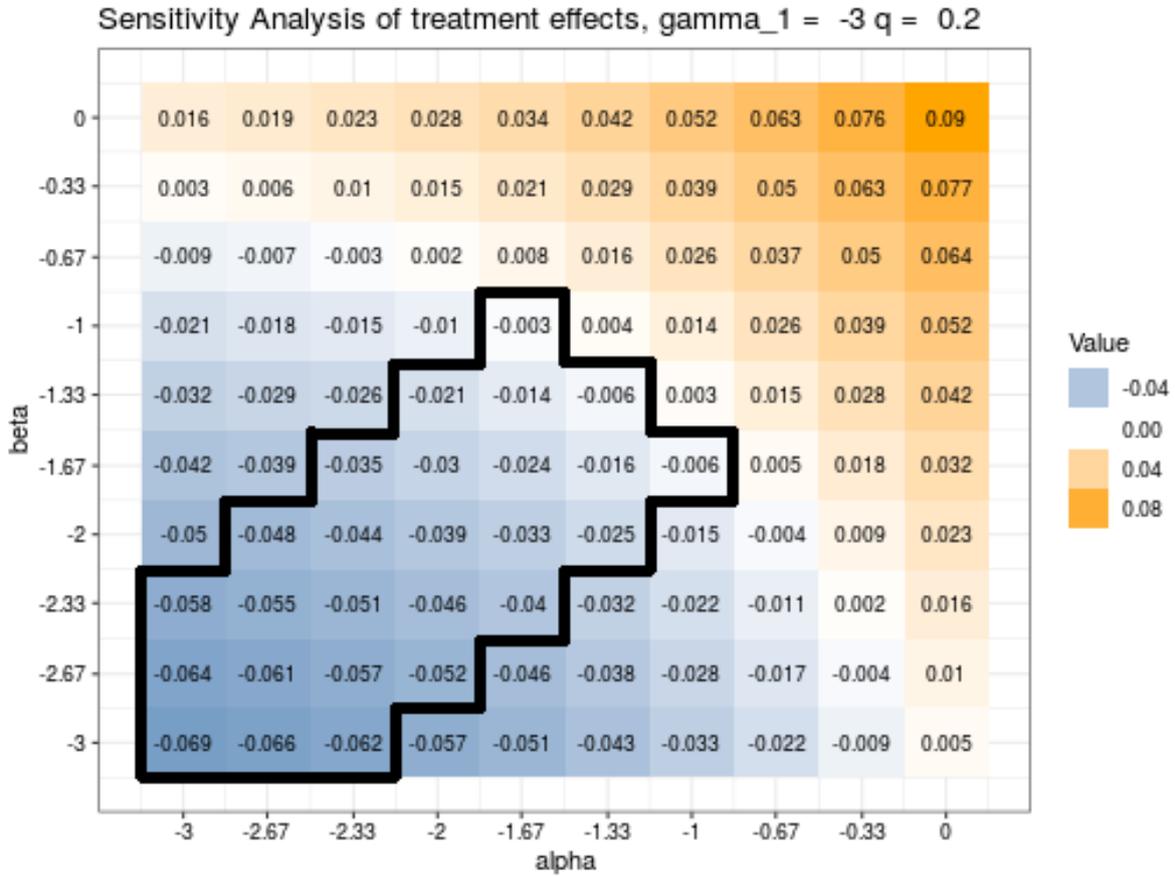


Figure (4.3) Sensitivity Analysis of the evaluation of vocational training by Bernhard (2016). Odds ratios regarding alcoholism do not differ by more than a factor of two between the treatment and the matched control group in the area highlighted by black lines.

to higher percentages for q we can see that the treatment effect diminishes. When q is close to zero (or one), treatment and control groups are still well balanced regarding the unobserved variable q , just because there are not enough subjects with $u = 1$ (or $u = 0$) that can be selected into either group. Thus, the effects of changing γ are small for small values of q . This changes as q increases. More variation in the outcome is now attributed to the unobserved covariate. Still, the treatment effect does not drop below 0.046 for the considered settings (gray box in Figure 4.2).

We can also explore which conditions would be necessary to alter the sign of the estimated treatment effect based on the unconfoundedness assumption. Imbens and Rubin (2015,

506) remark that there is a trade-off between the different parameter: more extreme values for γ_1 do not require quite as extreme value for α_1 and β_1 to see the same shift in the estimated treatment effect. We thus choose rather extreme values for the fixed parameters. In Figure 4.3, we fix q at 20% and γ_1 at -3, which means that the odds for treatment are about twenty times higher for German men without drinking problems compared to subjects with drinking problems. But even with such a strong selection mechanism, the estimated treatment effect only changes its sign if α_1 and β_1 get very small (blue tiles indicate negative treatment effects). If we also take into account that it seems unrealistic that the odds ratios differ by more than a factor of two between the treatment and the matched control group, the range of admissible values for the combination of the two parameters is limited further. Admissible value combinations that would lead to a negative treatment effect are highlighted by black lines in Figure 4.3. The figure shows that the values for both parameters would need to be smaller than -1.33 (odds ratio less than 0.26) to change the sign of the estimated treatment effect. This is a much smaller odds ratio than the odds ratio of 0.54 found by Devaux and Sassi (2015). Thus, to obtain a negative treatment effect, we would need to assume that the target population for the training program is a highly selective group, for which the odds ratio of finding a job for heavy drinkers is much smaller, and for which the prevalence of alcoholism is much higher than in the reference population. We can conclude that the findings in Bernhard (2016) are robust to the assumption that alcoholism is an unobserved confounder, which influences both the probability of assignment to the training program and the probability of being employed two years after baseline.

4.6. Discussion

The sensitivity analysis proposed by Rosenbaum and Rubin (1983) allows flexible modeling of the violation of the unconfoundedness assumption through four different parameters.

4. *TippingSens R Shiny Application*

In their initial illustrative application, Rosenbaum and Rubin worked with forking tables to convey their idea and their results. The table format forced the authors to evaluate only a limited set of combinations for the parameters. With the R Shiny app introduced in this chapter, we simplify the sensitivity analysis by creating interactive visualizations instead. Since it is typically unrealistic to assume that exact values are known for all four sensitivity parameters, the possibility of specifying ranges of plausible values for the parameters is a major advantage of the app compared to previously proposed solutions, such as using the most extreme relationships found in the observed data. With the app, the estimated treatment effects are visualized over the two-dimensional space spanned by the range of plausible values for the free parameters. Color coding helps to identify the relationships between the different parameters and the treatment effect. It is up to the user to decide which parameters should be held fixed, and which intervals should be considered for the free parameters. Once the data are loaded, all settings can be changed easily: parameters can be exchanged between the axes or from being fixed to being free, and vice versa, ranges of plausible values can be adjusted independently, and the values of the fixed parameters can be modified using separate sliders for each of the parameters (see also the step-by-step illustration in the Appendix). The interactive flexibility of the app also provides quick insights which (possibly extreme) sets of parameter combinations would be required to substantially change the research findings derived under the unconfoundedness assumption.

The illustrative application based on a quasi-experimental evaluation of vocational training in Germany discussed in Section 4.5 highlights the benefits of the TippingSens app. Evaluating different plausible scenarios regarding the association between the confounder and the outcome and the treatment, we found that the substantial findings in Bernhard (2016) are robust regarding the effects of alcoholism as a possible confounder. Of course, the evaluation has important limitations. It only focuses on one confounder.

4. *TippingSens R Shiny Application*

Other health problems might also have negative effects on both the outcome and the treatment assignment. Thus, it would not be appropriate to conclude based on this limited sensitivity analysis that the findings in Bernhard (2016) are robust to any form of confounding. However, the application illustrates the basic idea. Similar sensitivity analyses could be conducted for other health parameters in practice.

Still, it must be noted that the applicability of the app is limited by the requirements of the Rosenbaum-Rubin sensitivity analysis: both, the outcome, as well as the unobserved confounder, have to be univariate and binary. While it has been argued that this framework might still apply in contexts with more than one confounder (Liu et al. 2013), the assumption of a binary confounder might be too restrictive in other contexts. For example, with continuous confounders, it seems more realistic to assume that the effect of the confounder changes (non-)linearly with the value of the confounder instead of assuming only a single change in the effect at a certain threshold value. Whether the app could be extended to allow obtaining useful insights in this more general context beyond the Rosenbaum-Rubin approach would be an interesting area for future research.

Bibliography

- Ashenfelter, O. (1978). Estimating the Effect of Training Programs on Earnings. *Review of Economics and Statistics*, 6(1):47–57.
- Bayer, P., Ross, S., and Topa, G. (2008). Place of Work and Place of Residence: Informal Hiring Networks and Labor Market Outcomes. *Journal of Political Economy*, 116(6):1150–1196.
- Becker, S. and Caliendo, M. (2007). Sensitivity Analysis for Average Treatment Effects. *Stata Journal*, 7(1):71–83.
- Bernhard, S. (2016). Berufliche Weiterbildung von Arbeitslosengeld-II-Empfängern. Langfristige Wirkungsanalysen. *Sozialer Fortschritt*, 65(7):153–161.
- Blundell, R., Gosling, A., Ichimura, H., and Meghir, C. (2007). Changes in the Distribution of Male and Female Wages Accounting for Employment Composition Using Bounds. *Econometrica*, 75(2):323–363.
- Card, D., Kluve, J., and Weber, A. (2010). Active Labour Market Policy Evaluations: A Meta-analysis. *The Economic Journal*, 120:F452–F477.
- Cornfield, J., Lilienfeld, A. M., Hammond, E. C., Wynder, E. L., Shimkin, M. B., and Haenszel, W. (1959). Smoking and Lung Cancer: Recent Evidence and a Discussion of some Questions. *JNCI: Journal of the National Cancer Institute*, 22(1):173–203.

Bibliography

- Devaux, M. and Sassi, F. (2015). The Labour Market Impacts of Obesity, Smoking, Alcohol Use and Related Chronic Diseases. *OECD Health Working Papers*, 86:1–50.
- Dorner, M., Heining, J., Jacobebbinghaus, P., and Seth, S. (2010). The Sample of Integrated Labour Market Biographies. *Schmollers Jahrbuch*, 130(4):599–608.
- Fredriksson, P. and Johansson, P. (2008). Dynamic Treatment Assignment. *Journal of Business and Economic Statistics*, 26(4):435–445.
- Gangl, M. (2004). RBOUNDS: Stata Module to Perform Rosenbaum Sensitivity Analysis for Average Treatment Effects on the Treated. Statistical Software Components, Boston College Department of Economics.
- Gastwirth, J. L., Krieger, A. M., and Rosenbaum, P. R. (1998). Dual and Simultaneous Sensitivity Analysis for Matched Pairs. *Biometrika*, 85(4):907–920.
- Greenland, S. (1996). Basic Methods for Sensitivity Analysis of Biases. *International Journal of Epidemiology*, 25(6):1107–1116.
- Harding, D. J. (2003). Counterfactual Models of Neighborhood Effects: The Effect of Neighborhood Poverty on Dropping Out and Teenage Pregnancy. *American Journal of Sociology*, 109(3):676–719.
- Heckman, J., Ichimura, H., Smith, J., and Todd, P. (1998). Characterizing Selection Bias Using Experimental Data. *Econometrica*, 66(5):1017–1098.
- Heckman, J., Ichimura, H., and Todd, P. (1997). Matching as an Economic Evaluation Estimator: Evidence From Evaluating a Job Training Programme. *The Review of Economic Studies*, 64(4):605–654.
- Heckman, J., Stixrud, J., and Urzua, S. (2006). The Effects of Cognitive and Noncognitive

Bibliography

- Abilities on Labor Market Outcomes and Social Behavior. *Journal of Labor Economics*, 24(3):411–482.
- Heckman, J. J. (1989). Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training. Working Paper 2861, National Bureau of Economic Research.
- Henkel, D. (2011). Unemployment and Substance Use: A Review of the Literature (1990-2010). *Current Drug Abuse Reviews*, 4(1):4–27.
- Hof, S. (2014). Does Private Tutoring Work? The Effectiveness of Private Tutoring: A Nonparametric Bounds Analysis. *Education Economics*, 22(4):347–366.
- Holland, P. W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association*, 81(396):945–960.
- Ichino, A., Mealli, F., and Nannicini, T. (2008). From Temporary Help Jobs to Permanent Employment: What Can We Learn From Matching Estimators and Their Sensitivity? *Journal of Applied Econometrics*, 23(3):305–327.
- Imbens, G. W. (2003). Sensitivity to Exogeneity Assumptions in Program Evaluation. *American Economic Review*, 93(2):126–132.
- Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, chapter 22, pages 496—510. Cambridge University Press, Cambridge.
- Johansson, E., Alho, H., Kiiskinen, U., and Poikolainen, K. (2007). The Association of Alcohol Dependency With Employment Probability: Evidence From the Population Survey 'Health 2000 in Finland'. *Health Economics*, 16(7):739–754.

Bibliography

- Kang, C. (2011). Family Size and Educational Investments in Children: Evidence From Private Tutoring Expenditures in South Korea. *Oxford Bulletin of Economics and Statistics*, 73(1):59–78.
- Keele, L. (2010). Rbounds: An R Package for Sensitivity Analysis with Matched Data. R package.
- Kitahata et al., M. M. (2009). Effect of Early versus Deferred Antiretroviral Therapy for HIV on Survival. *New England Journal of Medicine*, 360(18):1815–1826.
- Lechner, M. (1999). Earnings and Employment Effects of Continuous Off-the-Job Training in East Germany After Unification. *Journal of Business and Economic Statistics*, 17(1):74–90.
- Lechner, M. and Wunsch, C. (2013). Sensitivity of Matching-based Program Evaluations to the Availability of Control Variables. *Labour Economics*, 21(C):111–121.
- Liu, W., Kuramoto, S. J., and Stuart, E. A. (2013). An Introduction to Sensitivity Analysis for Unobserved Confounding in Nonexperimental Prevention Research. *Prevention Science*, 14(6):570–580.
- Liublinska, V. and Rubin, D. B. (2014). Sensitivity Analysis for a Partially Missing Binary Outcome in a Two-Arm Randomized Clinical Trial. *Statistics in Medicine*, 33(24):4170–4185.
- MacDonald, Z. and Shields, M. A. (2004). Does Problem Drinking Affect Employment? Evidence From England. *Health Economics*, 13(2):139–155.
- Manski, C. F. (1990). Nonparametric Bounds on Treatment Effects. *The American Economic Review*, 80(2):319–323.

Bibliography

- Manski, C. F. and Pepper, J. V. (2000). Monotone Instrumental Variables: With an Application to the Returns to Schooling. *Econometrica*, 68(4):997–1010.
- Manski, C. F. and Pepper, J. V. (2009). More on Monotone Instrumental Variables. *The Econometrics Journal*, 12(S1):200–216.
- Mueller, G. and Plug, E. (2006). Estimating the Effect of Personality on Male and Female Earnings. *Industrial Labor Relations Review*, 60(1):3–22.
- Mullahy, J. and Sindelar, J. (1996). Employment, Unemployment, and Problem Drinking. *Journal of Health Economics*, 15(4):409–434.
- Pannenberg, M. (2010). Risk Attitudes and Reservation Wages of Unemployed Workers: Evidence From Panel Data. *Econometric Letters*, 106(3):223–226.
- Pearl, J., Glymour, M., and Jewell, N. P. (2016). *Causal Inference in Statistics: A Primer*. John Wiley & Sons.
- Popovici, I. and French, M. T. (2016). Substance Use and School and Occupational Performances. In Sher, K. J., editor, *The Oxford Handbook of Substance Use and Substance Use Disorders: Volume 2*. Oxford University Press, Oxford.
- Rosenbaum, P. R. (1995). *Observational Studies*. Springer New York.
- Rosenbaum, P. R. (1999). Reduced Sensitivity to Hidden Bias at Upper Quantiles in Observational Studies with Dilated Treatment Effects. *Biometrics*, 55(2):560–564.
- Rosenbaum, P. R. (2003). Does a Dose-Response Relationship Reduce Sensitivity to Hidden Bias? *Biostatistics*, 4(1):1–10.
- Rosenbaum, P. R. (2007). Sensitivity Analysis for m-Estimates, Tests, and Confidence Intervals in Matched Observational Studies. *Biometrics*, 63(2):456–464.

