Advanced Survey Designs for Planned Missing Data

Mehboob Ali

Dissertation
an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig–Maximilians–Universität
München

2021

Advanced Survey Designs for Planned Missing Data

Mehboob Ali

Dissertation
an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig–Maximilians–Universität
München

12.01.2021

Erstberichterstatter: Prof. Dr. Göran Kauermann

Zweitberichterstatter: Prof. Dr. Timo Schmid

Drittberichterstatterin: Prof. Dr. Christian Heumann

Tag der Disputation: 27.05.2021

Zusammenfassung

Das Beispiel, das unsere Forschung motiviert hat, stammt aus einer Umfrage zu Wohnungsmieten, die regelmig in allen groen Stdten in Deutschland durchgefhrt wird. Die Datenerhebung mittels langer Fragebgen ist teuer und/oder zeitaufwendig. Dies legt nahe, eine Umfrage zu konzipieren, die sowohl die Kosten reduziert als auch die Belastung der Befragten verringert. Der Prozess, absichtlich fehlende Werte in das Umfragedesign einzubeziehen, wird blicherweise als geplantes Missing Data Design oder Missing by Survey Design bezeichnet. Zu diesem Zweck verwenden wir zwei geplante Designs fr fehlende Daten, bei denen der Forscher oder die Umfrageorganisation die fehlenden Werte in der Planungsphase absichtlich in eine Umfrage einfgt. Im ersten Erhebungsdesign verwenden wir eine zweistufige Stichprobenstrategie. In der ersten Phase wird eine groe Zufallsstichprobe aus der Grundgesamtheit gezogen und eine kontinuierliche Antwortvariable Y mit einem Satz von kontinuierlichen Kovariaten x erfasst. In der zweiten Phase wird eine kleine Zufallsstichprobe aus der ersten Stichprobe gezogen, um teure kategoriale Kovariaten z zu erfassen. Wir schlagen einen approximativen Schtzansatz fr die semiparametrische Regression unter Verwendung der verfgbaren Informationen vor, der einen kleineren mittleren quadratischen Vorhersagefehler im Vergleich zu verschiedenen Methoden der multiplen Imputation liefert. Wenn wir annehmen, dass die Umfrage im Laufe der Zeit wiederholt gezogen wird, knnen wir eine effiziente Stichprobe der zweiten Phase auswhlen, die die kleinste Schtzvariabilitt fr die Koeffizienten des linearen Regressionsmodells liefert.

Das zweite Design ist ein geteiltes Fragebogenerhebungsdesign, bei dem die gemeinsamen Variablen Y und x von allen Befragten beobachtet werden, aber die Informationen der teuren kategorialen Kovariaten z und w in zwei verschiedenen Komponenten (Stichproben) des Fragebogens beobachtet werden. Die erste Komponente sammelt Informationen ber z, whrend die zweite Komponente Informationen ber w sammelt. Die Informationen werden so gesammelt, dass keine einzelne Stichprobeneinheit die Informationen ber z und w gleichzeitig liefert. Wir nehmen eine bedingte Unabhngigkeit zwischen den spezifischen Variablen (z und w) bei gegebenen gemeinsamen Variablen (Y und x) an und passen drei separate Regressionsmodelle mit derselben Antwortvariablen fr jede Regression an. Anschlieend haben wir diese Schtzungen mit dem vorgeschlagenen Ansatz kombiniert, der einen kleinen mittleren quadratischen Vorhersagefehler liefert. Diese kumulative Dissertation enthlt drei Beitrge und sie werden wie folgt zusammengefasst.

Beitrag 1 beschreibt das Szenario einer zweiphasigen Stichprobe, bei der teure Kovariaten in der ersten Phasenstichprobe fehlen und nur in der zweiten Phasenstichprobe beobachtet werden. Wir nehmen zustzlich an, dass die Umfrage im Laufe der Zeit wiederholt gezogen wird und sowohl die billigen als auch die teuren Variablen aus der vorherigen Umfrage verfgbar sind. Wir imputieren die fehlenden Werte der teuren Kovariablen, indem wir die Daten der ersten Phase mit den zuvor verfgbaren Daten kombinieren. Wir ziehen die Stichprobe der zweiten Phase, die die Designmatrix der Kovariaten fr die imputierte Stichprobe der ersten Phase unter Verwendung der Matrixnorm maximiert. Das vorgeschlagene Stichprobenverfahren whlt eine effiziente Stichprobe der zweiten Phase, die die kleinste Schtzvariabilitt fr die Koeffizienten des Regressionsmodells liefert.

Beitrag 2 befasst sich mit der Situation, in der einige teure Kovariaten in einer relativ groen Erstphasenstichprobe designbedingt fehlen. Die Stichprobe der zweiten Phase wird aus der Stichprobe der ersten Phase gezogen und beobachtet ebenfalls teure Kovariaten. Wir erweitern die Idee von Little (1992) in Richtung nicht-linearer Regression unter Verwendung von Daten mit zwei Phasen-Stichproben. Wir schlagen einen approximativen Schtzungsansatz vor, der nichtparametrische Mittelwert- und Varianzregressionsmodelle fr die erste Phasenstichprobe und ein semi-parametrisches Mittelwertregressionsmodell fr die zweite Phasenstichprobe verwendet. Der vorgeschlagene Ansatz erfordert nicht, die fehlenden Werte zu imputieren.

Beitrag 3 beschreibt das geteilte Fragebogendesign im Kontext des statistischen Matchings. Hier sind einige gemeinsame Variablen fr alle Stichproben (Fragebgen) verfgbar, whrend spezifische Ko-

variaten nur fr die spezifische Stichprobe erfasst werden. Da spezifische Variablen nicht gemeinsam mit gemeinsamen Variablen beobachtet werden, stehen wir also vor einem Identifikationsproblem, um die gemeinsame Verteilung aller interessierenden Variablen zu schtzen. Um das Identifikationsproblem zu lsen, nehmen wir an, dass die spezifischen Variablen angesichts der gemeinsamen Variablen bedingungslos unabhngig sind. Der vorgeschlagene Ansatz schtzt das interessierende Regressionsmodell mit den verfgbaren Daten und erfordert keine Imputation der fehlenden Werte.

Summary

The example which motivated our research comes from a survey on rents for apartments regularly conducted in all large cities in Germany. The data collection by the means of long questionnaire is expensive and/or time consuming. This suggests to design a survey which reduces the cost as well as lessen the respondent's burden. The process of including on purpose missing values in survey design is usually known as planned missing data design or missing by survey design. For this purpose, we use two planned missing data designs, where researcher or survey organization intentionally put the missing values in a survey at planning stage. In first survey design, we use two phase sampling strategy, in the first phase, a large random sample is drawn from the population and a continuous response variable Y with a set of continuous covariates x are recorded. In the second phase, a small random sample out of the first sample is drawn to record expensive categorical covariates z. We propose an approximate estimation approach for semi-parametric regression using the available information which provides smaller mean squared prediction error as compared to different multiple imputations methods. If we assume the survey is drawn repeatedly over time then we can select an efficient second phase sample which provides smallest estimation variability for the coefficients of linear regression model.

The second design is split questionnaire survey design, where common variables Y and x are observed from all respondents but information of expensive categorical covariates z and w is observed in two different components (samples) of the questionnaire. The first component collects information about z while the second component collects information about w. The information is collected in such a way that no single sampling unit provides the information about both z and w simultaneously. We assume conditional independence between specific variables (z and w) given common variables (Y and x) and fit three separate regression models with same response variable for each regression. We then combined these estimates with the proposed approach which provides small mean squared prediction error. This cumulative dissertation contains three contributions and they are summarized as follow.

Contribution 1 describes the scenario of two phase sampling where expensive covariates are missing in first phase sample and observed only in second phase sample. We additionally assume that the survey is drawn repeatedly over time, and both the cheap and expensive variables are available from previous survey. We impute the missing values of expensive covariates by combining the first phase data with previously available data. We draw second phase sample which maximize covariates design matrix for first phase imputed sample using matrix norm. The proposed sampling scheme selects an efficient second phase random sample which provides smallest estimation variability for the coefficients of regression model.

Contribution 2 deals the situation where some expensive covariates are missing by design in a relatively large first phase sample. The second phase sample is drawn from first phase sample and also observed expensive covariates. We extend the idea of Little (1992) towards non-linear regression using two phase sampled data. We propose an approximate estimation approach using non-parametric mean and variance regression models for first phase sample and a semi-parametric

mean regression model for second phase sample. The proposed approach does not require to impute the missing values.

Contribution 3 describes the split questionnaire survey design in the context of statistical matching. Here some common variables are available for all samples (questionnaires) while specific covariates are recorded only for the specific sample. Since specific variables are not observed jointly with common variables, therefore, we face an identification problem to estimate the joint distribution of all the variables of interest. To solve the identification problem, we assume that the specific variables are conditionally independent given the common variables. The proposed approach estimates the regression model of interest with available data and does not require the imputation of the missing values.

Acknowledgement

- I feel a matter of great pleasure to express my sincerest feelings of gratitude for my worthy supervisor Dean Prof. Dr. Göran Kauermann. I am utmost grateful for his guidance and advice for my research throughout the years. Thank you, for your supervision, motivation and your trust.
- I am thankful to Prof. Dr. Timo Schmid and Prof. Dr. Christian Heumann for review of this dissertation.
- I extend my thanks to Prof. Dr. Helmut Küchenhoff and Prof. Dr. Thomas Augustin for being members of the examination committee.
- I gratefully acknowledge the financial support of Punjab Higher Education Commission (PHEC), Lahore, throughout the study period.
- My special thanks go to Amjad Ali Jutt, Assistant Director, Punjab Planning and Development Department, Lahore and Syed Musa Hassan, Director HRD, PHEC, Lahore.
- I express my gratitude to my parents, brothers Arshad Ali, Amjad Ali, Dr. Maqsood Ali and sisters for their prayers and wishes, which have been a great source of comfort and ease during my whole educational period.
- Last but not the least, thanks to my wife Dr. Humera Razzak, my daughter Zarish Mehboob and son Saim Mehboob.

Contributions

The present cumulative dissertation is composed of the following three contributions. They are arranged according to the order they appear in this dissertation.

Contribution 1:

Ali, M. and Kauermann, G. (2021a). Second phase sample selection for repeated survey. Technical Report 237, Department of Statistics, LMU Munich. Available online at:
https://epub.ub.uni-muenchen.de/74729/

The large part of this manuscript is written by Mehboob Ali and the sampling procedure, other valuable input and proofreading of the manuscript is done by Göran Kauermann.