Bibliography

- Rosenbaum, P. R. (2010). *Design of Observational Studies*, chapter 14, pages 257–274. Springer, New York.
- Rosenbaum, P. R. (2014). Two R Packages for Sensitivity Analysis in Observational Studies. *Observational Studies*, 1(1):1–17.
- Rosenbaum, P. R. (2018). Sensitivity Analysis for Stratified Comparisons in an Observational Study of the Effect of Smoking on Homocysteine Levels. *Annals of Applied Statistics*, 12(4):2312–2334.
- Rosenbaum, P. R. and Rubin, D. B. (1983). Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study With Binary Outcome. *Journal of the Royal Statistical Society. Series B (Methodological)*, 45(2):212–218.
- Rubin, D. B. (1974). Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology*, 66(5):688–701.
- Rüb, F. and Werner, D. (2007). Typisierung von SGB II-Trägern. *Institute for Employment Research. Research Report*, (1/2007):1–35.
- Smith, J. A. and Todd, P. E. (2001). Reconciling Conflicting Evidence on the Performance of Propensity-Score Matching Methods. *American Economic Review*, 91(2):112–118.
- Terza, J. V. (2002). Alcohol Abuse and Employment: A Second Look. *Journal of Applied Econometrics*, 17(4):393–404.
- VanderWeele, T. J. and Arah, O. A. (2011). Bias Formulas for Sensitivity Analysis of Unmeasured Confounding for General Outcomes, Treatments, and Confounders. *Epidemiology*, 22(1):42–52.
- Zubizarreta, J. R., Cerdá, M., and Rosenbaum, P. R. (2013). Effect of the 2010 Chilean Earthquake on Posttraumatic Stress. *Epidemiology*, 24(1):79–87.

5. Better Together? Regression

Analysis of Complex Survey Data

After Ex-post Survey Harmonization

5.1. Introduction

What can researchers do if information from a single data source is insufficient to investigate a particular research question? An increasing number of researchers then pool, harmonize, and analyze survey data from different survey providers for their research questions (SDR 2020; IPUMS 2020; CLOSER 2020; MAELSTROM 2020; MTUS 2020). They are then able to study heterogeneity between groups over a long period, pick up subtle differences between populations, or examine smaller subgroups.

In this context, harmonization refers to procedures aimed at improving the comparability of different surveys and measures (Granda et al. 2010, 315). Both ex-ante and ex-post harmonization have a strong tradition in cross-cultural surveys. *Ex-ante* harmonization refers to harmonization during survey design and before data collection. *Ex-post* harmonization to harmonization after data collection. While the harmonization of variables is already a complex task (Granda et al. 2010, 331), the analysis of *complex* survey data after variable harmonization is not necessarily easier.

5. Better Together?

Surveys in the social sciences often rely on complex sampling designs, i.e., sample members do not have an equal selection probability. They are widely used in the social sciences, and ex-post harmonization projects will most likely also include surveys with complex sampling designs. Ignoring the complex sampling design can lead to biased population inferences in population means as well as regression coefficients (DuMouchel and Duncan 1983; Pfeffermann and Sverchkov 2009; Solon et al. 2013). For regression models that are widely used in quantitative sociological research, analysts often try to account for the sampling design through additional predictors. With ex-post survey harmonization, this often becomes impossible since it could require the analyst to harmonize many other predictors. Besides, these predictors might not be observed for several of the surveys, creating severe missing data problems. Another possible limitation would be that the variables related to the sampling design are not suitable for the analysis due to their associations with focus variables (Lumley 2010, 105).

There is, however, a second option to account for the complex sampling design: it is possible to estimate a survey-weighted regression. In this article, we explore this second option in the context of harmonized complex survey data and compare *different approaches for regression analyses of survey data after ex-post harmonization and how to incorporate survey weights and survey weighting*.

The idea of combining or pooling data has also surfaced in areas other than sociology and cross-cultural studies. The terms to describe the general idea of combining data differ widely between disciplines and even projects, terms used include, for example, individual person data (IPD) meta-analysis (MA) in medicine and psychology mainly (Riley et al. 2010; Burke et al. 2017), mega-analysis (Boedhoe et al. 2019) or just merely describing the strategy as “pooling raw data” (Korn and Graubard 1999). The IPD MA literature, in particular, proposes two general approaches for the regression analysis of pooled raw data: (1) synthesizing regression coefficients estimated from the single

data sets (two-stage approach) or (2) estimating a regression on the combined data sets (one-stage approach). This article will study which method is more suited for the analysis of complex survey data after ex-post harmonization.

This article pioneers in bridging the gap between survey harmonization, survey statistics, and meta-analysis, making research results and approaches from one area useful to the other. In particular, we add to the literature in the following ways: (1) We introduce two approaches from the IPD-meta-analytical literature for dealing with ex-post harmonized complex survey data, i.e., one-stage and two-stage IPD meta-analysis (see Burke et al. 2017 for an overview). More importantly, we study the performance of these two approaches in the context of ex-post survey harmonization. (2) We can show that the distribution of survey weights in the respective data set will play an important role when applying these two meta-analytical approaches. We demonstrate that unless the coefficient of variance for the survey weights is small, the assumption of known within-study variances for two-stage analysis is problematic and can result in biased point estimates. (3) Finally, we add a real-world example further illustrating the findings of the simulation.

5.2. Regression analysis of complex survey-based data after harmonization

5.2.1. Overview and literature review

We will draw on two different analysis approaches developed for individual participant data (IPD) meta-analysis (MA) in the following, one-stage and two-stage meta-analysis. These distinct ways of conducting an IPD MA have emerged over the last few years. The more common approach, which is also closer to a more classical (aggregate person data) meta-analysis, is the so-called two-stage approach. Meta-analysts first analyze each of the $k = 1, \dots, K$ studies separately to obtain study-specific effect sizes. Then,

5. Better Together?

they combine the K independent effect sizes by calculating a weighted (often inverse error-variance-based) average. The second approach is the so-called one-stage IPD MA approach. Here the combined data is analyzed simultaneously. This flexible approach requires different modeling decisions like the possible inclusion of separate intercepts for different surveys or the addition of random intercepts or slopes. These one-stage analyses with random effects are examples of hierarchical or multi-level models (Simmonds et al. 2005).

Regarding the comparability of these two approaches, one-stage and two-stage approaches often lead to similar estimates of treatment effects (Olkin and Sampson 1998; Stewart et al. 2012). However, one-stage analyses improve the power of detecting effects for continuous and binary data (Lambert et al. 2002; Simmonds et al. 2005). Burke et al. (2017) give an overview of previous results and a list of ten key reasons why one-stage and two-stage approaches may lead to different results in practice. *We will add another reason to that list, connected to the variances of survey weights (see Section 5.3.2).*

Few authors have tackled the topic of combining complex survey data from an (ex-post) survey harmonization or (IPD) meta-analytical perspective. Roberts and Binder (2009) briefly discuss the two main meta-analytical approaches – one-stage and two-stage meta-analysis – in the context of survey data. They stress to either frame the research question in a design-based or model-based way. Korn and Graubard (1999) suggest methods to adjust survey weights when pooling surveys with different numbers of observations. Besides, survey methodologists and statisticians have combined data from surveys though they generally do not refer to it as meta-analysis. For instance, Leslie Kish focused on combining data from rolling or periodic samples and non-overlapping probability samples from the same population (Kish 1979; Kish and Verma 1986; Kish 1994, 1999).

We, however, use surveys that were not designed to be analyzed together in the first place. Prospectively planned pooled analyses such as extensive international cross-

cultural studies are usually not called meta-analyses, although they also are a method for summarizing the evidence. Joye et al. (2019) discuss the pooling and weighting of surveys in the context of harmonized cross-cultural studies.

Following Kish, statisticians have tackled different types and aspects of pooling data. While Cochran (2007) and Fuller and Burmeister (1972) concentrated on combining sub-populations, Kalton and Anderson (1986) and Skinner and Rao (1996) among others worked on the topic of multiple frames for the same target population. Lohr and Raghunathan (2017) and Fox (2011) provide overviews over other related topics like dual-frame problems, statistical matching, combining survey data with data from other surveys, small-area-estimation or re-weighting data with several samples.

5.2.2. Two-stage IPD meta-analytical approaches

As a next step, we take a closer look at the two analytical options for combining complex survey data after survey harmonization. We start with two-stage meta-analysis. Suppose each study k ($k = 1, \dots, K$), that we want to include into our analysis, has n_k observations i ($i = 1, \dots, n_k$). We are interested in the effect of a single independent variable on the dependent variable. A simple approach would be to estimate regression models in each study separately and then, in the second stage, to combine the estimates obtained in the first stage (Burke et al. 2017, 856).

If we follow this approach, we estimate regression coefficients for every survey. We can either estimate a weighted or an unweighted estimate (DuMouchel and Duncan 1983). After estimating the regression coefficients for the different surveys, we reach the second stage of the two-stage analysis (Burke et al. 2017, 857). In this stage, we combine the regression coefficients, similar to a classical meta-analysis. If we assume between-study homogeneity, each survey point estimate's weight depends solely on its estimate's variance (a so-called fixed-effect model).

5. Better Together?

In some cases, it may be reasonable to assume that a common effect exists (for example, medical studies using the same treatment for a very similar group of patients and using the same outcome measures) and estimating a fixed-effect (FE) model. However, such an assumption of homogeneity can rarely be made for most studies (Viechtbauer 2007), especially in the social sciences. Studies are likely to have systematic differences, e.g., the measurements used for the independent and dependent variable (Elliott et al. 2018, 3). We then speak of study heterogeneity. If study heterogeneity is present, random-effects (RE) models are considered more appropriate for the second stage of the analysis. The underlying assumption is that there is not one common effect for all the studies, but instead, one assumes a distribution of effects across studies.

When calculating the combined RE estimate, we again take an inverse-variance weighting approach, but we now incorporate an estimate of the between study-variance $\hat{\tau}^2$ (Borenstein et al. 2009, section 2). There are many methods of estimating $\hat{\tau}^2$, for example, the method of moments estimator of DerSimonian and Laird (1986, 2015), also called DL estimator. For continuous outcomes, Veroniki et al. (2016, 55) advocate using REML estimation as a preferable alternative to the DL estimator, which we also use in our simulations.

5.2.3. One-stage IPD meta-analytical approaches

When conducting a one-stage analysis, we first combine surveys and then analyze the combined data set. An abundance of modeling options characterizes the one-stage (IPD meta-analytical) approach. For all included predictors, separate fixed or random study effects can be included.

We start with the simplest model, a one-stage model without separate fixed or random study effects. Let us still assume we have K ($k = 1, \dots, K$) studies for the IPD MA with n_k ($i = 1, \dots, n_k$) observations each. We are interested in the effect of the independent

5. Better Together?

variable X on the dependent variable Y .

1. One-stage model without separate fixed study effects (linear model):

$$y_{ik} = \alpha + \beta x_{ik} + e_{ik}, e_{ik} \sim N(0, \sigma^2) \quad (5.1)$$

Here we treat the data as if they came from one extensive survey. We do not model study heterogeneity. However, it is generally not recommended to treat the combined studies as one big survey, ignoring the clustered structure. Researchers instead include fixed or random study effects to model the clustering.

2. One-stage model with separate fixed study effects:

Adding a separate fixed intercept term α_k per study allows for different survey means in the dependent variable (when controlling for the independent variable(s)) and the clustering of observations in studies (Burke et al. 2017, 859).

$$y_{ik} = \alpha_k + \beta x_{ik} + e_{ik}, e_{ik} \sim N(0, \sigma^2) \quad (5.2)$$

It is also possible to allow varying slopes for the surveys by adding interaction effects of survey indicators and independent variables.

3. One-stage model with one or more random study effects:

The separate fixed study effects for the intercept can also be replaced with a random effect, e.g., when one is interested in the baseline intercept α or in a measure of between-study heterogeneity (Burke et al. 2017, 859). Another advantage is that we reduce the number of parameters compared to the fixed effects model (instead of estimating a fixed effect for each survey, we only estimate two parameters α and τ_α^2). One should note that the inclusion of random effects demands a sufficiently

high number of included surveys (Snijders and Bosker 1993).

$$y_{ik} = \alpha_k + \beta x_{ik} + e_{ik}, e_{ik} \sim N(0, \sigma^2) \text{ and } \alpha_k \sim N(\alpha, \tau_\alpha^2) \quad (5.3)$$

As in the fixed effect case, it is possible to add random study effects for the independent variables, i.e., random *slopes*. The random slope model allows the explanatory variable to vary between studies. Readers should be aware that the inclusion of weights into multi-level models like the RE one-stage model here is still a research question that is not entirely resolved (see Asparouhov 2006 and Carle 2009).

5.2.4. Survey weights and the use of weights in regression analysis

Since we are studying the analysis of pooled complex data with survey weights, we now briefly introduce the three types of survey weights most widely used in practice. The weights are design weights, nonresponse weights, and post-stratification weights.

Design weights are common when observations are drawn dependent on the primary sampling unit (PSU). In the case of stratified random sampling, the entire population is divided into homogeneous groups, which are called strata s . If the stratum size in the sample is not proportional to the stratum size in the population, we speak of disproportionate sampling. In the case of disproportionate sampling, design weights for design-based inference are often provided and used. Calculation of the design weight for an observation i is simple enough; we use the inverse of the probability $p_{i,selection}$ to be selected (Horvitz and Thompson 1952 and Lumley 2010, 4): $w_i = (p_{i,selection})^{-1}$.

Nonresponse weights are closely related to sampling weights. Units are weighted by the inverse of their response propensities $p_{i,response}$ (Little and Vartivarian 2003). $p_{i,response}$ can be estimated if researchers have information about respondents and non-respondents as in the European Social Survey (Blom 2009, 29). Interviewers were asked to record

5. Better Together?

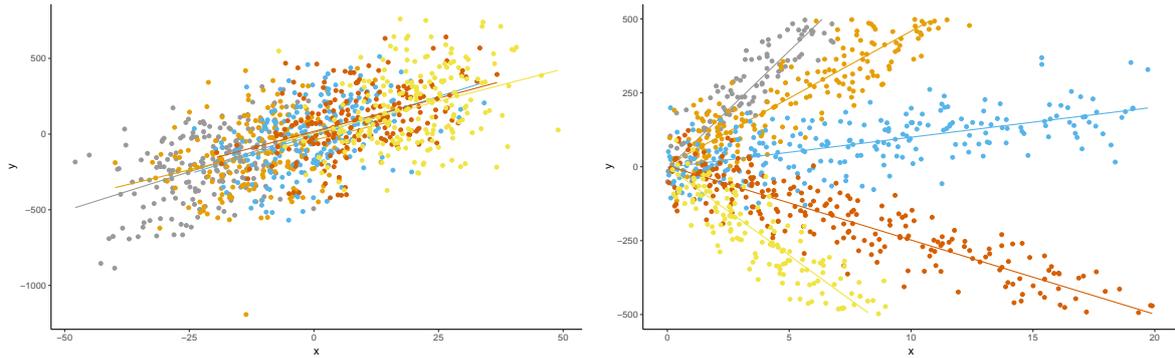
variables such as sex and approximate age, or the house's appearance (litter, graffiti, etc.) for respondents and non-respondents. This allows practitioners to estimate response propensities for respondents and, through these, to calculate nonresponse weights (Blom 2009; Kreuter et al. 2010; Krueger and West 2014). Nonresponse weights and design weights are often combined. One then calculates the combined inclusion probability $p_{i,inclusion}$, this is the product of selection probability $p_{i,selection}$ and response propensity $p_{i,response}$: $w_i = (p_{i,inclusion})^{-1} = (p_{i,selection} \cdot p_{i,response})^{-1}$.