Contribution 2:

Kauermann, G. and Ali, M. (2020). Semi-parametric regression when some (expensive) covariates are missing by design. Accepted for publication in *Journal of Statistical Papers* and available online at: https://doi.org/10.1007/s00362-019-01152-5

The large part of this manuscript is written by Göran Kauermann and Mehboob Ali has written Sections on simulation and real data example. The discussion of this manuscript is written by both the authors and proofreading is done by Göran Kauermann. Both the authors have contributed in the revised version of this manuscript.

Contribution 3:

Ali, M. and Kauermann, G. (2021b). A split questionnaire survey design in the context of statistical matching. Accepted for publication in *Journal of Statistical Methods and Applications* and available online at:
https://doi.org/10.1007/s10260-020-00554-2

The manuscript is written by Mehboob Ali and Göran Kauermann has added several valuable inputs and proofread it. Both the authors have contributed in the revised version of this manuscript.

Contents

1 Introduction

1.1 Overview

Some surveys are conducted regularly to collect the current data and draw inference from data which then play a central role in research in every field of life. Unfortunately, to collect all the intended data is rarely possible and every survey suffers from a common problem of missing data for many reasons which significantly affect on the conclusions of the survey results. There are two main reasons of this missingness, it may be either due to unforeseen reasons or the researcher intentionally put in the survey at planning stage (see also Figure 1). The earlier one is not under the control of the researcher or survey organization, i.e. the survey participants may refuse to provide some information of target questions. However, the later one is commonly referred to as missing data by survey design or planned missing data, which is under the control of researcher and provides reliable results for some studies (Graham et al., 2006). These survey designs do not provide any systematic bias in the statistical results due to control of researcher on planned missingness. When the missingness in a survey is induced by the random sampling design, this is either be missing completely at random or missing at random constellation. In earlier one, the missing values are independent from the fully observed and the unobserved data while in later one, the missing values depend on fully observed variables, that is, the missing values are associated with observed data.

In this cumulative dissertation, we use various estimation methods for planned missing data where the missing values are intentionally put in data by different survey designs. We assume that there is no unforeseen missing data and every participant provides full response to all the required questions of the survey. There are multiple reasons to use planned missing data design. Some of them include the budget constraints, the researcher needs detailed information without increasing burden on participants and the survey costs (Graham et al., 2006), improvement in the quality of sample data (Rässler, 2004), to reduce the non-response rate and the respondent's burden by splitting the long questionnaire (Raghunathan and Grizzle, 1995) and some covariates may be expensive to measure in a survey (Kauermann and Ali, 2020).

The expensive covariates may be obtained through different planned missing data designs as shown in Figure 1 (a) and (b). For example, one can use two phase sampling scheme, where cheap variables are observed in a relatively large random sample and a sub-sample is selected randomly from large sample to record the expensive covariates, see Figure 1 (a). The other planned missing data design is split questionnaire where expensive

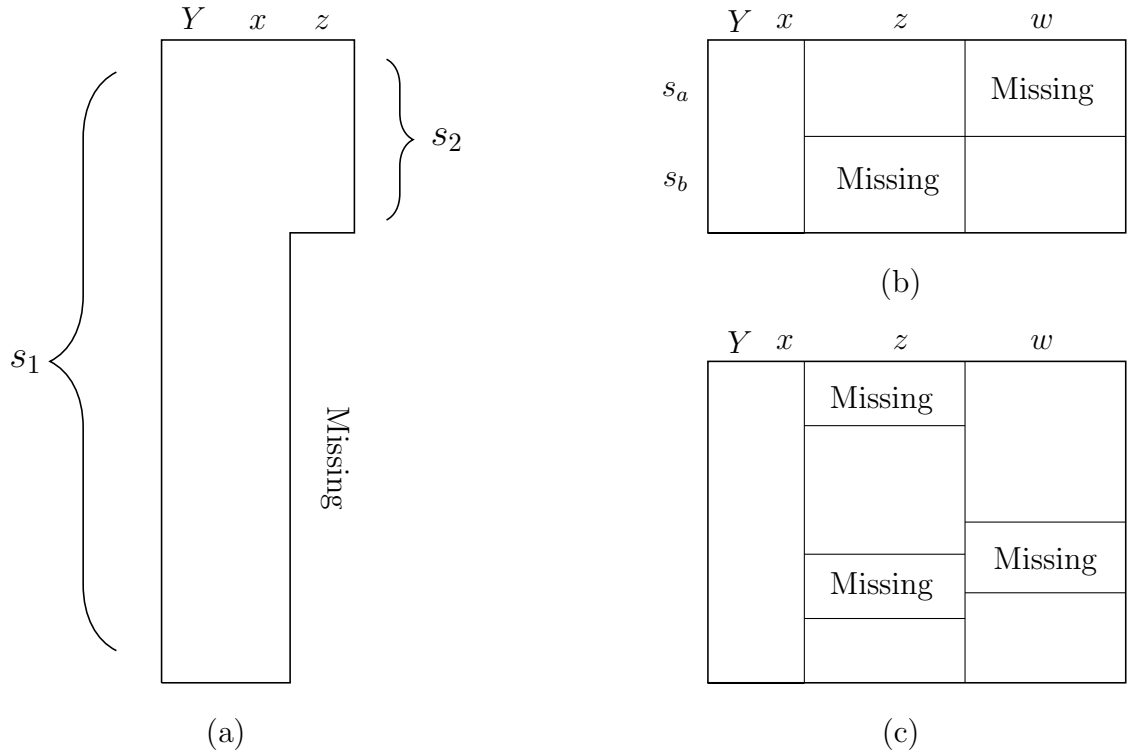

Fig. 1 (a) Missing data pattern for two phase sampling, (b) a split questionnaire survey design and (c) general missing data pattern

specific covariates are observed into two independent samples (questionnaires) and the cheap common variables are recorded for both the samples, see Figure 1 (b). In both the survey designs, the researcher can control the amounts of missing data by intentionally putting the missingness in survey. However, there is one limitation that the statistical power is reduced due to decrease number of observations. The power can be increased by increasing the sample size of survey. The Figure 1 (c) shows unplanned missing data pattern.

The Chapter 1 is organized as follows. Section 1.1 starts with introduction of missing data and its various types. In Section 1.2, we review two phase sampling where the second phase sample is selected with equal and unequal probabilities. Following this, the short review about missing covariates in a relatively large sample is provided in Section 1.3. Section 1.4 describes the different split questionnaire designs and short review about statistical matching. Section 1.5 contains the detail about rental guide data.

1.2 Two phase sampling

When observing certain variables from a target population is time consuming and/or expensive by personal interviews through questionnaires, Neyman (1938) suggests to use

a two phase sampling design. In this sampling, the cheap variables are observed for a relatively large first phase sample and expensive variables are recorded only for a smaller second phase sample drawn from first phase sample. If we ignore the information of first phase when selecting the second phase random sample, this sampling is know as simple random sampling where each sampling unit of first phase has equal chances to be selected in second phase sample.

To utilize the information of first phase sample, the second phase sample can be drawn with inclusion probability involving available data. When each sampling unit has different selection probability, the sampling becomes the unequal probability sampling (see Hanif and Brewer, 1980; Tille, 2006). In case of unequal probability sampling, design weights are assigned to fit regression model on complex survey data to ensure the unbiased results of the model parameters (Pfeffermann and Sverchkov, 2009). A common practice is to use the inverse of sample selection probability as design weight (Pfeffermann, 1993; 1996). These probabilities may depend on covariates or response variable or both of them. If these probabilities depend only on available information of cheap covariates of first phase sample then the sampling design is known as covariate dependent sampling (Kauermann and Ali, 2020) and design is non-informative. In contrast, if the probabilities depend on response variable included in the second phase sample, then the design is known as informative design or outcome/response dependent sampling (Zhou et al., 2007; McIsaac and Cook, 2014). The informative design may include both response and covariates to select a second phase sample. For example, if the collected information of first phase sample include a response variable and some cheap covariates then one may use the residuals of regression model (obtained from running the response variable on available covariates) to select second phase sample (Derkach et al., 2015; Kauermann and Ali, 2020). For detailed discussion about informative and non-information sampling designs, we refer to Pefferemann (2011) or Kim and Skinner (2013).

1.3 Short review when expensive covariates are missing in a large sample

Little (1992) describes the missing covariates problem for linear regression where the response variable Y and covariate x are observed for large sample while a covariate z is missing and available only for sub-sample. He derives the regression coefficients and their variances using maximum likelihood (ML) method and assumes the multivariate normality of joint distribution of Y and z given x. White and Carlin (2010) extend the Little's (1992) work for multivariate z, and derive regression coefficients and their

variances with ML method.

Zhang and Rockette (2005) discuss a semiparametric ML method for the regression where some covariates are always observed but other contain possibly missing values. Zhao et al. (2009) study regression model when some expensive covariates are missing by design and these covariates are observed only in sub-sample. They use two phase sampling to collect data where the regression model is estimated using ML method. Two phase sampling designs are also used where some covariates are expensive to record (McIsaac, 2013; Mandallaz et al., 2013). Lumley (2017) considers the regression model where auxiliary response variable and covariate are observed for all sample data while some covariates are only observed in sub-sample because they are expensive to measure. The missing values of expensive covariates are imputed using available data of two phase sampled and the standard linear regression model is fitted on the imputed data to estimate its coefficient.

Most of the literature cited above deals with linear regression model. We extend the idea of Little (1992) towards non-parametric regression and categorical covariates. We apply semi-parametric regression $E(Y|x,z) = m(x) + z\beta_z$ by assuming the effect of continuous covariate x is a smooth function $m(x)$ while categorical covariates z are modeled parametrically. Penalized-spline smoothing technique can be used to estimate the model parameters.

1.4 A split questionnaire survey design and statistical matching

In split questionnaire design (SQD), each respondent answers a fraction of total questions while the information about common questions is collected from each respondent. The non-response rates are usually high in long questionnaire discourage potential respondents while split questionnaire reduces the respondent's burden (Raghunathan and Grizzle, 1995; Rässler, et al., 2002; Chipperfield and Steel, 2009). A long questionnaire often leads to a loss interest of participants in the survey, making the sample quality low (Peytchev and Peytcheva, 2017) and the SQD can increase quality of response or sample and decrease the non-response rate (Rässler et al., 2002; Stuart and Yu, 2019). Some variables may be expensive to measure from all the participants or may be researcher or survey organization wants to get more information without increasing the survey cost (Graham et al., 2006; Chipperfield et al., 2018).