Many surveys (Little and Vartivarian 2005, 161) do not provide design weights or nonresponse weights but instead or additionally post-stratification or raking weights. Post-stratification weights are used to adjust the sample to known population totals (Lumley 2010, 136). Post-stratification is also called cell weighting or adjustment weighting. This weighting approach can be considered when the distribution of auxiliary variables (e.g., sex, age groups) is known in the population and differs from those in the sample. For example, if we have one additional dichotomized covariate U , the calculation would be as follows: Using the sample proportion $prop_s$ and the population proportion $prop_p$ of this covariate, we can calculate the weights by taking the inverse of the probability $p_i(U_i = 1) = \frac{prop_s \cdot n}{prop_p \cdot N}$ or $p_i(U_i = 0) = \frac{(1-prop_s) \cdot n}{(1-prop_p) \cdot N}$ respectively (Little 1993, 1001). Weights also often go through several stages of weighting (design weighting should always precede post-stratification). Post-stratification grows more complicated with the number of post-stratification variables. If the joint distribution of the variables is not known but only the marginal distributions, it is necessary to use alternative closely related techniques like raking (Deville and Särndal 1992).

Survey-weighted regression analysis – Hybrid models

We will now cover the topic of survey-weighted regression. We will briefly examine survey-weighted regressions with just one data set in this section. We then move on to the case with pooled complex survey data in the next section. For the moment, we concentrate on

5. Better Together?



- (a) Exemplary data setting with different means per strata (altered colors) but same data-generating model for the dependent variable.
- (b) Exemplary data setting with different slopes per strata. Points with the same color belong to the same strata.

Figure (5.1)

three simple cases: (1) a simple linear regression meeting the Gauss-Markov assumptions and observations drawn with exogenous sampling, i.e., independent from the error term, (2) a simple linear regression where we are faced with endogenous sampling, also called selection on unobservables or informative sampling, and (3) a linear heterogeneity of effects model where data points have different sampling probabilities depending on the groups. While in the first case, the use of weights is usually discouraged, it can be beneficial in the second and third case.

Case 1: We start with a simple linear regression meeting the Gauss-Markov assumptions and exogenous sampling. We are interested in the regression of a variable Y on X . For the moment, we only have data from one survey. Let us assume that the population from which this survey was drawn has different strata s , and the population units have had differing sampling probabilities depending on the strata s . The coefficient β is the same in all strata. The error term has a normal distribution with mean 0 and variance σ^2 , e , and the strata are independent. Since the sampling is only dependent on the strata, it is also independent of the error term e and exogenous. Like the sampling probabilities, the means of X may differ per strata (see Figure 5.1 a).

The reader should note that the Gauss-Markov (GM) assumptions are still met. The

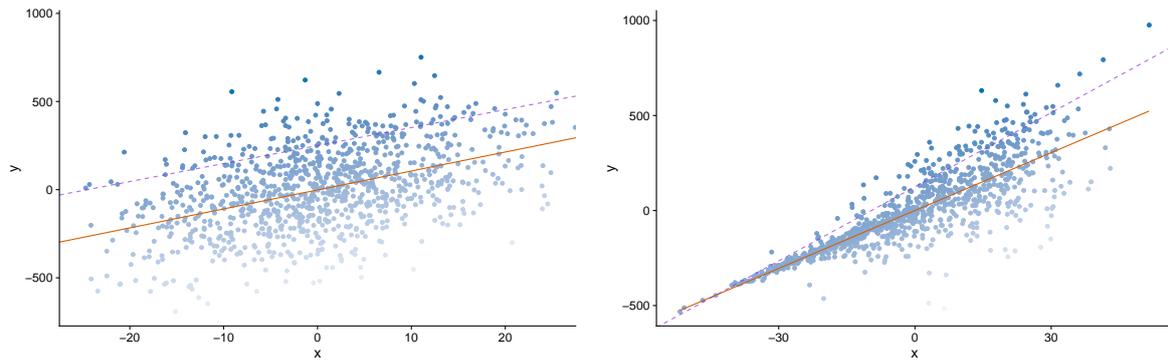
5. Better Together?

GM assumptions only concern the set of error terms (Verbeek 2004, 15): the error terms have to have mean zero, they must be homoscedastic, and distinct error terms must be uncorrelated. Since the GM assumptions are met, there is no reason to prefer the weighted estimate to the ordinary unweighted estimate. As an example, the unweighted OLS estimate $\hat{\beta}$ is already approximately unbiased and has minimum variance among all linear approximately unbiased estimators (DuMouchel and Duncan 1983, 536). This is true even though the means of a variable X differ per strata, and the sampling probabilities differ. While the weighted coefficients are also approximately unbiased, they are not the best linear approximately unbiased estimates (BLUE), whereas the OLS estimates are (Verbeek 2004, 16-17). Let us now move on to cases where weights are needed for approximately unbiased estimates.

Case 2: Again, we want to estimate a simple linear regression of Y on X . However, this time the sampling is endogenous/informative, e.g., depending on the error term. Error terms cannot be observed. For this reason, endogenous sampling is also called selection on unobservables. Estimates not weighted for the endogenous sampling are biased (Solon et al. 2013, 15).

A classic example of endogenous sampling is oversampling, which depends on the outcome variable and, therefore, the error term. For example, a researcher attempts to estimate a regression model of family income on years of education. If high-income families with a low number of schooling years were oversampled compared to middle and low-income families with the same number of years, then the error term is correlated with the sampling probabilities, and the regression estimates will be biased. We have visualized this situation in Figure 5.2 a. Less transparent colors represent higher sampling probabilities. One can easily see that the regression line estimated without weights will have a higher intercept than the population regression line in red. Plus, in the case of heterogeneous variance for the error term, endogenous sampling can shift the slope estimate as well (see Figure

5. Better Together?



- (a) Exemplary data setting with heterogeneous Y . Less transparent colors represent higher sampling probabilities. The red solid line is the regression line of Y on X in the population. The dashed purple line is the estimated unweighted regression line for a sample drawn with unequal sampling probabilities.
- (b) Exemplary data with different endogenous sampling probabilities. Less transparent colors represent higher sampling probabilities. The red solid line is the regression line of Y on X in the population. The dashed purple line is the estimated unweighted regression line for a sample drawn with unequal sampling probabilities.

Figure (5.2)

5.2 b).

In the case of endogenous sampling, one needs to use inverse-probability weights to achieve consistent estimates (Solon et al. 2013, 17). One can also transform this case into an example of endogenous nonresponse. If the high-income families have a higher response propensity than middle and low-income families with the same number of schooling years, we would need nonresponse weights to achieve approximately unbiased estimates.

However, we have to note that estimating nonresponse or post-stratification weights correcting for *all* the bias is highly unrealistic. The bias reduction depends on two associations: (1) on the correlation between the response propensity and the response propensities predictors and (2) on the correlation between the error term and the response propensities predictors (Little and Vartivarian 2003). Unfortunately, these correlations are often not very high (Kreuter et al. 2010, 405).

We return to the endogenous nonresponse example, where high-income families with a

5. Better Together?

low number of years of schooling are responding less often than others with the same number of years. If we regress income on education without weighting, the error term will be correlated with the response propensity. This leads to biased regression estimates. If we want to correct this bias through post-stratification, we need one or more variables correlated with the error term and the response propensity. E.g., let us assume that families with higher income tend to live in bigger cities than families with the same number of years of schooling but lower income. Then, we could use the size of the city as a post-stratification variable. Bias reduction would depend on the correlation between the error term of the regression and city size, as well as the association between inclusion probability and city size. However, both correlations will often be small in practical applications.

Case 3: Last but not least, there are linear heterogeneity of effects models where data points have different sampling probabilities depending on the groups (see Figure 5.1 b). It is not possible to give general recommendations for which estimate, unweighted $\hat{\beta}$ or weighted $\hat{\beta}_w$, to use when assuming a model with heterogeneous coefficients. Neither $\hat{\beta}$ nor $\hat{\beta}_w$ are generally approximately unbiased estimates of the average coefficient $\bar{\beta}$ (DuMouchel and Duncan 1983, 537). Solon et al. (2013, 19) show that there are two reasons why the unweighted estimate does not identify the average population effect $\bar{\beta}$. The first is easy to understand. Let us assume that children's effect on partnership quality is more negative in urban areas than in rural areas. If we oversample rural areas, we would not expect the unweighted estimate of the average coefficient to be the same as the average population effect. This is the first source of bias in the case of an unweighted estimate. However, Solon et al. (2013, 20) provide another reason why the OLS estimate is not consistent. Extreme values of the independent variables can have a considerable influence on the estimates. Therefore, the unweighted average also depends on the difference in the within-strata variance of the independent variables. The weights

5. *Better Together?*

deal with the first source of bias but not the second. If the within-variances are equal in all strata, WLS is thus consistent, and OLS is not. As Solon et al. (2013, 20) notice, this is the “knife-edge special case” and not true in general. They recommend comparing weighted *and* unweighted estimates to get an idea of possible bias.

Adding additional covariates to the model would be another possibility to account for nonresponse or oversampling and remove bias (Pfeffermann and Sverchkov 2009, 461). This approach would correspond to the classical model-based approach for inference by Fisher (1922). However, the covariates might not be available in some cases, but only the weights (Solon et al. 2013, 15–17). The inclusion of additional covariates might complicate the interpretation of focus model parameters (Sterba 2009, 727), for an example see Pfeffermann and Sverchkov (2009, 463). Also, in the case of harmonized surveys, conditioning on observed selection/nonresponse variables will be complicated since they are likely to differ between surveys. Therefore, we strongly prefer hybrid models that are primarily model-based and account for disproportionate selection through weighting instead of conditioning on all complex sampling features (see Sterba 2009 for a comparison of the model-based, design-based, and hybrid framework).

Conclusion

Out of the three cases we covered – (1) a simple linear regression meeting the Gauss-Markov assumptions and with exogenous sampling, (2) a simple linear regression where we are faced with endogenous sampling, and (3) a linear heterogeneity of effects model with different sampling probabilities depending on strata – we need to use survey-weighted regressions in the latter two. In the first case, survey weighting only leads to an inflation of the variance of the regression coefficients. However, whether a study falls under (1) is usually uncertain, and it is, therefore, advisable to compare weighted and unweighted estimates.

Survey-weighted regression analysis – The meta-analytical case

Moving on to combined and harmonized data sets, it is easy to see that the need for survey weighting in the case of endogenous sampling and heterogeneity of effects models (or a combination of these problems) does not simply vanish in the meta-analytical case. To make this explicit, let us remember that each data set is first analyzed separately in two-stage meta-analytical models. If these models' unweighted coefficients are biased, the combined coefficient will be biased, too (except in the unlikely case in which these biases cancel each other out).

For practical applications, this still leaves us wondering if we even need to use survey weights due to nonresponse or effect heterogeneity. Bollen et al. (2016) provides a review of various diagnostic tests used to determine if survey weights are necessary for regression analysis. These tests can be divided into two groups: the first group of tests examines the difference between weighted and unweighted regression coefficients; the second one whether, conditional on the independent variable, the dependent variable, and the weights are correlated (DuMouchel and Duncan 1983; Fuller 2009; Pfeiffermann 1993; Asparouhov and Muthen 2007). However, one should note that these tests were developed for OLS/WLS regression with a single data set.

Plus, it is uncertain and unlikely that *all* the surveys fall into the case of exogenous sampling and no effect heterogeneity. This problem is especially pressing with real complex surveys. Surveys can have multiple aspects of over-sampling across various stages, or it is unclear whether the outcome variables are exogenous. Therefore, we recommend comparing weighted and unweighted estimates in all cases to get an idea of possible bias.

5.3. A comparison of one-stage and two-stage approaches in case of pooled complex survey data

5.3.1. Introduction to the simulation design

We will now further explore the use of survey weights in the context of regression analysis of complex survey data after ex-post harmonization. We look at two questions that arise when conducting a regression analysis with weighted observations after pooling: Do survey-weighted one-stage and two-stage analysis perform differently? And is it possible to include random effects into the survey-weighted analysis, especially if we have to assume study heterogeneity?

When evaluating the performance of different analytical approaches, researchers, in general, look at the bias, the variance, the root mean square error (RMSE), and the coverage of these approaches. While bias, variance, and RMSE should be as small as possible, coverage should be around 0.95 in the case of standard 95%-confidence intervals. In our case, it is cumbersome to derive these four measures analytically for our multitude of approaches. Therefore, we turn to Monte Carlo simulations to calculate estimates of bias, variance, RMSE, and coverage (Cov.). We take care to model data settings of particular interest to all disciplines that work with non-experimental survey/observational data. We focused on relatively simple simulation scenarios to attribute any differences in performance to the different meta-analytical approaches. Simulated surveys are homogeneous, except for Simulation Nr. 3 and 4 in section 5.3.3. An overview of the data settings, the implemented sampling procedures, and the meta-analytical approaches we used are given in Tables 5.1 and 5.2.

Table (5.1) Simulation settings for simulations discussed in the article. All simulations were conducted with 1000 MC repetitions.

Sim. Nr	Special interest in	Pop. size N	Numb. surv k	n_k	Strata + Independent Var. Gen.	Dependent Var. Generation	Sampling	
1	One-stage vs. two-stage 1	10^6	5	10^3	Five strata $s = (1, 2, 3, 4, 5)$, equally large ($10^6/5$). $X_s \sim N(\mu_s, 10^2)$ with $\mu_s, \mu = (-20, -10, 0, 10, 20)$.	Strata slope component γ_s . $\gamma = (30, 15, 0, -15, -30)$. $\alpha = 0, \beta = 10$ $Y_s = (\beta + \gamma_s) * X_s + e_s, e \sim N(0, 70^2)$	Comb. of strata-specific prob. and endogenous samp, high variation in weights	
2	One-stage vs. two-stage 2	Same as Simulation Setting Nr. 1						Comb. of strata-specific prob. and endogenous samp, lower variation in weights
3	Study heterogeneity	10^6 (25 times)	25		Same as Simulation Setting Nr. 1	Superpopulation intercept distr. $\alpha_{super} \sim N(0, 400^2)$ plus strata interc. α_s , equal for all pop. $\alpha = (800, 450, 0, -450, -800), \beta = 10$ $Y_{super,s} = \alpha_{super} + \alpha_s + \beta X_{super,s} + e_{super,s}$ $e_{super,s} \sim N(0, 70^2)$	From each population one survey is drawn. Probabilities depending on strata s	
4	Study heterogeneity	10^6 (25 times)	25		Same as Simulation Setting Nr. 1	Superpopulation slope distr. $\beta_{super} \sim N(0, 20^2)$ plus strata-spec. slope γ_s , $\gamma = (30, 15, 0, -15, -30)$, equal for all pop. $\alpha = 0$. $Y_{super,s} = \alpha + (\beta_{super} + \gamma_s) X_{super,s} + e_{super,s}$ $e_{super,s} \sim N(0, 70^2)$.	From each population one survey is drawn. Probabilities depending on strata s	

Table (5.2) Models used in the various simulations.