There are many ways to split the questionnaire, we highlight a few of them in the following (see also Figure 2). Multiple matrix sampling approach introduced by Shoemaker (1973), randomly selects a small subset of the questions from the total questions and only one subset is asked from each respondent. Raghunathan and Grizzle (1995) introduce a SQD

Form	z_1	z_2	z_3
A		Missing	Missing
B	Missing		Missing
C	Missing	Missing	

(a) Shoemaker, 1973

Y x	z_1	z_2	z_3	z_4
A	M	M		
B	M		M	
C	M			M
D		M	M	
E		M		M
F			M	M

M=Missing

(b) Raghunathan and Grizzle, 1995

Form	Y x	z_1	z_2	z_3
A		Missing		
B			Missing	
C				Missing

(c) Graham et al., 2006

Y x	z_1	z_2
A		Missing
B	Missing	

(d) Kim et al., 2016

Fig. 2 Different situations of split questionnaire survey design

based on Shoemaker (1973) multiple matrix sampling scheme, where a questionnaire is divided into different components with near equal number of questions in each component. The bivariate association can be studied due to partial overlap in different components. This design provides similar results as a full questionnaire. Graham et al. (2006) split the questionnaire into four different components/forms, where each respondent answers some common questions as well as two of the three other forms. This design is known as 3-forms design and partial overlap also exists among different forms. Graham et al. (2006) use 3-forms design with the aim to increase the number of questions to get more information without increasing the respondent's burden and survey cost. The other commonly used SQD is that where questionnaire divided into different non-overlap parts and each respondent participates in one part and some certain portion of the questionnaire. This certain portion is asked to all the participants of the target survey. For example, Rässler (2004) uses a survey design where some common variables are observed from all the respondents while the specific variables are recorded in such a way that these variables have no common portion. Kim et al. (2016) use a SQD where a random sample s is selected from population. Further, this sample is splitted into two sub-samples s_a and s_b in such a way that $s_a \cup s_b = s$ and $s_a \cap s_b = \phi$.

To combine the two or more independent samples or data sources to estimate the joint distribution of all the variables of interest which is never jointly observed is usually known

as statistical matching. The other terms used for statistical matching in literature are: file concatenation (Rubin, 1986), data fusion (Rässler, 2002), file matching (Little and Rubin, 2002) or synthetical matching (D'Orazio et al., 2006a). Rubin (1986) considers statistical matching as a type of missing data problem. He studies the situation where variables of interest are present in two different surveys (i.e. information on some variables can be obtained from a specific survey whereas information on other variables can be observed from another survey) or it is not possible to observe complete information in one survey. Rubin uses multiple imputation method to impute the missing variables in order to study the relationship of all variables of interest present in different surveys. Moriarity and Scheuren (2001, 2003) use multivariate normal distribution for statistical matching in order to estimate the regression by assuming conditional independence of specific variables which are not observed simultaneously. Other methods used for statistical matching include a non-iterative Bayesian multiple imputation procedure (Rässler, 2004), excluding those variables which are not available simultaneously (Rendall et al., 2013) and fractional imputation method (Kim et al., 2016).

We follow a SQD similar to Kim et al. (2016) in the context of statistical matching as shown in Figures 1 (b) and 2 (d). By using SQD, we propose an approximate method which neither require imputation of the missing values nor the covariates model. The conditional independence is assumed for specific variables conditioned on common variables.

1.5 Rental guide data

The example which motivated our research comes from a survey on rents for apartments regularly run in all the large cities in Germany. These cities publish a rental guide for apartments which is used as an official instrument to decide the amounts of rent in the German apartments rental market (see e.g. Kneib et al., 2011; Fahrmeir et al., 2013; Fitzenberger and Fuchs, 2017). The guide provides information about the average rent of an apartment in a community or city. Such average rent is usually calculated by using regression model with net rent per square meter as dependent variable while the independent variables are the characteristics of the apartments such as the floor size, the floor type, the building type, the central heating, the bathroom equipment, the kitchen quality and the apartment location. The aim of this rental guide is to predict the rents of apartments based on its characteristics (covariates).

To fit the regression model and predict the rents, a random sample is drawn from all the relevant households and the data of dependent variable and several expensive covariates are obtained through questionnaires by personal interviews. To obtain this data through

long questionnaire is expensive and time consuming. Following the discussion in previous Sections, we proposed two planned missing data designs to obtain the rental guide data. In first design, we plan to use a two phase sampling scheme for data collection. The cheap variable Y, rent per square meter (in Euro) and covariate x_1, the floor space of the apartment, are obtained in first phase sample through telephone survey. The set of expensive covariates z describing the qualities and facilities of the apartment is recorded only in second phase sample. The data from second phase is obtained by personal interviews through questionnaires. The aim is to fit the regression model on two phase sampled data which provides minimum mean squared prediction error. To do that, we propose an approximate estimation approach using semi-parametric regression which provides smaller prediction error. To select an efficient second phase sample to obtain information about expensive variables, we propose a simple sampling procedure. Our proposed sampling procedure provides smallest estimation variability in the regression coefficients. Note that, for this purpose, we additionally observed x_2, the year of construction of apartment in first phase sample along with Y and x_1. Also assume that the survey is drawn repeatedly over time. The proposed sampling procedure is helpful in deciding which apartment is to select in second phase sample.

The second planned missing data design used for rental guide data is a split questionnaire, where we split a long questionnaire into two parts to reduce the length of questionnaire. We selected two independent samples from all the relevant households and observed Y, net rent per square meter, and covariate x_1, the floor space, from both the samples. The categorical covariates z and w describing the qualities and facilities of the apartments are recorded in first and second sample, respectively. With the assumption of conditional independence, we propose an approach to predict the rents of the apartments based on available splitted data.

2 Methodology

Section 2.1 starts with basic idea of parametric regression model and its performance measures. In Section 2.2, we describe non-parametric regression models, and the short review about polynomial and B-splines bases functions. Penalized splines based on B-splines are also described. Semi-parametric models are given in Section 2.3, which are a combination of parametric and non-parametric models. Further, the generalized additive models for location, scale and shape are given in Section 2.4. The matrix norm, which is used to maximize the covariates design matrix to get smaller variance of regression coefficients, is given in Section 2.5. Section 2.6 provides the fundamental approaches of statistical matching, i.e. macro and micro. In Section 2.7, conditional independence assumption is defined to handle the problem of identification of joint distribution. Section 2.8 discusses missing data and its various mechanisms. Some missing data handling methods are also given, for example, complete case analysis, multiple imputation, multivariate imputation by chained equations, classification and regression trees, and random forest.

2.1 Parametric regression

Linear regression model describes the linear relationship of a continuous response variable Y with one or more covariates x. These covariates may be continuous and/or categorical variables. The standard simple linear regression model is

$$Y = \beta_0 + x\beta_1 + \varepsilon, \tag{1}$$

where ε is independently and identically distributed error term with homogeneous variance. The parameters (β_0, β_1) are unknown quantities and need to be estimated. The most commonly used methods to estimate these unknown parameters are ordinary least squares (OLS) and maximum likelihood (ML). The equation (1) can be written in matrix form as

$$Y = X\beta + \varepsilon, \tag{2}$$

where X is a matrix of covariates including intercept term and β is corresponding vector of parameters. The OLS estimator $\hat{\beta}$ is obtained by minimizing

$$(Y - X\beta)^T(Y - X\beta),$$

with respect to β. After differentiation and equating to zero, we get

$$\hat{\beta} = (X^T X)^{-1} X^T Y. \tag{3}$$

Now the estimated version of (**??**) is

$$\hat{Y} = X\hat{\beta}.$$

The variance estimate of $\hat{\beta}$ is

$$\text{var}(\hat{\beta}) = \sigma_\epsilon^2 (X^T X)^{-1}, \tag{4}$$

where σ_ϵ^2 is the variance of residuals ($\varepsilon = Y - \hat{Y}$). If $(X^T X)$ becomes as large as possible, the var($\hat{\beta}$) will be the smallest. The matrix norm can be used to maximize $(X^T X)$ which will be defined in Section 2.5.

2.1.1 Performance measures of regression

2.1.1.1 Mean squared error (MSE)

To see how close the estimated values are to the true observations, we need to define some performance measure in quantitative form. Most of the performance measures are calculated based on residuals of the regression model. The commonly used performance measure of a regression model is mean squared error and given as

$$\hat{MSE} = \frac{1}{n} \sum_i^n (Y_i - \hat{Y}_i)^2.$$

To compare two regression models based on MSE, the model with smaller value of MSE is better than the other. Smaller MSE means that the model predicted values ($\hat{Y}$) are close to the observed values (Y). The other performance measures are bias and variance.

2.1.1.2 Bias

The difference between expected value of estimator (E($\hat{\beta}$)) from the corresponding true value (β) is known as bias, i.e.

$$\text{bias}(\hat{\beta}) = \text{E}(\hat{\beta}) - \beta,$$

where $\text{E}(\hat{\beta}) = \frac{1}{m} \sum_i^m \hat{\beta}_i$ and m is number of simulations.

2.1.1.3 Variance

Variance is expected value of squared difference of values from their mean or true value, i.e.

$$\text{est.var}(\hat{\beta}) = \frac{1}{m}\sum_{i}^{m}(\hat{\beta}_i - \beta)^2 \quad \text{and} \quad \text{E}(\text{var}(\hat{\beta})) = \frac{1}{m}\sum_{i}^{m}\text{var}(\hat{\beta}_i)$$

where est.var($\hat{\beta}$) is estimated variance based on m simulations and E(var($\hat{\beta}$)) is average of model based variances of the regression coefficient like (3).

2.2 Non-parametric regression

A parametric regression model tries to explain the dependency of response variable explained by linear covariates or becomes linear after some transformation (i.e. inverse or squared transformation). If the response variable is linearly related to the covariates, the linear regression is perfectly good (estimators are unbiased and efficient) satisfying all the other assumptions of the model. On the other hand, if the relationship is not linear, it is necessary to use more flexible model which describes this relation well, like non-parametric regression model. In non-parametric regression, the form of the model is not specified explicitly but determined from the data and the parameters are subset of the infinite dimensional vector space. Let assume that the data of a continuous response variable Y and a continuous covariate x are given. Then the standard univariate non-parametric regression model can be written as

$$Y = \beta_0 + m(x) + \varepsilon, \tag{5}$$

where ε is error term and we can make some assumptions about it just like classic linear regression. The $m(.)$ is unspecified smoothing function and different assumptions of this function leads to different modeling choices and different basis functions can be used. Before defining penalized splines, we here briefly describe polynomial and B-splines bases.