Name	Model Formula	Survey Weights
2 St. (FE) unweighted	<i>First stage:</i> $y_{ik} = \alpha + \beta x_{ik} + e_{ik}, e_{ik} \sim N(0, \sigma_k)$ for every survey k	No weighting
2 St. (FE) weighted	<i>Second stage:</i> $\hat{\beta}_{\bullet} = (\sum_{k=1}^K \hat{\beta}_k w_k) * (\sum_{k=1}^K w_k)^{-1}, w_k^* = (var(\hat{\beta}_k))^{-1}$	Inverse-probability weights
	<i>First stage:</i> $y_{ik} = \alpha + \beta x_{ik} + e_{ik}, e_{ik} \sim N(0, \sigma_k)$ for every survey k	
2 St. RE weighted	<i>Second stage:</i> $\hat{\beta}_{\bullet} = (\sum_{k=1}^K \hat{\beta}_k w_k) * (\sum_{k=1}^K w_k)^{-1}, w_k^* = (var(\hat{\beta}_k))^{-1}$	Inverse-probability weights
	<i>First stage:</i> $y_{ik} = \alpha + \beta x_{ik} + e_{ik}, e_{ik} \sim N(0, \sigma_k)$ for every survey k	
<i>Second stage:</i> $\hat{\beta}_{\bullet} = (\sum_{k=1}^K \hat{\beta}_k w_k) * (\sum_{k=1}^K w_k)^{-1}, w_k^* = (var(\hat{\beta}_k) + \hat{\tau}^2)^{-1}$		
1 St. unweighted	$y_{ik} = \alpha + \beta x_{ik} + e_{ik}, e_{ik} \sim N(0, \sigma^2)$	No weighting
1 St. weighted	$y_{ik} = \alpha + \beta x_{ik} + e_{ik}, e_{ik} \sim N(0, \sigma^2)$	Inverse probability weights
1 St. FE unweighted	$y_{ik} = \alpha_k + \beta x_{ik} + e_{ik}, e_{ik} \sim N(0, \sigma^2)$	No weighting
1 St. FE weighted	$y_{ik} = \alpha_k + \beta x_{ik} + e_{ik}, e_{ik} \sim N(0, \sigma^2)$	Inverse probability weights
1 St. RE unweighted	$y_{ik} = \alpha_k + \beta x_{ik} + e_{ik}, e_{ik} \sim N(0, \sigma^2), \alpha_k \sim N(\alpha, \tau_\alpha^2)$	No weighting
1 St. RE (int.) weighted A	$y_{ik} = \alpha_k + \beta x_{ik} + e_{ik}, e_{ik} \sim N(0, \sigma^2), \alpha_k \sim N(\alpha, \tau_\alpha^2)$	Inv. prob. weights, transformed with method 'A' (Carle 2009)
1 St. RE (int.) weighted B	$y_{ik} = \alpha_k + \beta x_{ik} + e_{ik}, e_{ik} \sim N(0, \sigma^2), \alpha_k \sim N(\alpha, \tau_\alpha^2)$	Inv. prob. weights, transformed with method 'A' (Carle 2009)
1 St RE (slp.) weighted A	$y_{ik} = \alpha + \beta_k x_{ik} + e_{ik}, e_{ik} \sim N(0, \sigma^2), \beta_k \sim N(\beta, \tau_\beta^2)$	Inv. prob. weights, transformed with method 'B' (Carle 2009)
1 St. RE (slp.) weighted A	$y_{ik} = \alpha + \beta_k x_{ik} + e_{ik}, e_{ik} \sim N(0, \sigma^2), \beta_k \sim N(\beta, \tau_\beta^2)$	Inv. prob. weights, transformed with method 'B' (Carle 2009)

5.3.2. Methodological decisions and differences between one-stage/two-stage meta-analysis

State of the Art

In the following, we shift our focus away from the more general questions ‘When do we have to weight and which weights should we use?’ to more specific questions. We are interested in differences between the two general approaches – one-stage vs. two-stage analysis. We have already summed up the known differences between one-stage and two-stage approaches in section 5.2.3.

We will now add another new reason why a one-stage meta-analysis might outperform a two-stage analysis in practice. We will demonstrate in our next simulation that with a moderately high coefficient of variation for the weights, the assumption of known within-study variances is highly problematic and can result in biased point estimates.

Simulation and Results

In the first two simulations, we use two different heterogeneity of slope models for data generation. While the sampling is endogenous for both of them, we change the range of sampling probabilities, therefore changing the weights’ variance. In Simulation Nr. 1 (see Table 5.1), the coefficient of variation CV for the weights amounts to approx. 0.76. This CV value is not unrealistic; it corresponds to the coefficient of variation for the weights in the second wave of the German family survey ‘Familiensurvey’ (Deutsches Jugendinstitut (DJI) 2003). The coefficient of variation CV for the weights in Simulation Nr. 2 is smaller; it is only half this size. This difference leads to differences in the performance of one-stage and two-stage meta-analyses. The weighted *two-stage* slope estimates are *biased* in case of a high CV (Simulation 1), whereas the weighted *one-stage* estimates are *approximately unbiased* (see Figure 5.3). The two-stage results will only be approximately unbiased when – all other things being equal – we have low variance in the weights (Simulation 2, not shown separately here).

5. Better Together?

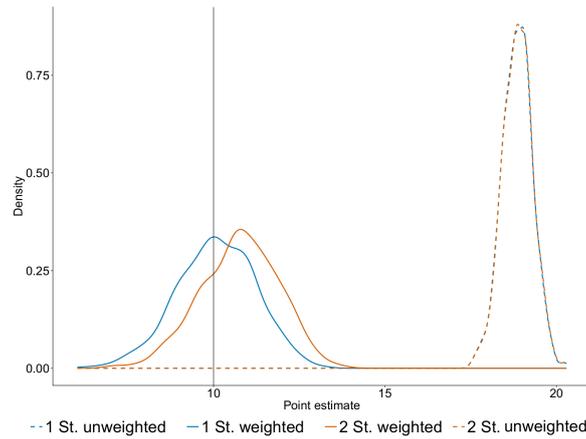
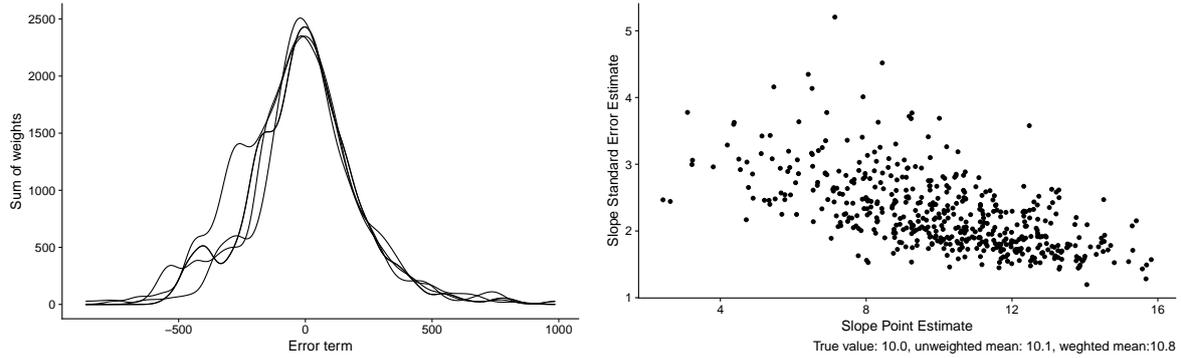


Figure (5.3) Density plot of slope point estimates. 1000 Monte Carlo repetitions. For simulation setting see Simulation Nr. 1 in Table 5.1.

To understand the mechanism behind this non-intuitive result, we look at the distribution of the survey-weighted error terms from several surveys when the coefficient of variation for the weights is high; we see that some of the surveys have bumps at negative values of the error term (see Figure 5.4 a).

This happens when observations with low inclusion probability are included. Since these observations have very high inverse-probability weights, they also increase the variance estimate. Simultaneously, point estimates are lower in these cases since the highly influential observations drag the point estimate down. If we now weight the point estimates by their variance estimates in the second stage, the point estimates which are smaller and have larger variance estimates get a lower (meta-analytical) weight than the larger point estimates with smaller variance estimates (see Figure 5.4 b). Therefore, we overestimate the coefficient in our two-stage meta-analysis. The extent of the bias will depend on the correlation between the added variance due to the weights and the point estimates. Providing corrections of the variance regarding the amount of variance that is due to weighting is difficult. If we used our knowledge that the variance before weighting is equal for all surveys in our simulation, we could get an approximately unbiased estimate. Still, we can not know the distribution between the two variance sources with certainty

5. Better Together?



- (a) Kernel ‘density’ plot of weighted error terms for five different surveys. The area under the curves sum up to the population size. For simulation setting see Simulation Nr. 1 in Table 5.1.
- (b) Scatter plot of slope point estimates against standard error estimates for 500 surveys. For simulation setting see Simulation Nr. 1 in Table 5.1.

Figure (5.4)

in real-world applications. Corrections are difficult to implement (Kish 1965; Le et al. 2002).

Instead of correcting the two-stage meta-analytical weights, we recommend using a one-stage meta-analysis, especially in the case of a high coefficient of variation for the weights. A high coefficient of variation for the weights can easily occur in the case of post-stratification with many sparsely occupied cells. However, we acknowledge that this might not always be possible, e.g., if researchers want to include information on strata/primary sampling units into their model (Rabe-Hesketh and Skrondal 2006). This would further complicate a one-stage meta-analytical model and is an important argument in favor of two-stage models.

5.3.3. Study heterogeneity

We now take a closer look at study homogeneity/heterogeneity and introduce a transformation needed in the case of survey-weighted one-stage meta-analyses with RE. We speak of study homogeneity when all studies in a meta-analysis were undertaken in the same way with the same measurement, target population, etc. However, we will seldom

5. Better Together?

be able to assume study homogeneity. If this assumption is violated, e.g., the studies used different measurements for the same construct, or the target population differs slightly between surveys, we have to deal with and model so-called study heterogeneity (Viechtbauer 2007).

Study heterogeneity is a problem since reflecting the heterogeneity of the studies in our analysis may be complex and challenging. However, it is often crucial since variables and target populations will somewhat differ between included surveys. A conventional approach is to use a model with random effects for the surveys. This can be done regardless of whether we estimate a one-stage or two-stage analysis. Estimating a two-stage analysis with random effects or including a random effect in a one-stage analysis is equivalent to assuming that the true coefficient is not the same for all studies. Instead, the true coefficient has a distribution over surveys. Researchers often choose a normal distribution for this distribution (Burke et al. 2017, 859).

Unfortunately, the use of random effects is tricky in the case of a weighted one-stage analysis. It is known that weights cannot be used in their ‘raw’ form in weighted multi-level/hierarchical analysis since the point and standard error estimates will be biased (Asparouhov 2006, 442, 445). Asparouhov (2006), building on earlier work by Pfeffermann et al. (1998), recommended two methods ‘A’ and ‘B’ (see Equation 5.4 and 5.5) for transforming the weights (Asparouhov 2006, 443).

Method ‘A’ scales the weights so that the transformed weights sum to the cluster sample size (in our case the survey sample size n_k):

$$w_{ik}^* = w_{ik} \left(\frac{n_k}{\sum_i w_{ik}} \right). \quad (5.4)$$

Method ‘B’ scales the weights so that the transformed weights sum to the effective sample size:

5. Better Together?

$$w_{ik}^* = w_{ik} \left(\frac{\sum_i w_{ik}}{\sum_i w_{ik}^2} \right). \quad (5.5)$$

‘A’ and ‘B’ become equivalent when cluster sizes and cluster sample sizes are constant (Asparouhov 2006, 445). Method ‘A’ is better suited if we want to examine regression coefficients, and method ‘B’ is recommended when we are interested in estimates of heterogeneity (Carle 2009, 10).

Multi-level models that incorporate survey weights use a pseudo maximum likelihood (PML) approach, and the weights w_{ik}^* , transformed with either method ‘A’ or ‘B’ (Asparouhov 2006) are included then into the PML (see Rabe-Hesketh and Skrondal 2006, 806). Carle (2009) compares a variety of software programs and software implementations. However, even Carle (2009, 3) and Asparouhov (2006) concede that the transformation methods ‘A’ and ‘B’ are not ideal and recommend analyzing with both scaling methods and then comparing the results.

Simulation and Results

We conducted the two simulations Nr. 3 and 4 to explore modeling study heterogeneity and the inclusion of random effects in a weighted one-stage analysis. In Simulation Nr. 3, we created 25 artificial populations, each with a different intercept. These finite populations come from a theoretical super-population. Each finite population has five strata. These five strata have different intercept components in the data generating model. The sampling probability also differs by strata. We draw an equally sized survey from each of the populations. What is crucial is that we now have study heterogeneity, which has to be reflected in the meta-analytical model. So far, we only looked at homogeneous studies (same population, same measurement, etc.). In this simulation, we compare the weighted one-stage analysis, which does not model the study’s heterogeneity (no random effect), and different one-stage analyses incorporating either a random intercept or a random slope. Of course, we also compare the two transformation approaches (Method

5. Better Together?

Table (5.3) Performance measures. For the simulation setting see Simulation Nr. 3 in Table 5.1. Description of all analysis models in Table 5.2. SE=Empirical standard error, RMSE= root-mean-square error, cov.= coverage.

	Intercept estimate				Slope estimate			
	Bias	SE	RMSE	Cov.	Bias	SE	RMSE	Cov.
One-stage weighted	0.02	5.81	5.81	0.97	0.01	0.64	0.64	0.94
One-stage RE (int.) w. A	0.57	5.36	5.39	1.00	0.01	0.53	0.53	0.82
One-stage RE (int.) w. B	0.57	5.36	5.39	1.00	0.01	0.53	0.53	0.82

Table (5.4) Performance measures. For the simulation setting see Simulation Nr. 4 in Table 5.1. Description of all analysis models in Table 5.2. SE=Empirical standard error, RMSE= root-mean-square error, cov.= coverage.

	Intercept estimate				Slope estimate			
	Bias	SE	RMSE	Cov.	Bias	SE	RMSE	Cov.
One-stage weighted	0.07	2.73	2.73	0.95	0.01	0.43	0.43	0.96
One-stage RE (int.) w. A	0.07	2.73	2.73	0.93	0.04	0.42	0.42	0.65
One-stage RE (int.) w. B	0.07	2.73	2.73	0.93	0.06	0.43	0.43	0.63
One-stage RE (slp.) w. A	0.05	2.01	2.01	0.89	0.08	0.33	0.34	1.00
One-stage RE (slp.) w. B	0.05	2.01	2.01	0.89	0.08	0.33	0.34	1.00

‘A’ and ‘B’).

Moving on to the results for the first simulation with heterogeneous intercepts between studies (see Table 5.3), we see that the RMSE for the intercept (the intercept now differs between super-populations) is slightly better for the weighted one-stage MA with RE than without RE. The intercept coverage is above 0.95 for all weighted approaches. However, the coverage is below 0.9 for the slope of the one-stage RE model (which does not differ between super-populations). We could not find differences between the two transformation methods for our case (Method ‘A’ and ‘B’, see Equation 5.4 and 5.5). This was to be expected since our study sample sizes are constant.

Social-science researchers are often interested in estimating the influence of one variable on another and are therefore often concentrating on slope estimates. Consequently, we created another simulation (Nr. 5), where we again created 25 artificial populations but now with a different slope per study instead of a different intercept.

The results (see Table 5.4) are similar to the first simulation. When we include a random effect for the slope into our meta-analysis, the coverage for the slope is very high but not optimal for the common intercept. Including an RE for the intercept instead of a slope RE does not solve the problem; the slope coverage is meager.

To sum up, we did not observe important differences between the two transformation methods discussed by Carle (2009) and Asparouhov (2006). However, it is important to note that if we *have* to include RE into the meta-analytical model due to study heterogeneity, good results for the parameters affected by study heterogeneity go hand in hand with less optimal results for homogeneous parameters. Both transformation do not lead to optimal results.

5.4. A practical example: Same-sex couples and their satisfaction with family life in Germany

We will conclude this chapter with a brief practical example conducting a small meta-analysis with data from three German surveys, to further illustrate the methodological issues we handled in the simulations.

In the following, we will look at the satisfaction with family relationships for people in a same-sex relationship versus peoples in different-sex relationships. Previous research is not conclusive on whether a person in same-sex relationships experiences family relationships differently than other persons in relationships, and whether people in same-sex couples potentially experience unique stressors regarding their family relationships (Cramer and Roach 1988; Willoughby et al. 2008; Baiocco et al. 2014; Schneider et al. 2016; Lampis et al. 2020).

Our example will use German survey data to examine the satisfaction with family relationships for same-sex and opposite-sex couples from three German surveys. The data

5. *Better Together?*

sets used are the first wave of the pairfam panel, the second wave of the Generations and Gender Survey, and the wave “ba” of the Socio-economic Panel (SOEP). The German Family Panel pairfam (“Panel Analysis of Intimate Relationships and Family Dynamics”) was launched in 2008 and is a multi-disciplinary, longitudinal study for researching partnership and family dynamics in Germany (Brüderl et al. 2017). The Generations and Gender Survey (GGS) is a longitudinal study intended to provide information about the relationships between children and their parents and relationships in couples (Ruckdeschel et al. 2006). The German Socio-Economic Panel (SOEP) is a longitudinal panel that started in 1984 and includes questions on household composition, occupation, employment, earnings, health, and life satisfaction (Wagner et al. 2007).

These surveys all included questions on (1) the sex of the current partner (which allowed to identify people in same-sex and opposite-sex relationships) and (2) satisfaction with family life (in case of the GGS they asked after the satisfaction with the relationship to various family members whose average we took for the satisfaction with the family). The surveys and panel waves were all conducted between 2008 and 2010.

We do not have direct information on the sexual orientation of respondents. However, as already mentioned, we can infer who is living in a same-sex partnership since respondents were asked for their partner’s sex. Concerning the legal situation in Germany, in 2008/2010 (and since 2001), same-sex couples could enter a civil partnership but not marriage. This lasted until 2017, when marriage was also opened for same-sex couples in Germany.

Our target population is people in Germany older than the age of 15 that are either married, in a civil partnership, or have been in a partnership for longer than six months (LATs or in cohabitation). We will reflect the subgrouping (people in relationships) in the variance estimation of the survey-weighted regression (West et al. 2008). We also reflect the fact that while the original target populations of the three survey are largely overlapping between the GGS (people in Germany aged 20-83) and SOEP (people in

5. Better Together?

Table (5.5) One-stage and two-stage meta-analysis with data from from GGS wave 3, pairfam wave 1, SOEP wave “y” (In all three surveys subsamples were taken: persons in partnerships).

<i>One-stage model, weighted observations</i>	
GGS Intercept	8.41*** (0.03)
pairfam Intercept	8.67*** (0.03)
SOEP Intercept	7.85*** (0.04)
Same-sex Partnership	-0.81*** (0.23)
Num. obs.	17579
<i>Two-stage model (same-sex partnership estimate), weighted observations</i>	
Same-sex Partnership	-0.40** (0.17)

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

Germany older than 16), pairfam has a particular population (cohorts born 1971-73, 1981-83, and 1991-93, ergo 15-17, 25-27 and 35-37 years old).

We will mirror the various overlaps through composite factors that are also often used in dual-frame surveys to correct for the overlap of the frames. We used the composite factor proposed by Xia et al. (2010).

Concerning the two-stage analysis, in the first stage, we estimate simple linear regressions of the family satisfaction regressed on the same-sex partnership indicator (and an intercept) for each survey separately. We use a fixed-effect (meta-analytical) model for the second stage (with the estimated regression coefficients from the first stage as effect estimates) since the number of surveys is insufficient for random effects.

The one-stage analysis model has three separate intercepts for each of the three different surveys and the same-sex partnership indicator as independent variables. We used the survey package (Lumley 2016) for the weighted estimation of the one-stage model and the first stage of the two-stage models.

The estimated (meta-analytical) coefficients can be seen in Table 5.5, the estimated

coefficients for the three surveys in the first stage of the two-stage approach are available in the Appendix (see Tables A.9, A.10 and A.11). We see quite a big difference between the two estimates. The two-stage model is strongly driven towards zero by the estimate from the GGS, which is not significantly different from zero. Since its standard error is of the same order as those of SOEP and pairfam, it greatly influences the two-stage model.¹ Of course, survey weighting might not be the only source of differences between the one-stage and two-stage estimates (see Burke et al. 2017). Factors that also (could) influence the estimates are the clustering of estimates in surveys, the specification of residual variances and the “unbalanced” samples (regarding the proportion of same-sex couples). However, the example underlines the point made by Burke et al. (2017) that one-stage and two-stage estimates can differ quite substantially and our point that survey weights are a possible factor that leads to these differences.

5.5. Discussion

Results

The main research question of this article was when and how survey weighting should be used in regression-based meta-analyses. We have answered several crucial questions about the inclusion of survey weights in the regression analysis of complex surveys after ex-post harmonization. We first covered *when* survey weights should be included in the analysis. Survey weights are required for approximately unbiased estimates in the case of endogenous sampling. They are also helpful in the heterogeneity of effects models when strata differ not only in their sampling probability but also in their coefficients. However, weights can also increase the variance of estimates.

Another crucial topic is the difference between the one-stage and two-stage approaches.

¹The coefficient of variation is non-negligible, 77% for the GGS, for pairfam 85% and 106% for the SOEP (after correcting for the overlap), so in a similar range or higher than in Section 5.3.2.

5. Better Together?

Burke et al. (2017) gave ten general recommendations for using one-stage or two-stage IPD meta-analyses and reasons why results may differ between the two methods. We added another one to that list: If we have a medium to high coefficient of variation of the survey weights, the assumption of known within-study variances cannot be upheld for the two-stage meta-analytical approach. Two-stage meta-analytical estimates will be biased.

Another point that we covered is the inclusion of random effects in a one-stage MA. Carle (2009) and Asparouhov (2006) demonstrated that even the performance of RE models with transformed weights is not always optimal. We confirmed this observation.

Implications

Researchers analyzing pooled complex survey data cannot brush aside the topic of survey weights. They have to examine carefully what purpose the weights in their combined surveys serve. Are they design weights used to correct for unequal sampling probabilities? Then the researcher has to carefully check if effects are expected to differ between sampling groups or if endogenous sampling might be a problem. If yes, observations have to be weighted.

While weighting observations can lead to bias for two-stage point estimates in case of high variance for the survey weights, one-stage analysis avoids this trap. However, pooling complex survey data for one-stage analysis is not without its pitfalls. The inclusion of random effects in a one-stage analysis to model study heterogeneity is another critical issue. Using the scaling methods proposed by Asparouhov (2006) is essential for computation in this case, but does not guarantee satisfying results.

Limitations and Future Research

The question of using weights will concern many analyses in the social sciences with survey data. In our simulations, we explored linear regressions. More research has to be conducted to check if conclusions drawn for linear models are also applicable to other

5. *Better Together?*

models from the class of generalized linear models, e.g., logistic regressions. Another interesting class of models would be survival models, e.g., proportional hazards models. We also only briefly covered the topic of the inclusion of weights in a one-stage meta-analysis with random effects. We were able to draw valuable lessons from the neighboring field of weighted multi-level analysis. However, optimal scaling methods are not yet available for weighted RE models like one-stage RE analysis (Asparouhov 2006; Carle 2009).

Last but not least, we only covered the use of survey weights to avoid bias. Weighting is, however, also used in regressions to correct for heteroscedasticity and improve efficiency. The use of weights to improve efficiency in meta-analyses is another open research question.

Conclusions

This article has conducted the first explorations into the field of weighted analysis of pooled complex survey data. We have identified several settings where survey weights are needed to achieve approximately unbiased estimates. We also found differences in performance between the two main (meta-)analytical approaches – one-stage and two-stage analysis. In a two-stage analysis, bias can be introduced through the weighting procedure. Fortunately, by avoiding the step of estimating point and variance estimates for the single surveys, a weighted one-stage analysis remains approximately unbiased. Even though we recommend using the one-stage analysis approach in the case of weighted complex survey data, researchers should take care of necessary weights transformations and model study heterogeneity appropriately.

Bibliography

- Asparouhov, T. (2006). General Multi-Level Modeling with Sampling Weights. *Communications in Statistics - Theory and Methods*, 35(3):439–460.
- Asparouhov, T. and Muthen, B. (2007). Testing for Informative Weights and Weights Trimming in Multivariate Modeling With Survey Data. Proceedings, Section on Survey Research Methods, American Statistical Association.
- Baiocco, R., Fontanesi, L., Santamaria, F., Ioverno, S., Marasco, B., Baumgartner, E., Willoughby, B. L. B., and Laghi, F. (2014). Negative Parental Responses to Coming Out and Family Functioning in a Sample of Lesbian and Gay Young Adults. *Journal of Child and Family Studies*, 24(5):1490–1500.
- Blom, A. G. (2009). *Measuring, Explaining and Adjusting for Cross-Country Differences in Unit Nonresponse: What Can Process Data Contribute?* A Thesis Submitted for the Degree of Doctor of Philosophy in Applied Social and Economic Research, Institute for Social and Economic Research. University of Essex.
- Boedhoe, P. S. W., Heymans, M. W., Schmaal, L., Abe, Y., Alonso, P., Ameis, S. H., Anticevic, A., Arnold, P. D., Batistuzzo, M. C., Benedetti, F., Beucke, J. C., Bollettini, I., Bose, A., Brem, S., Calvo, A., Calvo, R., Cheng, Y., Cho, K. I. K., Ciullo, V., Dallspezia, S., Denys, D., Feusner, J. D., Fitzgerald, K. D., Fouche, J.-P., Fridgeirsson, E. A., Gruner, P., Hanna, G. L., Hibar, D. P., Hoexter, M. Q., Hu, H., Huyser, C.,

Bibliography

- Jahanshad, N., James, A., Kathmann, N., Kaufmann, C., Koch, K., Kwon, J. S., Lazaro, L., Lochner, C., Marsh, R., Martínez-Zalacaín, I., Mataix-Cols, D., Menchón, J. M., Minuzzi, L., Morer, A., Nakamae, T., Nakao, T., Narayanaswamy, J. C., Nishida, S., Nurmi, E. L., O'Neill, J., Piacentini, J., Piras, F., Piras, F., Reddy, Y. C. J., Reess, T. J., Sakai, Y., Sato, J. R., Simpson, H. B., Soreni, N., Soriano-Mas, C., Spalletta, G., Stevens, M. C., Szeszko, P. R., Tolin, D. F., van Wingen, G. A., Venkatasubramanian, G., Walitza, S., Wang, Z., Yun, J.-Y., Thompson, P. M., Stein, D. J., van den Heuvel, O. A., and and, J. W. R. T. (2019). An Empirical Comparison of Meta- and Mega-Analysis With Data From the ENIGMA Obsessive-Compulsive Disorder Working Group. *Frontiers in Neuroinformatics*, 12(102).
- Bollen, K. A., Biemer, P. P., Karr, A. F., Tueller, S., and Berzofsky, M. E. (2016). Are Survey Weights Needed? A Review of Diagnostic Tests in Regression Analysis. *Annual Review of Statistics and Its Application*, 3(1):375–392.
- Borenstein, M., Hedges, L. V., Higgins, J. P. T., and Rothstein, H. R. (2009). *Introduction to Meta-Analysis*. Wiley, Chichester, U.K.
- Brüderl, J., Hank, K., Huinink, J., Nauck, B., Neyer, F. J., Walper, S., Alt, P., Borschel, E., Buhr, P., Castiglioni, L., Friedrich, S., Finn, C., Garrett, M., Hajek, K., Herzig, M., Huyer-May, B., Lenke, R., Müller, B., Peter, T., Schmiedeberg, C., Schütze, P., Schumann, N., Thönnissen, C., Wetzels, M., and Wilhelm, B. (2017). The German Family Panel (pairfam). Technical Report ZA5678 Data file Version 8.0.0, GESIS Data Archive, Cologne.
- Burke, D. L., Ensor, J., and Riley, R. D. (2017). Meta-Analysis Using Individual Participant Data: One-Stage and Two-Stage Approaches, and Why They May Differ. *Statistics in Medicine*, 36(5):855–875.

Bibliography

- Carle, A. C. (2009). Fitting Multilevel Models in Complex Survey Data with Design Weights: Recommendations. *BMC Medical Research Methodology*, 9(49):1–14.
- CLOSER (2020). The Home of Longitudinal Research. <https://www.closer.ac.uk/>. [Online; accessed 20-February-2020].
- Cochran, W. (2007). *Sampling Techniques*. A Wiley Publication in Applied Statistics. Wiley, 3rd edition.
- Cramer, D. W. and Roach, A. J. (1988). Coming Out to Mom and Dad: A Study of Gay Males and Their Relationships With Their Parents. *Journal of Homosexuality*, 15(3-4):79–92. PMID: 3235830.
- DerSimonian, R. and Laird, N. (1986). Meta-analysis in Clinical Trials. *Controlled Clinical Trials*, 7(3):177–188.
- DerSimonian, R. and Laird, N. (2015). Meta-analysis in Clinical Trials Revisited. *Contemporary Clinical Trials*, 45(Part A):139–145.
- Deutsches Jugendinstitut (DJI) (2003). *Wandel und Entwicklung familialer Lebensformen - 3. Welle (Familiensurvey)*. München. ZA3920 Datenfile Version 1.0.0.
- Deville, J.-C. and Särndal, C.-E. (1992). Calibration Estimators in Survey Sampling. *Journal of the American Statistical Association*, 87(418):376–382.
- DuMouchel, W. H. and Duncan, G. J. (1983). Using Sample Survey Weights in Multiple Regression Analyses of Stratified Samples. *Journal of the American Statistical Association*, 78(383):535–543.
- Elliott, M. R., Raghunathan, T. E., and Schenker, N. (2018). Combining Estimates From Multiple Surveys. In *Wiley StatsRef: Statistics Reference Online*, pages 1–10. American Cancer Society.

Bibliography

- Fisher, R. (1922). On the Mathematical Foundations of Theoretical Statistics. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 222(594-604):309–368.
- Fox, K. (2011). *A Framework for the Meta-Analysis of Survey Data*. Queen’s University. A Thesis submitted to the Department of Mathematics and Statistics.
- Fuller, W. (2009). Analytic Studies. In *Sampling Statistics*, chapter 6, pages 341–390. Wiley-Blackwell.
- Fuller, W. A. and Burmeister, L. F. (1972). Estimators for Samples Selected From Two Overlapping Frames. Proceedings of the Social Statistics Section, American Statistical Association. pages 245–249.
- Granda, P., Wolf, C., and Hadorn, R. (2010). Harmonizing Survey Data. In *Survey Methods in Multinational, Multiregional, and Multicultural Contexts*, chapter 17, pages 315–332. John Wiley & Sons, Ltd.
- 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(260):663–685.
- IPUMS (2020). Integrated Public Use Microdata Series. <https://www.maelstrom-research.org/>. [Online; accessed 20-February-2020].
- Joye, D., Sapin, M., and Wolf, C. (2019). Weights in Comparative Surveys? A Call for Opening the Black Box. *Harmonization: Newsletter on Survey Data Harmonization in the Social Sciences*, 5(2):2–16.
- Kalton, G. and Anderson, D. W. (1986). Sampling Rare Populations. *Journal of the royal statistical society. Series A (general)*, 149(1):65–82.