2.2.1 Polynomial splines

Looking again at right hand side of model (1), this is a linear combination of 1 and x. This is usually known as basis functions of the model and can be written as

$$B_1(x) = 1 \quad \text{and} \quad B_2(x) = x.$$

If we add x^2 in model (1) then the model becomes

$$Y = \beta_0 + x\beta_1 + x^2\beta_2 + \varepsilon. \tag{6}$$

Now the basis functions for model (4) are

$$B_1(x) = 1, B_2(x) = x \quad \text{and} \quad B_3(x) = x^2$$

Both (1) and (4) are polynomial models, the only difference is that model (1) of first degree while model (4) is of second degree, depending on the power of covariate. For p degree polynomial, the model is

$$Y = \beta_0 + x\beta_1 + x^2\beta_2 + ... + x^p\beta_p + \varepsilon.$$

This model has $p + 1$ basis functions and can be written with its basis functions as

$$Y = B_1(x)\beta_0 + B_2(x)\beta_1 + \ldots + B_{p+1}(x)\beta_p + \varepsilon. \tag{7}$$

2.2.2 B-splines

The polynomial basis is easy to use for spline based regression but sometimes this is not numerically stable for a large number of knots (the values of covariate x where the pieces meet are known as knots). The alternative basis function which has some numerically superior properties is the B-splines. To explain the B-splines, the equation (5) can be represented as

$$m(x) = \sum_{j=1}^{q} u_j B_j(x),$$

where u_j is the coefficient of the basis function B_j. For detailed review of B-splines, we refer to Boor (2001). The basic idea is that by plugging pieces of a certain polynomial degree onto each other we can obtain a smooth function. For each polynomial of p degree, the function $m(.)$ is continuously differentiates $(p-1)$ times. For the given set of c knots, a B-spline basis function of zero degree can be defined as

$$B_j^p(x) = \begin{cases} 1, & \text{if } c_j \leq x \leq c_{j+1}, \\ 0, & \text{otherwise,} \end{cases}$$

where $j = 1, \ldots, C-1$ and knots c_j. For the given set of c knots, a general B-spline basis function of degree $p \geq 1$ is given as

$$B_j^p(x) = \frac{c_{j+1} - x}{c_{j+1} - c_{j+1-p}} B_j^{p-1}(x) + \frac{x - c_{j-p}}{c_j - c_{j-p}} B_{j-1}^{p-1}(x).$$

Alternately saying, that the B-spline of degree p - 1 can be used to construct each B-spline of degree p and each B-spline of arbitrary degree can be traced back to a B-spline of degree zero. Now design matrix with basis functions which can help to define penalized splines, can be written as

$$\mathbf{B} := \begin{pmatrix} B_1(x_1) & \dots & B_c(x_1) \\ \vdots & \ddots & \vdots \\ B_1(x_n) & \dots & B_c(x_n) \end{pmatrix}.$$

2.2.3 Penalized splines

Penalized splines also known as P-splines, is a very popular non-parametric technique originally proposed by O'Sullivan (1986), and Eilers and Marx (1996). The term P-splines is first used by Eilers and Marx (1996) who describe the numerical practicability and flexibility of this approach. A general introduction and flexibility of this approach is given in the book by Ruppert, et al. (2003) and in software (see Wood, 2017; Stasinopoulos et al., 2017). The basic idea of P-splines is to replace the smoothing function $m(x)$ by B-spline basis representation, i.e. replace $m(x)$ in model (**??**) with $B(x)u$, this make whole model parametric where we need to estimate the spline coefficients u. Eilers and Marx (1996) propose to use a large number of knots and imposed difference penalty on coefficients of adjacent B-splines to achieve a smooth fit. Instead of minimizing $(Y - \mathbf{B}u)^T(Y - \mathbf{B}u)$ itself, we can introduce an additional penalty. If we assume a symmetric penalty matrix C, then

$$(Y - \mathbf{B}u)^T(Y - \mathbf{B}u) + \lambda u^T C u, \tag{8}$$

where $\lambda \geq 0$. This equation can be minimized with respect to u using penalized least squares criterion and λ controls the amount of smoothness. Differentiating equation (8) with respect to u and solving the normal equations provide us

$$\hat{u} = (\mathbf{B}^T\mathbf{B} + \lambda C)^{-1}\mathbf{B}^T Y.$$

Note that this equation reduces to (2) when $\lambda = 0$, otherwise both the estimators are different, i.e. we need additional term λC in equation (2). Now the estimated values of $\hat{Y}$ are

$$\hat{Y} = \mathbf{B}(\mathbf{B}^T\mathbf{B} + \lambda C)^{-1}\mathbf{B}^T Y.$$

2.3 Semi-parametric regression

There are often the cases when the functional form of some covariates is not known or fully non-parametric regression model does not perform well, one may use semi-parametric regression model in those situations. Semi-parametric regression models described by Ruppert et al. (2003, 2009) are a fusion between parametric and non-parametric components. These models are smaller than the fully non-parametric regression models. For example, one may be interested in only finite dimensional parameters of semi-parametric regression model. In contrast, in completely non-parametric regression model, the primary interest may be to estimate the infinite dimension of the parameters. The semi-parametric regression models are more flexible than parametric regression models because they can deal with both the parametric and non-parametric components simultaneously. A large class of regression models fall into semi-parametric models such as generalized additive model (GAM) and generalized additive models for location, scale and shape (GAMLSS). These models can accommodate linear and/or non linear effects of the covariates with linear response variable in regression analysis, hence, the models are semi-parametric regression. The commonly used semi-parametric models are partial linear regression models, and a simple such model can be written as

$$Y = \beta_0 + m(x) + z\beta_z + \varepsilon. \tag{9}$$

This model contains the non-parametric part in the form of unspecified smoothing function $m(.)$ and $z\beta_z$ is finite dimensional parameter part and ε is the error term. The smoothing function $m(.)$ is often in form of additive structure in one dimensional non-parametric function.

2.4 Generalized additive models for location, scale and shape

Generalized additive models for location, scale and shape (GAMLSS) are introduced by Rigby and Stasinopoulos (2005) as a general class of statistical models for regression. These models are more flexible and overcome some shortcomings of traditional GAMs. The response variable in GAMLSS is univariate and has different choices in statistical modeling. For example, we can relax the assumption of exponential family for response variable and various other distributions of discrete, continuous, high skewed and/or kurtotic type can be used. The GAMLSS includes both parametric and semi-parametric regression models. The distribution of the response variable is parametric and if the covariates are linear, then model becomes parametric. And when the covariates are non

linearly related with response, the model becomes semi-parametric. The structure of covariates in GAMLSS is additive and various functional forms of covariates are possible, i.e. parametric or non-parametric based on penalized splines. Usually the penalized likelihood estimation method is used to estimate the model coefficients (see Rigby and Stasinopolos, 2005). The GAMLSS have nice feature which allows simultaneously modeling of various parameters of the response distribution. For example, we can run regression model for mean, variance, skewness and kurtosis parameters of the response distribution as a function of covariates and each one of these can be modelled separately with its own covariates. The mean and variance models for response variable Y, non linearly related with a covariate x, can be written as

$$\mu = \beta_0 + m(x) \tag{10}$$

$$\sigma = \beta_{0\sigma} + m(x)_\sigma \tag{11}$$

2.5 Matrix norms

The norm of a matrix measures how large the entries are in a matrix and if a matrix is notated with W then the function of this matrix norm is denoted by ||W||. There are various types of a matrix norm, but all have following features in common:

1. $||W|| \geq 0$ and $||W|| = 0$ if and only if the matrix $W = 0$,

2. $||hW|| = |h|.||W||$, for any scalar h,

3. $||W + U|| \leqslant ||W|| + ||U||$, where U is also a matrix like W,

4. $||WU|| \leqslant ||W||.||U||$.

For optimization problems, the most commonly used matrix norms are: 1-norm, ∞-norm and Frobenous norm, and these matrix norms can be computed as

1-norm

$$||W||_1 = max_j \sum_i^m |w_{ij}|$$

∞-norm

$$||W||_{\infty} = max_i \sum_{j}^{n} |w_{ij}|$$

Frobenius-norm

$$||W||_F = (\sum_{i}^{n} \sum_{j}^{m} |W_{ij}|^2)^{1/2} = (Trace(W^T W))^{1/2}$$

Here w_{ij} denotes the elements of $W_{m.n}$.

2.6 Macro and micro approaches in statistical matching

The purpose of statistical matching is to obtain the joint information about the specific variables which are not jointly observed. According to D'Orazio et al. (2006a), the term joint information can be of two types; first, it can be of joint density or any of its characteristics (marco approach); second, it may refers to complete but synthetic data (micro approach). As mentioned in introductory Chapter 1 (Figure 1 (b)), which shows that specific variables of interest, z and w, are not simultaneously observed and to estimate their joint distribution, we can use marco approach. This approach provides the estimates directly from available data without imputing the missing values of specific variables. On the other hand, if the aim is the estimation of missing information (of specific variables), the micro approach can be used to impute these missing values. This approach provides complete but synthetic data to estimate the parameters of interest. The application of both approaches simultaneously is known as mixed approach of statistical matching (D'Orazio et al., 2006b). In mixed approach, we first estimate the parameters of joint distribution then construct the complete data with synthetic values using hot deck methods (usually). In the next Section, we describe conditional independence assumption and assume macro approach.

2.7 Conditional independence assumption

In statistical matching, an important assumption is often made to analysis the specific variables which are not simultaneously observable. This traditional assumption is commonly known as conditional independence assumption. First we provide the definition of independence and dependence of two random variables, w and z, then we will define conditional independence. The variables w and z are dependent if the probability of w is not equal to the probability of w given z, that is

$$P(w) \neq P(w|z).$$

If both the variables are independent then the probability of w is equal to the probability of w given z as

$$P(w) = P(w|z).$$

To define conditional independence, we need to consider at least three variables w, z and x. Suppose the variables z and w are independent but they both depend individually on a third variable x, i.e.

$$P(w, z|x) = P(w|x).P(z|x),$$

then z and w are said to be *conditionally* independent given x. Their conditional independence is written as

$$z \perp\!\!\!\perp w|x,$$

where $\perp\!\!\!\perp$ sign shows independence.

Refer to Figure 1 (b), the data of common variables y and x are available in both samples s_a and s_b, and the specific variables z and w are not jointly observed. If we use data from sample s_a only then the information of w is ignored and if we use sample s_b only then the information of z is ignored. And if both the samples are used together and we are interested to estimate the joint distribution, the problem of identification is faced because variables z and w are not jointly observed. To overcome this limitation in statistical matching, the conditional independence assumption can be used. This assumption factorizes joint distribution of z and w into marginal distributions: z given y and x; and w given y and x. We assume that variables z and w are conditionally independent given y and x. The joint distribution of (z, w, y, x) with the chain rule of conditional independence can be written as

$$f(z, w, x, y) = f(z|x, y)f(w|x, y)f(y|x)f(x), \tag{12}$$

where $f(z|x, y, w)$ becomes $f(z|x, y)$. The conditional independence assumption enables us to estimate the joint distribution with different sub factors. Using this factorization, we can estimates some factors from sample s_a and the others from sample s_b or from both of them. Hence, there is no identification problem and this assumption makes a unique estimation of $f(z, w, x, y)$. A disadvantage of this assumption is that it is not testable

with available data and provides biased estimate of the joint distribution when it does not hold true. If the conditional independence assumption is true, the available data of both samples, s_a and s_b is sufficient to estimate (8) (Roszka, 2015). If the common variables are closely related to the specific variables then this assumption is reasonable. And this assumption is often made on the view of that the common variables have rich enough information to explain the relation between the specific variables of interest.