Bibliography

- Kish, L. (1965). *Survey Sampling*. Wiley Classics Library. J. Wiley.
- Kish, L. (1979). Samples and Censuses. *International Statistical Review/Revue Internationale de Statistique*, 47(2):99–109.
- Kish, L. (1994). Multipopulation Survey Designs: Five Types With Seven Shared Aspects. *International Statistical Review/Revue Internationale de Statistique*, 62(2):167–186.
- Kish, L. (1999). Cumulating/Combining Population Surveys. *Survey Methodology*, 25(2):129–138.
- Kish, L. and Verma, V. (1986). Complete Censuses and Samples. *Journal of Official Statistics*, 2(4):381.
- Korn, E. L. and Graubard, B. I. (1999). Analyses Using Multiple Surveys. In Korn, E. L. and Graubard, B. I., editors, *Analysis of Health Surveys*, chapter 8, pages 278–303. Wiley-Blackwell.
- Kreuter, F., Olson, K., and Wagner, J. and Yan, T. and Ezzati-Rice, T. M. and Casas-Cordero, C. and Lemay, M. and Peytchev, A. and Groves, R. M. and Raghunathan, T. E. (2010). Using Proxy Measures and other Correlates of Survey Outcomes to Adjust for Non-response: Examples From Multiple Surveys. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 173(2):389–407.
- Krueger, B. S. and West, B. T. (2014). Assessing the Potential of Paradata and Other Auxiliary Data for Nonresponse Adjustments. *Public Opinion Quarterly*, 78(4):795–831.
- Lambert, P., Sutton, A., Abrams, K., and Jones, D. (2002). A Comparison of Summary Patient-level Covariates in Meta-regression With Individual Patient Data Meta-analysis. *Journal of Clinical Epidemiology*, 55(1):86 – 94.

Bibliography

- Lampis, J., Simone, S. D., and Belous, C. K. (2020). Relationship Satisfaction, Social Support, and Psychological Well-Being in a Sample of Italian Lesbian and Gay Individuals. *Journal of GLBT Family Studies*, 0(0):1–14.
- Le, T., Brick, J., and Kalton, G. (2002). Decomposing Design Effects. In *Section on Survey Research*, New York.
- Little, R. and Vartivarian, S. (2003). On Weighting the Rates in Non-response Weights. *Statistics in Medicine*, 22(9):1589–1599.
- Little, R. and Vartivarian, S. (2005). Does Weighting for Nonresponse Increase the Variance of Survey Means? *Survey Methodology*, 31(2).
- Little, R. J. A. (1993). Post-Stratification: A Modeler’s Perspective. *Journal of the American Statistical Association*, 88(423):1001–1012.
- Lohr, S. L. and Raghunathan, T. E. (2017). Combining Survey Data With Other Data Sources. *Statistical Science*, 32(2):293–312.
- Lumley, T. (2010). *Complex Surveys*. John Wiley and Sons, Inc.
- Lumley, T. (2016). survey: Analysis of Complex Survey Samples. R Package Version 3.32.
- MAELSTROM (2020). Maelstrom Research. <https://www.maelstrom-research.org/>. [Online; accessed 20-February-2020].
- MTUS (2020). Multinational Time Use Study. <https://www.timeuse.org/mtus/>. [Online; accessed 19-July-2020].
- Olkin, I. and Sampson, A. (1998). Comparison of Meta-Analysis Versus Analysis of Variance of Individual Patient Data. *Biometrics*, 54(1):317–22.

Bibliography

- Pfeffermann, D. (1993). The Role of Sampling Weights When Modeling Survey Data. *International Statistical Review / Revue Internationale de Statistique*, 61(2):317–337.
- Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H., and Rasbash, J. (1998). Weighting for Unequal Selection Probabilities in Multilevel Models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 60(1):23–40.
- Pfeffermann, D. and Sverchkov, M. (2009). Chapter 39 - Inference Under Informative Sampling. In Rao, C., editor, *Handbook of Statistics*, volume 29 of *Handbook of Statistics*, pages 455 – 487. Elsevier.
- Rabe-Hesketh, S. and Skrondal, A. (2006). Multilevel Modelling of Complex Survey Data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169(4):805–827.
- Riley, R. D., Lambert, P. C., and Abo-Zaid, G. (2010). Meta-analysis of Individual Participant Data: Rationale, Conduct, and Reporting. *BMJ*, 340:c221.
- Roberts, G. and Binder, D. (2009). Analyses Based on Combining Similar Information From Multiple Surveys. In *Proceedings. Section on Survey Research Methods. JSM 2009*, pages 2138–2147.
- Ruckdeschel, K., Ette, A., Hullen, G., and Leven, I. (2006). *Generations and Gender Survey: Dokumentation der ersten Welle der Hauptbefragung in Deutschland*, volume 121a of *Materialien zur Bevölkerungswissenschaft*. Bundesinstitut für Bevölkerungsforschung (BIB), Wiesbaden.
- Schneider, B. W., Glover, J., and Turk, C. L. (2016). Predictors of Family Satisfaction Following a Child’s Disclosure of Sexual Orientation. *Journal of GLBT Family Studies*, 12(2):203–215.
- SDR (2020). Survey Data Recycling Project. <https://www.asc.ohio-state.edu/dataharmonization/>. [Online; accessed 20-February-2020].

Bibliography

- Simmonds, M. C., Higginsa, J. P. T., Stewartb, L. A., Tierneyb, J. F., Clarke, M. J., and Thompson, S. G. (2005). Meta-analysis of Individual Patient Data From Randomized Trials: A Review of Methods Used in Practice. *Clinical Trials*, 2(3):209–217.
- Skinner, C. J. and Rao, J. N. (1996). Estimation in Dual Frame Surveys with Complex Designs. *Journal of the American Statistical Association*, 91(433):349–356.
- Snijders, T. A. B. and Bosker, R. J. (1993). Standard Errors and Sample Sizes for Two-Level Research. *Journal of Educational Statistics*, 18(3):237–259.
- Solon, G., Haider, S. J., and Wooldridge, J. (2013). What Are We Weighting For? Working Paper 18859, National Bureau of Economic Research.
- Sterba, S. (2009). Alternative Model-Based and Design-Based Frameworks for Inference From Samples to Populations: From Polarization to Integration. *Multivariate Behavioral Research*, 44(6):711–740.
- Stewart, G. B., Altman, D. G., Askie, L. M., Duley, L., Simmonds, M. C., and Stewart, L. A. (2012). Statistical Analysis of Individual Participant Data Meta-Analyses: A Comparison of Methods and Recommendations for Practice. *PLOS ONE*, 7(10):1–8.
- Verbeek, M. (2004). *A Guide to Modern Econometrics*. Wiley, Chichester, 2. ed., reprint. with corr. edition.
- Veroniki, A. A., Jackson, D., Viechtbauer, W., Bender, R., Bowden, J., Knapp, G., Kuss, O., Higgins, J. P., Langan, D., and Salanti, G. (2016). Methods to Estimate the Between-study Variance and its Uncertainty in Meta-Analysis. *Research Synthesis Methods*, 7(1):55–79.
- Viechtbauer, W. (2007). Accounting for Heterogeneity via Random-Effects Models and Moderator Analyses in Meta-Analysis. *Zeitschrift für Psychologie/Journal of Psychology*, 215(2):104–121.

Bibliography

- 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). Technical report.
- West, B. T., Berglund, P., and Heeringa, S. G. (2008). A Closer Examination of Subpopulation Analysis of Complex-Sample Survey Data. *The Stata Journal*, 8(4):520–531.
- Willoughby, B. L. B., Doty, N. D., and Malik, N. M. (2008). Parental Reactions to Their Child’s Sexual Orientation Disclosure: A Family Stress Perspective. *Parenting*, 8(1):70–91.
- Xia, K., Pedlow, S., and Davern, M. (2010). Dual-Frame Weights (Landline and Cell) for the 2009 Minnesota Health Access Survey. *American Statistical Association - Proceedings of the Survey Research Methods Section*, pages 3912–3922.

6. Conclusion and Discussion

My dissertation had two goals: first, to tackle several understudied missing data problems that are important in ex-post survey harmonization; second, to demonstrate that these problems and the solutions developed are not only applicable to this narrow field but are also of value to adjacent methodological and substantive fields.

In the course of the studies conducted within the framework of this dissertation, I worked on different missing data problems: I tackled sporadically missing data (item nonresponse) and data that are systematically missing either for a unit (unit nonresponse) or a variable (systematically missing data). Because I drew conclusions in the individual main chapters, I will summarize the findings of the four studies only briefly here and discuss their role for ex-post survey harmonization and other research fields.

The first study (Chapter 2) was on the challenges of imputing missing values in (time-invariant) covariates for the discrete-time event model, which is popular in fields like family research and survival analysis in medicine. The added difficulty in that case was that the outcome is usually only partly observed due to censoring, and that it consists of two variables – the event variable and the time-to-event data. In addition, researchers have to decide whether to impute in person-oriented format or person-period format. I tested different imputation approaches based on the popular packages `mice` and `smcfcs`. SMC-FCS in person-period format performed best, both in Monte Carlo simulations with various degrees of heterogeneity and in a real-world example with pairfam data. An important limitation is that the tested MI approaches are not strictly compatible with

6. Conclusion and Discussion

the analysis model, and although they perform excellently in part, it is still desirable to implement a compatible imputation model. The next step to be done is to implement this compatible imputation model in `smcfcs` (Bartlett and Keogh 2019) and compare it with the simpler approximations.

Although the second study (Chapter 3) addressed a different type of missing data problem – systematically missing partner variables – it was also tackled with the help of MI. It was shown that with systematically missing partner variables, imputation approaches without additional bridging studies and based on the assumption of (conditional) independence do not perform adequately. These approaches lead to serious bias toward zero regarding the interdependence of anchor and partner variables. Bridging studies can help to make more appropriate assumptions regarding the strength of the interdependence. They are also used in other contexts, for example, measurement changes in official statistics, survey harmonization, or – implicitly – when using a matrix sampling design. An added difficulty in the case of the multiple imputation of systematically missing partner variables is the fact that heterogeneity between surveys hinders the transferability of relationships. Statistical associations are, in general, more stable than measures of central tendency of a single variable, such as, for example, mean values. Nevertheless, caution is warranted when replacing the assumption of conditional independence with the transferability of associations.

The third study (Chapter 4) dealt with a completely missing variable that could be an important confounder for the analysis the researchers wish to conduct. Rosenbaum and Rubin (1983) proposed a sensitivity analysis approach that allows the very flexible modeling of the plausible violation of the unconfoundedness assumption with four parameters. I developed an interactive R Shiny app named `TippingSens`, an easy-to-use tool to examine the resulting bias from different combinations of parameters. Compared with the tabular approach, it allows researchers to obtain an overview of the results for

6. Conclusion and Discussion

more parameter sets at the same time. Another advantage is that manipulating the values used can provide quick insights into the sets of parameter combinations that would be required to significantly alter the research conclusions derived under the unfoundedness assumption.

The fourth study (Chapter 5) was on the topic of weighted regression analysis after ex-post survey harmonization. Weighted regression analyses are usually employed to avoid so-called analytical errors (West et al. 2016) and to obtain unbiased population-averaged estimates. The need for this can arise both with a single data set and a pooled data set. For pooled data sets, I found differences in performance between the two main (meta-)analytical approaches – one-stage and two-stage analysis. In a two-stage analysis, bias can occur due to the weighting process. However, a weighted one-step analysis remains approximately unbiased.

This dissertation has produced the following results. First, it is clear that, despite benefits such as potentially increased sample size and analytical power, ex-post survey harmonization does not automatically avoid problems with missing data. High item nonresponse rates can still lead to bias and increased variance in complete case analyses; for completely missing variables, multiple imputation solutions must be found or sensitivity analyses performed where necessary. Moreover, new questions arise, for example, in the case of regression analyses with pooled survey data. My dissertation also demonstrates that methodological research in the area of ex-post survey harmonization also produces findings that are relevant to other areas. This applies, in particular, to the studies presented in Chapters 2–4. Discrete-time event data analysis is used to analyze all sorts of events and in many different fields – for example, in economics, the start of a new job after a period of unemployment; in medicine, deaths; or in family studies, the time to the birth of the first child. The SMC-FCS solution, which was shown to perform best (Chapter 2), is thus also usable beyond ex-post survey harmonization in

6. Conclusion and Discussion

partnership research. In multi-actor surveys (see Chapter 3), not only anchors and their partners are interviewed but also other persons with whom the anchors have an important connection, such as children and parents. For these additional respondents, the problem of systematically missing data on secondary respondents can also occur. The problem of missing confounders (see Chapter 4) is not only important for ex-post harmonization in the domain of family research. The proposed app for sensitivity analyses can also be applied in quite different fields where the goal is to make causal claims (e.g., medicine, economics, social sciences).

As limitations and further research steps have been addressed in detail in the conclusions of each chapter, I will mention here one limitation that is perhaps the most crucial question in research synthesis: Are we comparing apples with apples or apples with oranges (Sharpe 1997)? At the core of this critical question regarding study heterogeneity and comparability lie even more fundamental questions: Are study results generalizable (Sears 1986; Coppock et al. 2018; Aronow and Samii 2016)? Are research claims true, or are we looking at rotten apples (Ioannidis 2005)?

This limitation became most apparent in the study presented in Chapter 3, when I had to assume that the partial correlation between anchor and partner respondents was transferable between surveys. Moreover, the goal of weighting the different surveys for the analysis of pooled survey data is to improve the comparability between surveys and to allow inference with regard to a common target population. This dissertation thus points to the need to explore study heterogeneity further and more generally, especially in the field of social sciences. In medical science, systematic exploration of heterogeneity between studies has already been conducted, comparing different medical fields in their likely extent of study heterogeneity (Turner et al. 2012). In the social sciences, explorations of study heterogeneity have been carried out, for example, between probabilistic and non-probabilistic surveys or survey modes (Ansolabehere and Rivers

6. Conclusion and Discussion

2013; Ansolabehere and Schaffner 2014; Callegaro et al. 2014; Weinberg et al. 2014; Coppock 2019). However, these studies (like our own brief excursion in Chapter 3) are generally limited to a specific substantive field, such as election research, and to just a few studies. Further research on the amount of heterogeneity to be expected in different fields of social sciences, types of analyses, or populations/samples would be helpful in many regards. Apart from helping to judge the appropriateness of the use of bridging studies or the interpretation of measures of study heterogeneity in meta-analyses or ex-post survey harmonization projects with only a few studies, it would also facilitate the assessment of the importance of new substantive results and the identification of “rotten apples” in research. However, while such an analysis is an important goal on the horizon, the first step in that direction would be to provide a sufficient pool of meta-analyses and other forms of research synthesis that allows such a comparison of study heterogeneity under different circumstances, as Turner et al. (2012) did for the medical sciences. While there have been numerous projects and efforts to synthesize research in the social sciences, coverage comparable to that in medicine has not yet been achieved (see for example Čehovin et al. 2018 for an overview in survey methodology).

Despite its limitations, this dissertation contributes to facilitating further research (synthesis) by removing barriers erected by missing data. I conclude with a quote from Meng (2000): “The topic of missing data is as old and as extensive as statistics itself – after all, statistics is about knowing the unknown. [...] Much remains to be done, however.” This dissertation has closed some gaps in the field of ex-post survey harmonization and beyond, but much more remains to be done. Onward into the unknown!

Bibliography

- Ansolabehere, S. and Rivers, D. (2013). Cooperative Survey Research. *Annual Review of Political Science*, 16(1):307–329.
- Ansolabehere, S. and Schaffner, B. F. (2014). Does Survey Mode Still Matter? Findings From a 2010 Multi-Mode Comparison. *Political Analysis*, 22(3):285–303.
- Aronow, P. M. and Samii, C. (2016). Does Regression Produce Representative Estimates of Causal Effects? *American Journal of Political Science*, 60(1):250–267.
- Bartlett, J. and Keogh, R. (2019). *smcfcs: Multiple Imputation of Covariates by Substantive Model Compatible Fully Conditional Specification*. R Package Version 1.4.0.
- Callegaro, M., Villar, A., Yeager, D., and Krosnick, J. A. (2014). *A Critical Review of Studies Investigating the Quality of Data Obtained With Online Panels Based on Probability and Nonprobability Samples*, chapter 2, pages 23–53. John Wiley & Sons, Ltd.
- Coppock, A. (2019). Generalizing from Survey Experiments Conducted on Mechanical Turk: A Replication Approach. *Political Science Research and Methods*, 7(3):613–628.
- Coppock, A., Leeper, T. J., and Mullinix, K. J. (2018). Generalizability of Heterogeneous Treatment Effect Estimates Across Samples. *Proceedings of the National Academy of Sciences*, 115(49):12441–12446.

Bibliography

- Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLoS Medicine*, 2(8):e124.
- Meng, X.-L. (2000). Missing Data: Dial M for ??? *Journal of the American Statistical Association*, 95(452):1325–1330.
- Rosenbaum, P. R. and Rubin, D. B. (1983). Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study With Binary Outcome. *Journal of the Royal Statistical Society. Series B (Methodological)*, 45(2):212–218.
- Sears, D. O. (1986). College Sophomores in the Laboratory: Influences of a Narrow Data Base on Social Psychology's View of Human Nature. *Journal of Personality and Social Psychology*, 51(3):515–530.
- Sharpe, D. (1997). Of Apples and Oranges, File Drawers and Garbage: Why Validity Issues in Meta-Analysis Will Not Go Away. *Clinical Psychology Review*, 17(8):881 – 901.
- Turner, R. M., Davey, J., Clarke, M. J., Thompson, S. G., and Higgins, J. P. (2012). Predicting the Extent of Heterogeneity in Meta-Analysis, Using Empirical Data From the Cochrane Database of Systematic Reviews. *International Journal of Epidemiology*, 41(3):818–827.
- Weinberg, J. D., Freese, J., and McElhattan, D. (2014). Comparing Data Characteristics and Results of an Online Factorial Survey Between a Population-Based and a Crowdsource-Recruited Sample. *Sociological Science*, 1:292–310.
- West, B. T., Sakshaug, J. W., and Aurelien, G. A. S. (2016). How Big of a Problem is Analytic Error in Secondary Analyses of Survey Data? *PLOS ONE*, 11(6):1–29.
- Čehovin, G., Bosnjak, M., and Lozar Manfreda, K. (2018). Meta-Analyses in Survey Methodology: A Systematic Review. *Public Opinion Quarterly*, 82(4):641–660.

A. Appendix

A.1. Appendix for Chapter 2: Systematically Missing Partner Variables and Multiple Imputation Strategies: A Case Study with German Relationship Data

A.1.1. Weibull survival times

For the scenario based on the Weibull distribution, we used the method by Bender et al. (2005) for creating survival data. Bender et al. (2005) show how the cumulative baseline hazard H_0 can be inverted and individual-specific survival times be created by the following formula (in case of the Weibull distribution):

$$T = \left(-\frac{\log(U)}{\lambda \exp(\gamma'x)} \right)^{1/\nu}. \quad (\text{A.1})$$

U is a variable following a uniform distribution on the interval from 0 to 1.

We choose as γ -parameter vector $(0.8, 2.2, -0.5, 0.3, -1.4)$, as scale parameter $\lambda = 0.001$ and as shape parameter $\nu = 5$. All survival times $T > 15$ are censored at period $j = 15$.

We rounded the survival times to whole numbers.

A. Appendix

Table (A.1) Evaluation measures for the coefficient β_5 (log-odds scale) of X_5 . Simulation with Weibull survival times

FCS/ SMC-FCS	Data For- mat	Included Survival Time Vars	Bias	Rel. Bias	MSE	Cov	SE	CI len.	AME	
1	Full data		0.00	-0.00	0.05	0.96	0.05	0.21	-0.10	
2	Listwise deletion		-0.08	0.07	0.13	0.86	0.09	0.37	-0.09	
3	FCS	P	$\sum Y_j + E$	0.01	-0.01	0.07	0.92	0.07	0.26	-0.10
4	FCS	P	$\log(T)$	0.38	-0.32	0.39	0.00	0.06	0.25	-0.07
5	FCS	PP	$\sum P_j + E$	0.30	-0.25	0.30	0.00	0.06	0.22	-0.07
6	FCS	PP	$\sum Y_j + E$	0.35	-0.29	0.35	0.00	0.06	0.22	-0.07
7	SMC-FCS	PP	$\sum Y_j; E$	-0.05	0.05	0.08	0.86	0.06	0.25	-0.10
8	SMC-FCS	PP	$\sum P_j; E$	0.02	-0.01	0.07	0.91	0.06	0.25	-0.09

A.1.2. Computational times

Table (A.2) Data set is the reduced pairfam data set described in Section 2.5.

Imputation Package	Data For- mat	Included Survival Time Vari- ables	Time in Seconds for 5 Imputa- tions
mice	P	$\log(T)$	2
mice	P	$\sum Y_j + E$	2
mice	PP	$\sum P_j + E$	110
mice	PP	$\sum Y_j + E$	94
smcfcs	PP	$\sum Y_j + E$	65
smcfcs	PP	$\sum P_j + E$	65

**A.2. Appendix for Chapter 3: Systematically Missing
Partner Variables and Multiple Imputation
Strategies: A Case Study With German
Relationship Data**

**A.2.1. Overview over harmonized data sets and the mapping of
values to the variables used for analysis**

Number of kids	nkids bio birth: ch001- sumkids, bio birth: kids.birth days	Number of all kids born until time of interview, Year of birth kid	Number of births in total	Number of children	[0, ...]	[0, ...]	[0, 10]	[0, 17]	[0, 13]	I()	f(sum kids- kids .birth day >2012))
Big Five, Extra version	extravbdp: - bdp15112, bdp: bdp15102, bdp: bdp15108	Big Five, Extra version	Item for short item skala Big Five extra version	-	1 Niedrig, 5 Hoch	1, Trifft überhaupt nicht zu (Does not apply at all), 7 Trifft voll zu (Does fully apply)	[1, 5]	[1, 7]	-	I()	((8- bdp15112)+ bdp15102+ bdp15108)*5/21
Big Five, Agreeable ness	agree bdp: - bdp15113, bdp: bdp15103, bdp: bdp15106	Big Five, Agreeableness	Item for short item skala Big Five agreeable-ness	-	2 Niedrig, 5 Hoch	1, Trifft überhaupt nicht zu (Does not apply at all), 7 Trifft voll zu (Does fully apply)	[1, 5]	[1, 7]	-	I()	((8- bdp15103)+ bdp15113+ bdp15106)*5/21
Big Five, Neuroticism	neu bdp: - bdp15115, bdp: bdp15105, bdp: bdp15110	Big Five, Neuroticism	Item for short item skala Big Five neuroticism	-	3 Niedrig, 5 Hoch	1, Trifft überhaupt nicht zu (Does not apply at all), 7 Trifft voll zu (Does fully apply)	[1, 5]	[1, 7]	-	I()	((8- bdp15115)+ bdp15105+ bdp15110)*5/21
Big Five, Conscientiousness	con bdp: - bdp15111, bdp: bdp15107, bdp: bdp15101	Big Five, Conscientiousness	Item for short item skala Big Five conscientiousness	-	4 Niedrig, 5 Hoch	1, Trifft überhaupt nicht zu (Does not apply at all), 7 Trifft voll zu (Does fully apply)	[1, 5]	[1, 7]	-	I()	((8- bdp15107)+ bdp15111+ bdp15101)*5/21
Big Five, Openness	open bdp: - bdp15114, bdp: bdp15104, bdp: bdp15109	Big Five, Openness	Item for short item skala Big Five openness	-	5 Niedrig, 5 Hoch	1, Trifft überhaupt nicht zu (Does not apply at all), 7 Trifft voll zu (Does fully apply)	[1, 5]	[1, 7]	-	I()	(bdp15114+ bdp15104+ bdp15109)*5/21

A.2.2. Steps of imputation when using parameters from an external bridging study

Imputation of the systematically missing variable y by the normal model by an adaptation of the method defined by (Rubin 1987, 167) and used in the package `mice`.

y_{bridge} is a vector of values for the systematically missing variable from the original data set, which were however observed in the bridging study. X_{bridge} is the matrix of values for the independent variables in the bridging study. X_{sysmis} is the matrix of values for the independent variables in the original study.

1. Calculate the cross-product matrix $S = X'_{bridge}X_{bridge}$ from the additional bridging study
2. Calculate $V = (S + diag(S)\kappa)^{-1}$, with some small ridge parameter κ .
3. Calculate regression weights $\hat{\beta} = VX'_{bridge}y_{bridge}$ from the bridging study.
4. Draw a random variable $\dot{g} \sim \chi^2_v$ with $v = n_1 - q$.
5. Calculate $\dot{\sigma}^2 = (y_{bridge} - X_{bridge}\hat{\beta})'(y_{bridge} - X_{bridge}\hat{\beta})/\dot{g}$ from the bridging study
6. Draw q independent $N(0, 1)$ variates in vector \dot{z}_1 .
7. Calculate $V^{1/2}$ by Cholesky decomposition.
8. Calculate $\dot{\beta} = \hat{\beta} + \dot{\sigma}\dot{z}_1V^{1/2}$.
9. Draw n_0 independent $N(0, 1)$ variates in vector \dot{z}_2 .
10. Calculate the n_0 values $y_{imp} = X_{sysmis}\dot{\beta} + \dot{z}_2\dot{\sigma}$ with the data from the original data set with systematically missing values.

A.2.3. Univariate simulation results

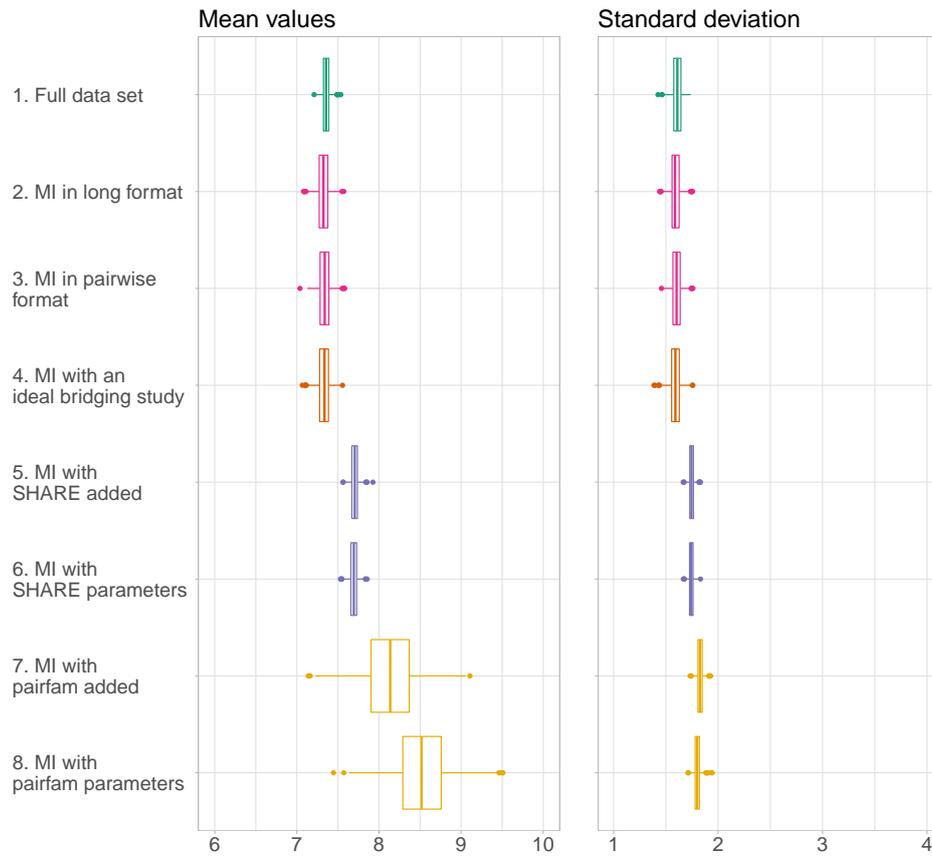


Figure (A.1) Mean and standard deviation values for the variable partner's life satisfaction

A.3. Appendix for Chapter 4: TippingSens: An R Shiny Application to Facilitate Sensitivity Analysis for Causal Inference Under Confounding

A.3.1. A step-by-step guide for the TippingSens R shiny application

In their illustrative application, Rosenbaum and Rubin worked with forking tables to convey their idea for a sensitivity analysis regarding an unobserved binary confounder. This approach is only helpful if a very limited set of plausible values needs to be evaluated per parameter (Rosenbaum and Rubin choose two values for α_1 , β_1 , and γ_1 and three values for q). More parameter sets would result in complex and hard-to-read tables. The app simplifies comparisons by creating visualizations instead of tables, which also allows looking into more possible parameter combinations. We borrow ideas from the plot design developed by Liublinska and Rubin (2014) and adapt it for our purpose for the implementation. Liublinska and Rubin (2014) developed tipping point plots to visualize how different assumptions about missing data affect statistics, such as the average treatment effect. Assuming a binary outcome, their plot's two axes represent the number of positive outcomes among the missing cases for the treatment and the control group, respectively. The plot is divided into tiles, and the statistic of interest is computed for each tile.

We adopt this design to plot the results of a Rosenbaum-Rubin sensitivity analysis. Since the sensitivity analysis depends on four parameters instead of the two parameters considered in Liublinska and Rubin (2014), we choose an interactive display, an R shiny app. The app allows selecting two of the four parameters for the axes. The values for the other two parameters are fixed at user-specified values when generating the plot. Drop-down menus allow manipulating which parameters are displayed on the

A. Appendix

axes. Sliders help to adjust the constant values for the fixed parameters and specify the range for the parameters displayed on the axes. The plot is updated automatically each time any of the specifications is adjusted. The TippingSens app is available at <https://tippingsens.shinyapps.io/TSApp/>.

The general setup

The default setup when loading the app is displayed in Figure A.2. The panel, which will eventually display the sensitivity plots (Panel 1 in Figure A.2), contains a brief summary regarding the interpretation of the different sensitivity parameters. A first sensitivity plot is created as soon as the user interacts with any menu or slider. Default data for this first plot are based on the illustrative example in Rosenbaum and Rubin (1983). In addition to Panel 1, we have a panel containing two drop-down menus to choose the sensitivity parameters printed on the axes (Panel 2). We also have two sliders to set the range limits for the parameters on the axes (Panel 3) and two sliders for the fixed sensitivity parameters (Panel 4). The user can upload data in Panel 5. The last drop-down menu in Panel 6 offers two different color fillings for the plot.

A. Appendix

TippingSens App for Rosenbaum-Rubin Sensitivity Analyses

Please move any slider to create your first sensitivity plot. Default data are taken from Rosenbaum and Rubin (1983).

Remember:
alpha - Log odds ratio of the confounder regarding the outcome in the treatment group
beta - Log odds ratio of the confounder regarding the outcome in the control group
gamma - Log odds ratio of the confounder regarding the treatment assignment
q - Prevalence of the binary confounder

1

[Download Plot](#)

Choose which sensitivity parameters should be displayed on the axes of the output (the other two parameters will be treated as fixed).

alpha

beta

2

Choose a range for the parameters displayed on the axes.

Range for alpha_1

Range for beta_1

3

Choose a value for the parameters treated as fixed.

Value for gamma_1

Value for q

4

Choose CSV File for the data (treatment and outcome).

Browse... No file selected

5

Take care to specify column names correctly. The column names should be 'Treatment' and 'Outcome'. No rownames or missing values are allowed. Separators are allowed to be comma, semicolon, or tab. Decimal separators are allowed to be comma or point.

An example file can be found here (<https://github.com/CaroHaensch/TippingSensExampleFiles>).