2.8 Fundamentals of missing data and multiple imputation

2.8.1 Missing data and its mechanisms

In Chapter 1, we discussed the various types of missing data, however, when dealing with it, it is important to know the reasons of this missingness. The researcher needs to understand the patterns of missing data because it affects the performance of missing data handling techniques. This pattern of missingness is usually known as missing data mechanism. Its probability may depend either on fully observed and/or unobserved variables or independent from both of them. Rubin (1976) classified missing data mechanisms into three parts: Missing completely at random (MCAR), Missing at random (MAR) and Missing not at random (MNAR).

To understand these mechanisms in terms of probability model, suppose the data in matrix Y with n observations in rows and p variables in columns. The values of this data matrix are denoted by y_{ij}, where $i = 1, ..., n$ and $j = 1, ..., p$. We divide Y into observed and unobserved (missing) components. The observed portion is denoted by Y_o while unobserved by Y_m. Rubin (1976) uses a random variable R whose dimensions are the same as of Y, where $R_{ij} = 1$ when element is observed while $R_{ij} = 0$ for unobserved element.

2.8.1.1 MCAR

If the missingness is induced by the simple random sampling design, it is known as missing completely at random constellation. Here, the missing values are independent from the fully observed and the unobserved data, mathematically speaking

$$P(R = 0|Y_o, Y_m) = P(R = 0).$$

2.8.1.2 MAR

The mechanism is missing at random if the missing values of Y depend on fully observed variables, that is, the missing values are associated with observed data. The probability

of missing values rely on observed values as

$$P(R = 0|Y_o, Y_m) = P(R = 0|Y_o).$$

2.8.1.3 MNAR

The third mechanism, MNAR, is not commonly used in practice. In this mechanism, the probability of missingness depends on missing values itself, i.e.

$$P(R = 0|Y_o, Y_m) = P(R = 0|Y_o, Y_m).$$

2.8.1.4 General points

In MCAR, the observed data is considered as random sample and probability of missing values is equal for each case. Any estimate obtained from this data provides unbiased results but a loss in statistical power due to removal of completely missing cases from data. The MAR is correlated with observed data and MNAR is correlated with unobserved data, therefore, the estimates with these mechanisms provide biased results. In MCAR and MAR, the missing data mechanism is ignorable because there is no need to explicitly specify model for missing data mechanism. In MNAR case, the missing data is used directly therefore it is not ignorable. All three mechanisms have different requirements to implement missing data handling methods. We here force on MCAR and MAR.

2.8.2 Complete case analysis

Missing data is a common problem in every survey. The traditional approach to handle is to delete the incomplete cases from the data and fit the model of interest on remaining complete cases, usually known as complete case (CC) analysis. The CC method is a first choice to solve the missing data problem as it is easy and available in almost every statistical software (Kang, 2013) and sometimes available as a default option in many statistical softwares (van Buuren, 2018). The main advantage of this method is that one can apply any standard statistical technique to estimate the parameters of the regression model of interest. This method assumes that fully observed complete cases are a random sub-sample of the original sample. If the missing data is missing completely at random, the complete case analysis provides unbiased estimates of the parameters (Little and Rubin, 2002; Kang, 2013). The disadvantage of this approach is that the statistical power is reduced by removing the missing cases completely from data. Hence, if a large portion is missing from the data, then CC reduces the sample size by removing those cases and we lose all the other information corresponding to missing cases. This method does not

utilize any available auxiliary information in statistical analysis and can not be used where we are interested to fit the regression model and the specific covariates of this regression are missing in such a way that both missing covariates are not observed together like statistical matching problem.

2.8.3 Multiple imputation

Multiple imputation (MI) introduced by Rubin (1987) is a commonly used method to handle general pattern of missing data or missing by survey design. The method of MI uses posterior predictive distribution of the missing data given the observed data to generate multiple ($K > 1$) estimated values against each missing value. One can run standard regression model of interest on each imputed data set, then combine the results to incorporate uncertainty in the imputations (within and between imputations variability). When MI is implemented correctly, it provides asymptotically unbiased and efficient estimates. This approach generally assumes that the data are missing at random, which means the probability of missing values depends only on fully observed values and not on the missing values.

As previously discussed the data matrix is denoted by Y with its observed components Y_o and missing components Y_m. Carpenter and Kenward (2013) describe the general MI procedure in following three steps:

1. The missing values are imputed independently K times using the distribution of missing values condition on observed data, that is, from $f(Y_m/Y_o)$.

2. Since, there are no missing values after imputations, the standard regression model can be fitted on each data set separately. Also, since each missing value is replaced by K different imputed values, the results would vary from one regression model to another.

3. The K independent estimates are combined using Rubin's rules to get an overall estimate. Suppose, the interest is to estimate regression coefficients, say β_k and their variances, say $\text{Var}(\beta_k)$, where $k = 1, ..., K$, then Rubin's rules can be applied as:

$$\hat{\beta}_R = \frac{1}{K}\sum_{k=1}^{K}\hat{\beta}_k,$$

and variance estimator

$$Var(\hat{\beta}_R) \quad = \hat{W} + (1 + \tfrac{1}{K})\hat{B},$$

where

$$\hat{W} = \frac{1}{K}\sum_{k=1}^{K}\hat{\sigma}_k^2 \quad \text{and} \quad \hat{B} = \frac{1}{K-1}\sum_{k=1}^{K}(\hat{\beta}_k - \hat{\beta}_R)^2.$$

The index R indicates the results are obtained using Rubin's rule, and $\hat{W}$ and $\hat{B}$ represent within and between variations from imputations, respectively.

2.8.4 Multivariate imputation by chained equations

There are two general approaches to impute the missing values: joint modeling and fully conditional specification also known as multivariate imputation by chained equations (mice). Schafer (1997) introduces several methods for imputation based on joint models under the conditional multivariate normal and log-linear models. The joint modeling approach specifies a joint model for the partially observed data given the fully observed data. This approach draws the sample values from their posterior predictive distribution (see van Buuren and Groothuis-Oudshoorn, 2011). The approach is suitable when the assumed multivariate distribution describes the data reasonable well.

The mice method separately defines the model for each missing variable, conditioned on the other observed variables. This approach may fail to specify a valid multivariate distribution due to a series of univariate conditional distributions. A nice, and perhaps, an important feature of this approach is that the researcher can model each missing variable according to its own distribution. For example, if a binary covariate is missing, the researcher can use logistic regression to impute the missing values and if a continuous covariate is missing, a normal regression model can be used.

Suppose the data of a continuous covariate x is fully observed and the categorical covariates z are partially observed, and z is a vector of covariates i.e. $z_1, z_2, ..., z_q$. The mice procedure first imputes the missing values of that covariate which has least missing values, then imputes for the second least and so on. If all the missing covariates have same number of missing values, like in missing by survey design, then this algorithm takes the order sequence to impute the missing values. For example, the covariate z_1 will be regressed on observed variable x and predictive values of z_1 are drawn randomly from the posterior predictive distribution given observed values of z_1. After filling out z_1, the covariate z_2 will be imputed by running the regression model for z_2 on both x and z_1 simultaneously and

predictive values of z_2 are drawn randomly, similar to z_1, and so on, for all the remaining missing covariates in z. This procedure provides one complete imputed data set. If this procedure is repeated multiple times independently, we will obtain multiple imputed data sets, we refer to van Buuren (2007; 2011) for details.

2.8.5 Tree-based imputation methods

There are two popular tree based methods to impute missing values with mice: one is classification and regression trees (CART), and the other is random forest. CART (Breiman et al., 1984) is a widely used technique in statistics and machine learning. CART models look for the cut point on predictors, which are used to split the sample into homogeneous two sub-samples. The procedure to split the sample is repeated on both splitted samples, it make a series of binary tree. If the variable of interest is discrete then classification tree is used to identify the most suitable class to fall of a target variable values. When the variable is continuous, regression tree is used.

Some features of CART make it an attractive imputation engine. For example, this approach is less restricted, that is, it can be flexibly used to fit nonlinear relations without taking into account the parametric assumptions. The literature shows that application of CART mice provides more reliable inferences as compared to just default mice settings (see Burgette and Reiter, 2010; Doove et al., 2014; van Buuren, 2018). CART can be used alongside mice, as default mice settings for binary missing variable is logistic regression and for continuous missing variable, the predictive mean match can be used. CART can be fast and effective imputation engine for categorical variables if they have a limited set of levels (see Akande et al., 2017).

Sometimes CART results are overfitting with low bias and high variance (Burgette and Reiter, 2010). This problem can be solved by using random forest technique, which is an extension of CART. Random forest is a non-parametric technique, which is very attractive to use when CART or the standard mice is no more feasible to impute missing values. For example, Shah et al. (2014) use random forest with mice and their estimates are more efficient and less biased than standard mice procedure. Random forest generates several trees and each tree is based on an independent and randomly drawn sub-sample. As the number of trees increases in forest, the random errors become small by taking the average of the trees. To avoid the overfitting trend of any single tree, random forest combines the results across the trees. The predictions have less variations as compared to single tree based method (Breiman, 2001).

3 Contributions

This Chapter is dedicated to the summary of our three contributions, the first two are based on two phase sampling and the third one on a split questionnaire design (SQD). In contribution 1, we propose a sampling scheme to select an efficient second phase random sample which provides smallest estimation variability for the coefficients of regression model. Contribution 2 deals with the situation where some expensive covariates are missing by survey design. We propose an approximate estimation approach using semi-parametric regression for two phase sampled data. In contribution 3, a SQD is considered in the context of statistical matching where common variables are available for all two samples while some specific variables are observed only in one specific sample and the other specific variables on second sample. To estimate the regression model of interest, we assume that the specific variables are conditionally independent given common variables.

3.1 Contribution 1

In this contribution, we consider two phase sampling to collect the data of the rents survey of Munich city in Germany. The first phase sample s_1 contains response variable Y and two continuous covariates x_1 and x_2 while a vector of categorical covariates z is missing, which is expensive to measure. Here, Y is the rent per square meter (in Euro) of an apartment, x_1 is its floor space in square meters and x_2 is its year of construction. The categorical covariates z describe the qualities and facilities of the apartment and are observed in second phase sample s_2 only. The primary goal is to fit a regression model Y on $x = (x_1, x_2)$ and z. we additionally assume that the survey is conducted repeatedly, that is the survey drawn regularly every two years and the data of Y, x and z are available from the previous surveys or past studies as well and are denoted by s_p.

The problem addressed in this contribution is how to select the second phase random sample which provides smallest possible estimation variability for the coefficients of regression model. We propose a simple three steps procedure to select an efficient second phase random sample. In first step, we combine sample s_1 with previous sample s_p to get a large data set, note that covariates z are missing form first phase sample s_1 but are available for s_p. In second step, we simulate/predict the missing values of covariates z for s_1 using R package `mice`, (see van Buuren and Groothuis-Oudshoorn 2011). And in the final step, a random sample s_2 is drawn from the first phase sample s_1, $s_2 \subset s_1$ and calculate matrix norm as

$$M := || \sum_{j \subset s_2} W_j^T W_j || \tag{13}$$

where W has columns $w_j = (x_j, z_j^*)$ for $j \subset s_2$, and x and z^* correspond to observed and imputed covariates in second phase sample, respectively. Now draw the second phase simple random sample repeatedly (1000 times) from the first phase imputed sample and calculate (9) for each sample. The sample which has maximum norm is our final selected sample $s_{2,b}$. With sample $s_{2,b}$, we obtain the information of expensive covariates z through questionnaires. The sample selected through this procedure ensures minimum possible variance of regression coefficients as shown in (3).

A simulation study is conducted to see performance of the proposed sampling procedure. We generate six dimensional multivariate normal data and use standard linear regression model as described in (**??**) to get the values for continuous response variable. We notate here four binary covariates with z, which are correlated with two continuous covariates x. The binary covariates are obtained using the R package `Binnor` (Demirtas, et al., 2014). The second phase random sample is selected with the proposed scheme. We compare the variances of regression coefficients from the data of second phase sample. The proposed method performs better than the standard simple random sample both in simulation study and the rent data example.

3.2 Contribution 2

The contribution 2 considers the scenario where some expensive covariates are missing at the planning stage of a survey and they are observed only for sub-sample to estimate the average rent of the apartments consider Munich rental guide data. The idea is similar to contribution 1 (two phase sampling) to collect the cheap variables in first phase sample s_1 and draw a second phase sample s_2 from s_1 to collect the information of expensive covariates. In first phase sample s_1, the response variable Y, rent per square meter (in Euros) and a covariate x, the floor space in square meters are observed through telephone while the expensive covariate z are missing. In second phase, a random sample s_2 is drawn from the first phase sample s_1 and covariates z, qualities and facilities of the apartment, are also recorded by personal interviews through questionnaires. This is the two phase nested sampling design, that is $s_2 \subset s_1$. The random sampling (simple random sampling) do not utilize available information of first phase sample to select the second phase sample. One can draw the second phase sample using the information of first phase sample. We use two types of unequal probability to select the second phase sample. The first one

is covariate dependent sampling and the second one is residuals dependent sampling. In first case, covariate x from first phase sample is used to select the second phase sample, while in the second case, we run the regression model of Y on x using s_1 information and residuals of this model are used to select the second phase sample. The first unequal probability design is non-informative sampling design because information of Y is not used while the later one is informative design because the response variable Y is included through residuals to select second phase sample s_2.

The question addressed in this contribution is how to use the available data to fit the regression model which provides the minimum mean squared prediction error and regression coefficients have smaller bias and estimated variance. We propose an approximate estimation approach using non-parametric mean and variance regression Y and x only and a semi-parametric mean regression model of Y on x and z. The response variable Y is non linearly related with x and linearly with z, hence the model is semi-parametric due to non-parametric component included x and linear effect with z. The idea extends the approach of Little (1992) towards non-normal data and non-linear regression. Little (1992) assumes that the response variable Y and covariates x are observed for large sample data while covariates z are only available for small sample. He assumes multivariate normality for joint distribution of Y and z given x and estimate as

$$f(y, z|x) = f(y|x)f(z|y, x). \tag{14}$$

We do not assume multivariate normality for $f(y, z|x)$ and instead use an approximate normal distribution. If we condition on z then conditional distribution of Y given x and z is

$$f(y|x, z) := \frac{f(y, z|x)}{f(z|x)} = f(y|x)\frac{f(z|y, x)}{(z|x)}. \tag{15}$$

We can transform the ratio in (11) to

$$\frac{f(z|y, x)}{f(z|x)} = \frac{f(y, z|x)}{f(z|x)f(y|x)} = \frac{f(y|z, x)}{f(y|x)}. \tag{16}$$

The (11) and (12) enable us to estimate the three regression models separately with available data and there is no need to impute the missing values of covariates z. First, we fit the non-parametric regression model of Y on x with heteroscedastic error term using data of $s_1 \setminus s_2$ and secondly, the same model is fitted for the sample s_2. The last model is run as Y on x and z for sample s_2 with homoscedastic error term. Note that, with the heteroscedastic model Y on x, we have induced an interaction between x and z. The penalized splines smoothing technique is used to estimate these models. The proposed

method is simple and easy to apply in practice. For example, one can use the function like the `gamlss()` to fit the models (see Stasinopoulos et al., 2017).

Simulation studies are conducted to compare the proposed method with some existing alternatives such as complete case analysis and multiple imputations methods. The continuous response variable is generated using semi-parametric model (**??**), where the continuous covariate is related non linearly with response while categorical covariates have linear relation with the response variable. We assume homoscedastic as well as heteroscedastic error terms for regression model in simulation studies. We use cross-validation to get prediction error of the fitted regression model. The population is divided into two parts, one is considered as train data to fit the regression model and second part is considered as test data to calculate the out of sample prediction error. The simulation results provide minimum mean squared prediction error for the proposed routine as compared to alternatives methods. The ratio prediction error deviates only with five percent margins as compared to complete case, multiple imputation methods and the hypothetical case where we assume all information on x and z is available for the first phase sample. We include the multivariate metrical variables in non-parametric part and results turn out to be same as for univariate case. The bias and estimated variance of the regression coefficients are also reported for simulated data. The proposed routine provides smaller values of bias and variance as compared to the alternative methods. The application of the approximate routine on the rent data example produces smaller predication error as compared to the alternatives routines.

3.3 Contribution 3

In contribution 3, we propose a split questionnaire design in the context of statistical matching for rents survey of Munich. We split a long questionnaire into two non-overlapping parts, that is, draw two independent random samples s_a and s_b from the same population. The response variable Y, rent per square meter, and the continuous covariate x, the floor space are observed in both the samples while the categorical covariates z and w, representing qualities and facilities of an apartment are recorded in such a way that z are collected only in sample s_a while w in sample s_b only. There is no sampling unit which has the information about both z and w simultaneously. To integrate two or more such samples is usually known as statistical matching problem. Statistical matching can be considered as a special missing data problem.

The objective of this contribution is to estimate regression model Y on x, z and w with split questionnaires data to get minimum mean squared prediction error compared to

alternatives methods, e.g., multiple imputations. We want to estimate the joint densities $f(y, x, z, w)$ and $f(x, z, w)$ but there is identification problem. To estimate both the distributions and overcome the identification problem, we assume the conditional independence of z and w. We condition on Y and x to estimate the joint density $f(y, x, z, w)$ and condition on x only for joint density $f(x, z, w)$. With this assumption and using chain rule, we can factorize the joint distribution into different small components as shown in (8). We use three regression models with same response variable for each model. First, we estimate the regression model Y on x and z using sample s_a, the second regression model Y on x and w is estimate with sample s_b and the third regression model is Y on x which uses information from both the samples. And finally, the separate estimates are combined to get the overall estimates of regression model Y on x, z and w. Note that numerically the proposed procedure is very simple and straightforward. It does not require to specify a distributional model for the covariates nor multiple imputation is needed.

To demonstrate the performance of the proposed approach in comparison to imputation alternatives, simulation studies are conducted. We use standard linear regression model (2) to generate the continuous response variable with error term following normal distribution with zero mean and constant variance. The continuous covariate x is generated with different distribution, e.g., uniform, log-normal and normal. The standardized form of this covariate is used in logistic model to obtain specific categorical covariates z and w. To compare the proposed routine with alternative methods, `mice` Package in R is used. To impute the missing values, mice assumes conditional independence between specific missing covariates given common variables. Three multiple imputations methods are used to impute the missing values (mice default setting, CART and random forest).

A simple random sample from a population is selected containing all variables. Further, we consider the data in this sample as a full questionnaire where we hypothetically assumed that there exists no missing values among all covariates. Two random samples are drawn according to the proposed method and the data of common variables and specific covariates is obtained. To compare the simulation results with alternative methods, we used cross-validation as described in contribution 2. That is, a part of the population data is used to fit the regression model and the other part of population is used to obtain the out of sample prediction error. The proposed method tends to produce smaller median and standard deviation compare to multiple imputation methods. We also calculated the bias and root mean squared error of regression coefficients. The results are again in support of the proposed routine. Application of the proposed method on rent data example also provides minimum mean squared prediction error as compared to the alternative methods.

4 Concluding remarks

This cumulative dissertation considers two different survey designs when some (expensive) covariates are missing by design. The first design involves the idea of two phase sampling and the second is based on a split questionnaire survey design in the context of statistical matching. Both the designs are illustrated through simulation studies as well as using an example of Munich rents data.

The contribution 1 is based on two phase sampling with an additional assumption that the survey is drawn repeatedly over time so that data from past can be utilized. We have shown that the proposed method to draw second phase random sample provides smallest estimation variability for the regression coefficients as compared to a standard simple random sample. This contribution helps to determine which apartment (sample unit) should be selected for second phase sample to collect information about expensive covariates through questionnaires.

In contribution 2, the proposed method extends the idea of Little (1992) towards categorical covariates z and non linear effect of continuous covariate(s) x. Penalized splines smoothing technique is applied to estimate the proposed semi-parametric regression model. The proposed method provides a small mean squared prediction error than the competing commonly used missing data methods, such as, complete case analysis and multiple imputations via full conditional specification. With this proposal, considerable resources can be saved for rental guide survey. For example, the administrative department of Munich collects data of nearly 3000 apartments through personal interviews based on questionnaires. Applying the proposed routine with two phase sampling, a relatively big amount of data can be collected through telephone at first phase and only a small number of questionnaires are needed to collect second phase data through personal interviews by questionnaires.