Separator

- Comma
- Semicolon
- Tab

Decimal separator

- Point
- Comma

The colour filling

range

6

When choosing range the largest value will be orange and the smallest white, when choosing zero to max negative values will be blue, positive ones orange and values near zero will be white.

Figure (A.2) Default appearance of the TippingSens app before interacting with any menu/slider. Red numbers added by the authors.

A.3.2. Using the app

To illustrate how to interact with the app, we use data from the illustrative example in Rosenbaum and Rubin (1983). The authors use data from a clinical study investigating the effects of two different treatments (coronary artery bypass surgery or medical therapy)

A. Appendix

on symptomatic relief from coronary artery disease. Of the 1515 patients contained in the study, 590 received surgery, while 925 underwent medical treatment. The outcome variable in the study is an indicator of improvement six months after cardiac catheterization (1 = improvement, 0 = no improvement). This data set is also used as a default if no other data are provided.

Uploading the data

Note that the app assumes that the data are well balanced, i.e., any steps to ensure the balance between the treatment group and control group regarding the observed covariates have been performed before loading the app. Under this assumption, the only information required is the data on the outcome and the treatment indicator. The TippingSens app expects that this information is provided in the form of a csv-file containing two columns: one labeled “Treatment” and the other labeled “Outcome”. Commas, semicolons, or tabs can be used as column separators, and points or commas can be used as decimal separators, but the selected separators need to be specified in the data upload panel (see Panel 5 in Figure A.2). An example, how the csv-file would look like for the data from Rosenbaum and Rubin (1983) is depicted in Figure A.3 (following the notation in the original article, surgery is coded as 1, and medical treatment is coded as 0). The data are also available at https://osf.io/35wvf/?view_only=d7950bfd15314e75ad12db9f4f751bd7.

The csv-file containing the data can be uploaded by clicking on the browse button and selecting the file from the appropriate folder.

Adjusting the parameters

Once the data from the (quasi-)experiment are uploaded, researchers have to think about appropriate sensitivity parameters. In their illustration, Rosenbaum and Rubin used three different values for the sensitivity parameter q : 0.1, 0.5, and 0.9. They also assumed that an unobserved variable could double or half the odds of a recovery in the treatment (surgery) and the control (medical therapy) group. Doubling the odds of recovery in the

A. Appendix

Branch: master TippingSensExampleFiles / RosenbaumRubinExampleAppendix / DataFrame.csv Find file Copy path

haenscaa Reorganizing in folders ffa5028 on 26 Mar

0 contributors

1516 lines (1516 sloc) | 15.2 KB Raw Blame History

Search this file...

	Treatment	Outcome
1		
2	1	0
3	2	0
4	3	0
5	4	0
6	5	0
7	6	0
8	7	0
9	8	0
10	9	0
11	10	0
12	11	0
13	12	0

Figure (A.3) Example of input data for the app in csv format containing two columns labeled “Outcome” and “Treatment”.

treatment group is equivalent to $e^{\alpha_1} = 2$, therefore the sensitivity parameter needs to be $\alpha_1 = \ln(2) = 0.693$. Reducing the odds by half is equivalent to $e^{\alpha_1} = 1/2$, therefore $\alpha_1 = \ln(1/2) = -0.693$. Finally, they assumed that the unobserved variable could double or triple the odds of getting the treatment.

To compare the results from the app with those of Rosenbaum and Rubin (1983), we keep α_1 and β_1 as the parameters to be displayed on the axes and change q and γ_1 from their default values to $q = 0.5$ and $\gamma_1 = 0.69$ (the app only allows two decimal places). We also limit the ranges for α_1 and β_1 to $[-0.69, 0.69]$ to allow for a direct comparison with the results presented Table 2 from Rosenbaum and Rubin (1983). The selected parameter settings are displayed in Figure A.4. The app generates a sensitivity plot, as displayed in Figure A.5. The plot can also be downloaded as a png-file to allow easy integration into

A. Appendix

Choose which sensitivity parameters should be displayed on the axes of the output (the other two parameters will be treated as fixed).

alpha

beta

Choose a range for the parameters displayed on the axes.

Range for alpha_1

Range for beta_1

Chose a value for the parameters treated as fixed.

Value for gamma_1

Value for q

Choose CSV File for the data (treatment and outcome).

Browse... DataFrame.csv

Upload complete

Take care to specify column names correctly. The column names should be 'Treatment' and 'Outcome'. No rownames or missing values are allowed. Separators are allowed to be comma, semicolon, or tab. Decimal separators are allowed to be comma or point.

An example file can be found here (<https://github.com/CaroHaensch/TippingSensExampleFiles>).

Separator

Comma

Semicolon

Tab

Decimal separator

Point

Comma

The colour filling

range

When choosing range the largest value will be orange and the smallest white, when choosing zerotomax negative values will be blue, positive ones orange and values near zero will be white.

Figure (A.4) Parameter settings used to generate output displayed in Figure A.5.

technical reports or research papers.

We can compare the TippingSens plot with the table from Rosenbaum and Rubin by subtracting the probability in the row labeled “M” (for medical treatment) from the probability in the row labeled “S” (for surgery) in Table A.8. Corresponding effect sizes are marked by letters a, b, c and d in Table A.8 and Figure A.6.

A final feature of the app, which we do not illustrate further, is the option to change the color filling (see Panel 6 in Figure A.2). When choosing **range**, the largest value will be orange, and the smallest value will be white, when choosing **zerotomax** negative values will be blue, positive values will be orange, and values near zero will be white (see Figure 3 in the main text for an illustration of the latter setting).

A.3.3. Limitations of the app

To ensure that the handling of the app is intuitive, we deliberately limited the flexibility

A. Appendix

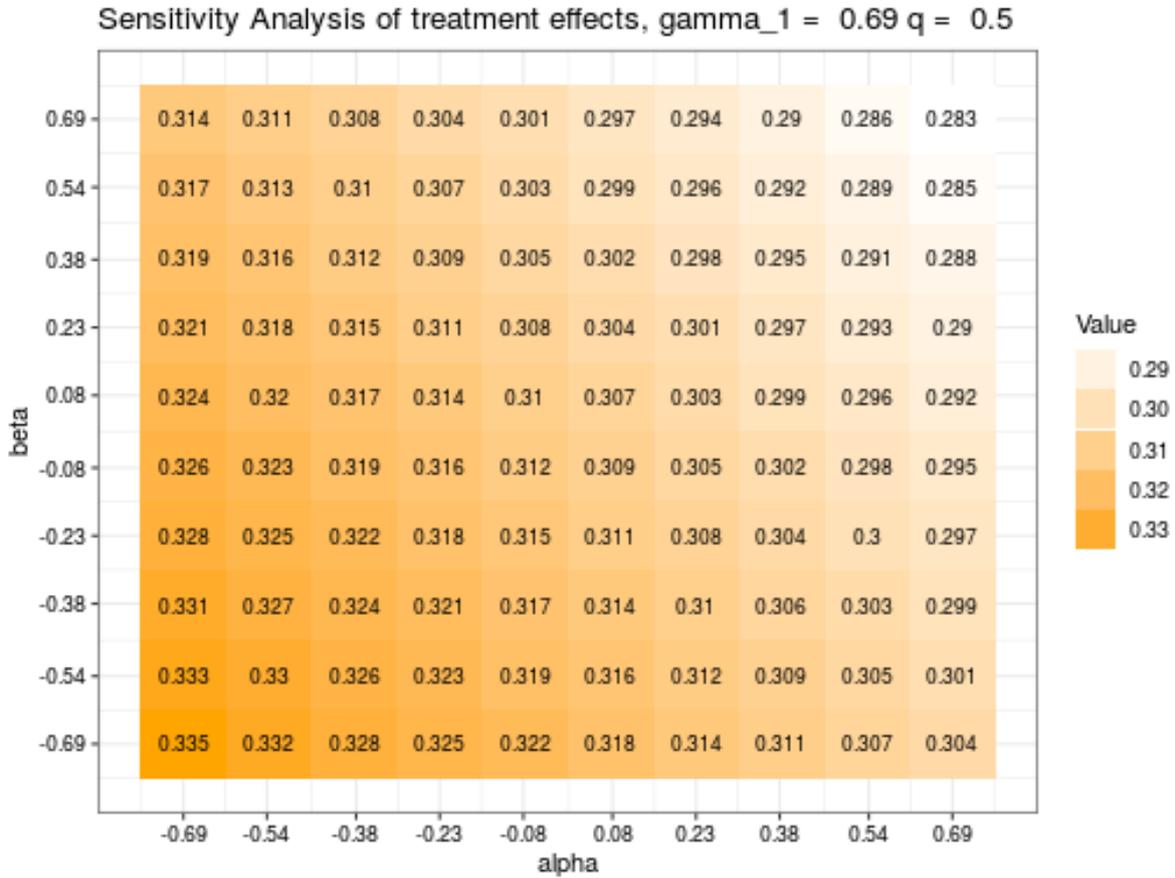


Figure (A.5) Output generated by the TippingSens app based on data from Rosenbaum and Rubin (1983). These data are also the default data used for the app.

of the parameter settings:

- The maximum range for the parameters α_1 , β_1 , and γ_1 is $[-10;10]$ (the range for q is naturally bounded between 0 and 1).
- Parameter values can only be specified up to the second decimal point.
- Two of the four parameters need to be fixed.

We feel that these limitations are necessary to ensure that useful and easily interpretable output can be generated with just a few clicks once the data have been uploaded. At the same time, we tried to strike a balance between user friendliness and broad applicability. For example, there is an inherent trade-off between the bounds for the parameters and

A. Appendix

Table (A.8) Upper half of original table containing sensitivity analysis results from Rosenbaum and Rubin (1983), Table 2, page 216. Results for $exp(\alpha) = 3$ omitted. Parameters $\alpha, \delta_0, \delta_1$ in the notation of Rosenbaum and Rubin are $\gamma_1, \beta_1, \alpha_1$ in the notation used in this paper. Red letters correspond to the results in Figure A.6 and were added by the authors.

Effect of u=1 vs u=0 on treatment assignment z	Effect of u=1 vs u=0 on response under M	Effect of u=1 vs u=0 on response under S	Fraction of patients with u=0: π						
			0.1		0.5		0.9		
Doubles the odds of surgery $exp(\alpha) = 2$	Halves the odds of improvement $exp(\delta_0) = \frac{1}{2}$	Halves the odds of improvement $exp(\delta_1) = \frac{1}{2}$	S	0.67	S	0.68	S	0.68	
			M	0.36	M	0.35	M	0.36	
			Doubles the odds of improvement $exp(\delta_1) = 2$	S	0.66	S	0.65	S	0.66
				M	0.36	M	0.35	M	0.36
			Doubles the odds of improvement $exp(\delta_0) = 2$	S	0.67	S	0.68	S	0.68
				M	0.36	M	0.37	M	0.36
		Doubles the odds of improvement $exp(\delta_1) = 2$	S	0.66	S	0.65	S	0.66	
			M	0.36	M	0.37	M	0.36	

the granularity that can be offered when specifying the values for the fixed parameters. Large bounds on the sliders will allow picking from a very wide range of plausible values. However, it will be more difficult to fix the parameters in specific settings. We feel that fixing the bounds at $[-10;10]$ offers a good compromise. The bounds imply that the assumed odds-ratios are bounded roughly between $4.55 \cdot 10^{-5}$ and 22,000. We believe that these bounds are sufficiently extreme for most practical purposes. At the same time, the bounds ensure that the users can conveniently pick any value between the bounds in incremental steps of 0.01 (users can click on the button of the slider and use the up and down arrow on the keyboard for fine-tuning). Of course, it would also have been possible to let the user specify the values of the fixed parameters directly. However, the interactive property of the app would have been lost. We believe that it is one of the attractive features of the app that the researchers can use the sliders to directly evaluate

A. Appendix

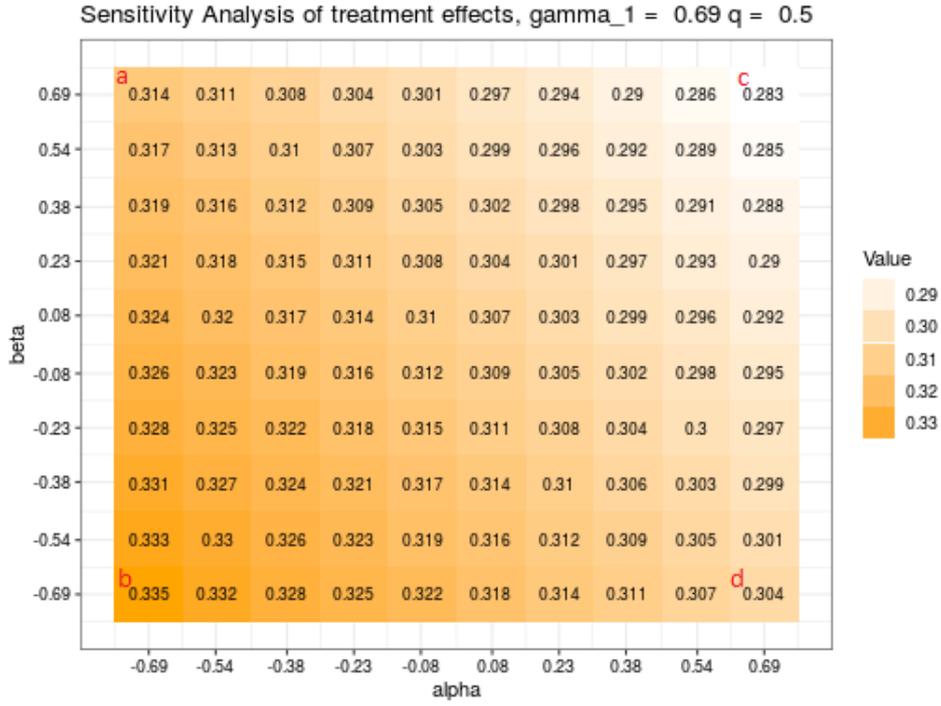


Figure (A.6) Output from the TippingSens app with data taken from Rosenbaum and Rubin (1983). Red letters correspond to the results in Figure A.8 and were added by the authors.

how increasing or decreasing the values of the fixed parameters impacts the estimated treatment effect.

A.4. Appendix for Chapter 5: Better Together?

Regression Analysis of Complex Survey Data After Ex-post Harmonization

Table (A.9) Data taken from pairfam wave 1, subsample filter: people in partnerships. Subsampling was accounted for in the estimation.

(Intercept)	8.67***
	(0.03)
Same-sex Partnership	-1.00**
	(0.35)
Num. obs.	7229

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

Table (A.10) Data taken from GGS wave 2, subsample filter: people in partnerships. Subsampling was accounted for in the estimation.

(Intercept)	8.39***
	(0.03)
Same-sex Partnership	0.10
	(0.26)
Num. obs.	2686

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

A. Appendix

Table (A.11) Data taken from SOEP wave “y”, subsample filter: people in partnerships. Subsampling was accounted for in the estimation.

(Intercept)	7.85***
	(0.04)
Same-sex Partnership	-0.73*
	(0.35)

Num. obs.	7664
-----------	------

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

Bibliography

- Bender, R., Augustin, T., and Blettner, M. (2005). Generating Survival Times to Simulate Cox Proportional Hazards Models. *Statistics in Medicine*, 24(11):1713–1723.
- Liublinska, V. and Rubin, D. B. (2014). Sensitivity Analysis for a Partially Missing Binary Outcome in a Two-arm Randomized Clinical Trial. *Statistics in Medicine*, 33(24):4170–4185.
- Rosenbaum, P. R. and Rubin, D. B. (1983). Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study With Binary Outcome. *Journal of the Royal Statistical Society. Series B (Methodological)*, 45(2):212–218.
- Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley.

The software code for this dissertation is available at the Open Science Framework:

https://osf.io/hy6j9/?view_only=c03d26c445264d469bab45fd4cb67a99.