In contribution 3, we propose a split questionnaire survey design to collect data for Munich rental guide. Through splitting a long questionnaire, each participant needs to answer only a fraction of total questions while some common questions are asked from everyone. Splitting the long questionnaire reduces the consumption of resources as well as the respondent does not feel burden to fill short questionnaire. Since, information of specific variables z and w is not obtained simultaneously from a participant it causes an identification problem to estimate the joint distribution. To overcome this problem, we assume conditional independence of specific variables given common variables. This is a strong assumption and is not testable with available data. If this assumption does not hold true, the estimates would be biased. However, if the common variables are closely related to

the specific variables, as in the example of Munich survey for the rents, then this is a reasonable assumption. In this example, covariate x, the floor space, is related to the covariates z and w, qualities and facilities of apartment. It makes sense, because if an apartment has large floor size, it is more likely that this apartment will also have other facilities as well.

To split the questionnaire, we use only two independent samples (questionnaires). The samples are simple random samples, however, the unequal probability sampling can be used like Kim et al. (2016) proposed the idea of a split questionnaire survey design in the context of statistical matching. They proposed to select a random sample and considered it as auxiliary information and used this information to select two non-overlapping sub-samples.

The present study limited itself to the Munich survey for the rents and consider that only categorical covariates are missing, however, the proposal for missing by survey design can be extended to a case when continuous coavraites are missing, to check and compare the performance of proposed methods.

References

Akande, O., Li, F., and Reiter, J. (2017). An empirical comparison of multiple imputation methods for categorical data. *The American statistician* 71(2):162–170.

Boor, C.D. (2001). *A Practical Guide to Splines.* New York: Springer.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and regression trees.* Wadsworth Publishing, New York.

Breiman, L. (2001). Random forests. *Machine learning* 45(1):5–32.

Burgette, L.F. and Reiter, J.P. (2010). Multiple imputation for missing data via sequential regression trees. *American Journal of Epidemiology* 172(9):1070-1076.

Carpenter, J.R. and Kenward, M.G. (2013). Multiple Imputation and its Application. John Wiley and Sons, Ltd.

Chipperfield, J.O. and Steel, D.G. (2009). Design and estimation for split questionnaire surveys. *Journal of Official Statistics* 25(2):227-244.

Chipperfield, J.O., Barr, M.L., and Steel, D.G. (2018). Split questionnaire designs: collecting only the data that you need through MCAR and MAR designs. *Journal of Applied Statistics* 45(8):1465-1475.

Cole, T.J. and Green, P.J. (1992). Smoothing reference centile curves: the LMS method and penalized likelihood. *Statistics in medicine* 11(10):1305–1319.

Demirtas, H., Amatya, A., and Doganay, B. (2014). Binnor: An R package for concurrent generation of binary and normal data. *Communications in Statistics-Simulation and Computation* 43(3):569-579.

Derkach, A., Lawless, J.F., and Sun, L. (2015). Score tests for association under response-dependent sampling designs for expensive covariates. *Biometrika* 102(4):988–994.

Doove, L.L., van Buuren, S., and Dusseldorp, E. (2014). Recursive partitioning for missing data imputation in the presence of interaction effects. *Computational Statistics & Data Analysis* 72:92–104.

D'Orazio, M., Di Zio, M., and Scanu, M. (2006a). *Statistical Matching: Theory and Practice.* Chichester, United Kingdom: Wiley. https://doi.org/10.1002/0470023554.

D'Orazio, M., Di Zio, M., and Scanu, M. (2006b). Statistical matching for categorical data: Displaying uncertainty and using logical constraints. *Journal of Official Statistics* 22:137–157.

Eilers, P.H.C. and Marx, B.D. (1996). Flexible smoothing with B-splines and penalties. *Journal of Statistical Science* 11(2):89–121.

Fahrmeir, L., Kenib, T., Lang, S., and Marx, B. (2013). *Regression-Models, Methods and Applications.* Springer.

Fitzenberger, B. and Fuchs, B. (2017). The residency discount for rents in Germany and the tenancy law reform act 2001: Evidence from quantile regressions. *German*

Economic Review 18(2):212-236.

Graham, J.W., Taylor, B.J., Olchowski, A.E., and Cumsille, P.E. (2006). Planned missing data designs in psychological research. *Psychological Methods* 11(4):323-343.

Hanif, M. and Brewer, K.R.W. (1980). Sampling with unequal probabilities without replacement: A review. *International Statistical Review* 48(3):317-335.

Kang, H. (2013). The prevention and handling of the missing data. *Korean Journal of Anesthesiology* 64(5):402–406.

Kauermann, G. and Ali, M. (2020). Semi-parametric regression when some (expensive) covariates are missing by design. *Statistical Papers* https://doi.org/10.1007/s00362-019-01152-5

Kim, J.K. and Skinner, C.J. (2013). Weighting in survey analysis under informative sampling. *Biometrika* 100(2):385-398.

Kim, J., Berg, E., and Park, T. (2016). Statistical matching using fractional imputation. *Survey Methodology* 42(1):19-40.

Kneib, T, Konrath, S., and Fahrmeir, L. (2011). High dimensional structured additive regression models: Bayesian regularization, smoothing and predictive performance. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 60(1):51– 70.

Little, R.J.A. (1992). Regression with missing X's: A review. *Journal of the American Statistical Association* 87(420):1227-1237.

Little, R.J. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data.* (2nd edition), Wiley. https://doi.org/10.1002/9781119013563.

Lumley, T. (2017). Robustness of semi-parametric efficiency in nearly-true models for two-phase samples. e-print arXiv:1707.05924

Mandallaz, D., Breschan, J., and Hill, A. (2013). New regression estimators in forest inventory with two phase sampling and partially exhaustive information: a design based monte carlo approach with applications to small area estimation. *Canadian Journal of Forest Research* 43(11):1023–1031.

McIsaac, M.A. and Cook, R.J. (2014). Response-dependent two-phase sampling designs for biomarker studies. *Canadian Journal of Statistics* 42(2):268–284.

McIsaac, M. (2013). Statistical methods for incomplete covariates and two-phase designs. PhD Thesis. http://hdl.handle.net/10012/7259

Moriarity, C. and Scheuren, F. (2001). Statistical matching: A paradigm for assessing the uncertainty in the procedure. *Journal of Official Statistics* 17:407-422.

Moriarity, C. and Scheuren, F. (2003). A note on Rubin's statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business and Educational Studies* 21:65–73.

Neyman, J. (1938). Contribution to the theory of sampling human populations. *Journal of American Statistical Association* 33(201):101-116.

O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems. *Journal of Statistical Science* 1(4):502–518.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review* 61(2):317-337.

Pfeffermann, D. (1996). The use of sampling weights for sampling data analysis. *Statistical Methods in Medical Research* 5:239-261.

Pfeffermann, D. and Sverchkov, M. (2009). Inference under informative sampling. *Handbook of Statistics; Sample Surveys: Inference and Analysis 29.*

Pfeffermann, D. (2011). Modelling of complex survey data: Why model? Why is it a problem? How can we approach it? *Survey Methodology* 37(2):115-136.

Peytchev, A. and Peytcheva, E. (2017). Reduction of measurement error due to survey length: Evaluation of the split questionnaire design approach. *Survey Research Methods* 11(4):361-368.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Raghunathan, T.E. and Grizzle, J.E. (1995). A split questionnaire survey design. *Journal of the American Statistical Association* 90(429):54-63.

Rässler, S. (2002). *Statistical Matching: A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches.* New York: Springer. https://doi.org/10.1007/978-1-4613-0053-3.

Rässler, S., Koller, F., and Mäenpää, C. (2002). A split questionnaire survey design applied to german media and consumer surveys. *Proceedings of the International Conference on Improving Surveys,* ICIS, Copenhagen.

Rässler, S. (2004). Data fusion: Identification problems, validity, and multiple imputation. *Austrian Journal of Statistics* 33(1–2):153–171.

Rendall, M.S., Dastidar, B.G., Weden, M.M., Baker, E.H., and Nazarov, Z. (2013). Multiple imputation for combined-survey estimation with incomplete regressors in one but not both surveys. *Sociological Methods and Research* 42(4):483–530.

Rigby, R.A. and Stasinopoulos, D.M. (1996a). A semi-parametric additive model for variance heterogeneity. *Statistics and Computing* 6:57–65.

Rigby, R.A. and Stasinopoulos, D.M. (1996b). Mean and dispersion additive models. In Statistical Theory and Computational Aspects of Smoothing (eds W. Härdle and M. G. Schimek), pp. 215–230. Heidelberg: Physica.

Rigby, R.A. and Stasinopoulos, D.M. (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 54(3):507–554.

Roszka, W. (2015). Some practical issues related to the integration of data from sample

surveys. *Statistika: Statistics and Economy Journal* 95(1):60-75.

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

Rubin, D.B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business and Economic Statistics* 4(1):87-94.

Rubin, D.B. (1987). *Multiple imputation for nonresponse in surveys.* New York: John Wiley & Sons.

Ruppert, D., Wand, M.P., and Carroll, R.J. (2003). *Semiparametric Regression.* Cambridge University Press.

Ruppert, D., Wand, M.P., and Carroll, R.J. (2009). Semiparametric regression during 2003-2007. *Electronic Journal of Statistics* 3:1193-1256.

Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data.* New York: Chapman and Hall.

Shah, A.D., Bartlett, J.W., Carpenter, J., Nicholas, O., and Hemingway, H. (2014). Comparison of random forest and parametric imputation models for imputing missing data using mice: A CALIBER study. *American Journal of Epidemiology* 179(6):764-774.

Shoemaker, D.M. (1973). *Principles and Procedures of Multiple Matrix Sampling.* Cambridge, MA: Ballinger Publishing Company.

Stasinopoulos, D.M., Rigby, R.A., Heller, G.Z., Voudouris, V., and De Bastiani, F. (2017). *Flexible regression and smoothing: Using GAMLSS in R.* Chapman and Hall/CRC.

Stuart, M. and Yu, C. (2019). A computationally efficient method for selecting a split questionnaire design. *Communications in Statistics - Simulation and Computation* https://doi.org/10.1080/03610918.2019.1697819

Tille, Y. (2006). *Sampling Algorithms.* Springer

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. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software* 45(3):1-67.

van Buuren, S. (2018). *Flexible Imputation of Missing Data.* Second edition, Chapman and Hall/CRC.

White, L.R. and Carlin, J.B. (2010). Bias and efficiency of multiple imputation compared with complete case analysis for missing covariate values. *Statistics in Medicine* 29(28):2920-2931.

Wood, S.N. (2017). *Generalized additive models - An introduction with R.* Second edition, CRC Press.

Zhang, Z. and Rockette, H.E. (2005). On maximum likelihood estimation in parametric regression with missing covariates. *Journal of Statistical Planning and Inference* 134(1):206-223.

Zhao, Y., Lawless, J.F., and Mcleish, D.L. (2009). Likelihood methods for regression

models with expensive variables missing by design. *Biometrical Journal* 51(1):123-136.

Zhou, H., Chen, J., Rissanen, T.H., Korrick, S.A., Hu, H., Salonen, J.T., and Longnecker, M.P. (2007). Outcome-dependent sampling: an efficient sampling and inference procedure for studies with a continuous outcome. *Epidemiology* 18(4):461–468.

Attached contributions

A.1:

Ali, M. and Kauermann, G. (2021a). Second Phase Sample Selection for Repeated Survey. Technical Report 237, Department of Statistics, LMU Munich.

Available online at: https://epub.ub.uni-muenchen.de/74729/

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INSTITUT FÜR STATISTIK

Mehboob Ali
Göran Kauermann

Second Phase Sample Selection For Repeated Survey

Technical Report Number 237, 2021
Department of Statistics
University of Munich

http://www.statistik.uni-muenchen.de

Second Phase Sample Selection For Repeated Survey

Mehboob Ali[1], **Göran Kauermann**[2]

Abstract

The paper describes the scenario of a survey where a relatively large random sample is drawn at a first phase and a response variable Y and a set of (cheap) covariates x are observed, while (usually expensive) covariates z are missing. In a second phase, a smaller random sample is drawn from the first phase sample where the additional covariates z are also recorded. The overall intention is to fit a regression model of y on both, x and z. The question tackled in this paper is how to select the second phase random sample. We assume further that the survey is drawn repeatedly over time, that is data on Y, x and z are available from previous studies. As example for such setting we consider rental guide surveys, regularly run in German cities. We propose to draw the second phase sample such that it minimizes the estimation variability in the underlying regression model. This step is carried out with imputation using the previous survey data. The norm of matrix can be used to find simulation based second phase sample which maximize design matrix of imputed data. The proposed sampling scheme is numerically rather simple and performs convincingly well in simulation studies as well as in the real data example.

Key Words: Two phase sampling; Repeated survey; Rental guide survey; Matrix norms

[1]Department of Statistics, Ludwig-Maximilians-University Munich, Germany. E-mail: mehboob.ali@stat.uni-muenchen.de

[2]Department of Statistics, Ludwig-Maximilians-University Munich, Germany. E-mail: goeran.kauermann@stat.uni-muenchen.de

1. Introduction

Assume we want to draw a survey where some of the quantities are cheap and easy to obtain while others are time consuming and/or expensive. We plan to use a two phase sampling scheme for data collection, following the aim to select an efficient second phase sample using the collected information of first phase. In the first phase, information on the inexpensive quantities is obtained from a large numbers of sampling units. Then, a subset of sampling units are drawn in a second phase sample from the first sample and the expensive covariates are also recorded. As example we consider rental guide surveys which are regularly run in German cities as an official instrument to control the rental market (see e.g. Fahrmeir et al., 2013 or Fitzenberger and Fuchs, 2017). Thomschke (2019) compares the rent for five German cities to explore the effect of official rent constraints. Breidenbach et al. (2019) study the regional variation in rent and elevate 2015 rent control policy for Germany. More recently, Kauermann et al. (2020) discuss about the data collection and sampling of rent index in Germany. They analysis the rent index practice with statistical perspective of the 30 cities in Germany. Their article particularly focus on three main aspects: Firstly, they made comparison of tenant and landlord surveys in order to find out which individuals are likely to be included in the sample frame. Secondly, they discuss the various forms of data collection, i.e. written questionnaires, interviews or combination of both. Lastly, they describe the sampling methods and designs used for rent index practice in Germany. Specially, they discuss the problem of un-availability of complete lists of all apartments related to the rent index. In addition, they describe the way how to get complete list of all households/apartments relevant to the rent index of Munich and draw first phase random sample from this list, i.e. the first phase sample can be drawn from residents' registration office of Munich.

We here extend the work of Kauermann et al. (2020) and answer the question how to select second phase random sample from first phase sample which provides smallest estimation variability for the regression model of interest. In our example, we consider the rental guide surveys for Munich only. The following quantities are easily obtained through a simple survey: the rent y (Euro per square meter), floor space x_1 (square meter) and year of construction x_2. These quantities are observed through the first phase sample, which

in Munich is carried out through a telephone survey. In a second phase sample additional quantities about quality and facilities of the apartment are investigated. The apartment facilities are recorded based on a personal interview, which apparently is time consuming and expensive. The overall goal is to fit a regression model

$$Y = x\beta_x + z\beta_z + \varepsilon, \tag{1}$$

where $x = (x_1, x_2)$ and $z = (z_1, ..., z_q)$ is the vector of covariates describing quality and facilities of the apartment. Let $w = (x, z)$ denote the joint vector of covariates of the design matrix for model (1). Applying ordinary least squares (OLS) give us

$$\hat{\beta} = (W^T W)^{-1} W^T Y, \tag{2}$$

where W is the design matrix with rows (x, z) of the second phase sample. The variance of $\hat{\beta}$ equals

$$\sigma^2 (W^T W)^{-1} \tag{Var}$$

and we intend to draw the second phase sample such that $(W^T W)$ is large (or even maximal) leading to a small variance of $\hat{\beta}$ (Imbriano, 2018).

We additionally assume that the survey is drawn repeatedly meaning that we have data of previous surveys on x and z (and y) available. When survey data are taken for the same population at different time this is commonly known as repeated surveys (Steel and McLaren, 2008). Scott and Smith (1974) discuss general terms of both, overlap and non-overlap surveys where they assume a time series models for the repeated surveys. More recently, Ismail et al. (2018) use time series methods for repeated surveys. The problem relate to the design and analysis of repeated surveys over time can be seen in (Duncan and Kalton, 1987).

The variance of the estimator can be reduced using past information available from previous data (Haslett, 1986; Steel and McLaren, 2008). Quality

and Tille (2008) proposed a method which accounts for sampling design. Kott (1994) uses linear regression on repeated survey data and estimates the variance of this fitted model coefficients under two cases, first when the primary sampling units (PSU) are the same across the survey time and secondly, when the PSU are not the same across the survey periods. Fuller (1990) reviewed least squared estimation for repeated surveys in which a portion of units are sampled at more than one time point.

A repeated survey is mostly run with regular frequency, for example monthly, quarterly, or annually. If a repeated survey is conducted at regular intervals, it is generally known as periodic survey (Duncan and Kalton, 1987). Usually, repeated sampling is a key reason to measure important changes in a population (Steel and McLaren, 2008). In our example, the rental guide survey is drawn every two years and we use cross sectional data where survey participants are not necessary the same as in previously drawn survey from the same population. This means that we select the participants independently across the time. We use the previous survey data for simulation and imputation of missing covariates values. That is we use observed information of inexpensive covariates of the first phase with the previous survey data to select an efficient second phase sample for the expensive covariates.

The paper is organized as follows. In Section 2, we briefly describe matrix norms and proposed sampling procedure, and sketch how to select the second phase random sample from first phase sample when data of previous survey is also available. In Section 3, we give simulation study and compare the performance of the sampling procedure on simulated data. We also compare the method on a real data example and report the results of the variance of the fitted model for simulated and real data example. Section 4 discusses our findings.

2. Matrix Norms and Sampling Procedure

2.1. Matrix Norms

How to make (W^TW) as large as possible which minimizes variance of $\hat{\beta}$ as given in (Var)? The commonly used method to maximize the matrix is the

norm of matrix (Steinberg, 2005). If W is a real number matrix, then the norm of a matrix is a non-negative number associated with W and have the following properties:

1. $||W|| \geq 0$ and $||W|| = 0$ if and only if the matrix $W = 0$,

2. $||hW|| = |h|.||W||$, for any scalar h,

3. $||W + U|| \leqslant ||W|| + ||U||$, where U is also a matrix like W,

4. $||WU|| \leqslant ||W||.||U||$.

The size of the matrix can be measure using any norm of W matrix and this size provides some useful information of design matrix in regression analysis (Horn and Johnson, 1990; Yuan, 2020).

2.2. Sampling Procedure

Let the population be indexed by $1, ..., N$ from which we draw the first phase sample $s_1 \subset \{1, ..., N\}$. The question tackled in this paper is how to draw a second phase sample $s_2 \subset s_1$ such that the fitted model (1) has small estimation variability. As motivated in the introduction we look at the case that the survey is drawn repeatedly. Assume therefore that we have data on y, x and z from a previous survey. This means that we have a sample s_p from a previous time-point of the population. Particularly we have data (x_j, z_j) for $j \subset s_p$. These data can be used to estimate the distribution function $F_p(x, z)$, where index p refers to the previous time point. Note that sample s_2 at the current time point should be drawn such that

$$\int ||w^T w|| dF(x, z)$$

where $w = (x, z)$ and $||.||$ stands for some matrix norm. The idea is to use the estimate of $F_p(,)$ as estimate of $F(,)$. Note that with sample s_1 we have already

drawn information about x, so that we condition on sample s_1 and consider the marginal distribution of x as given through the empirical distribution in sample s_1. That is we aim to maximize

$$\sum_{i \subset s_1} \int ||w^T w|| dF_1(x_i, z)$$

where $F_1(x, z)$ is the distribution function with marginal $F_1(x) = \frac{1}{n} \sum_{i \subset s_1} 1\{x_i \leq x\}$. Based on the observed x values in sample s_1 we can predict (or simulate) the corresponding z value using the previous year distribution $F_p(x, z)$. Numerically this can be done in three steps. First we pool the samples s_p and s_1 leading to the large data set where x and y is observed for all pooled observations while z has missing values for all data from sample s_1. This is sketched in Figure 1. As second step we drop column y and apply single imputation for z using the entire pooled data set and use the R package `mice` for imputation, (see Van Buuren and Groothuis-Oudshoorn, 2011). As third step we draw a simple random sample s_2 out of s_1 and calculate

$$M := || \sum_{j \subset s_2} W_j^T W_j || \tag{3}$$

where W has columns $w_j = (x_j, z_j^*)$ for $j \subset s_2$. We repeat this step B times leading to B samples $s_{2,b}$ with $b = 1, ..., B$. For each sample we calculate M_b from (3) for the bth leading to $M_1, ..., M_B$. We then propose to take sample $s_{2,b}$ that maximizes (3), that is take sample $s_{2,b}$ with $b =$argmax$\{M_l, l = 1, ..., B\}$. This sample provides a simulation based small variance, if we take the previous survey distribution of x and z into account.

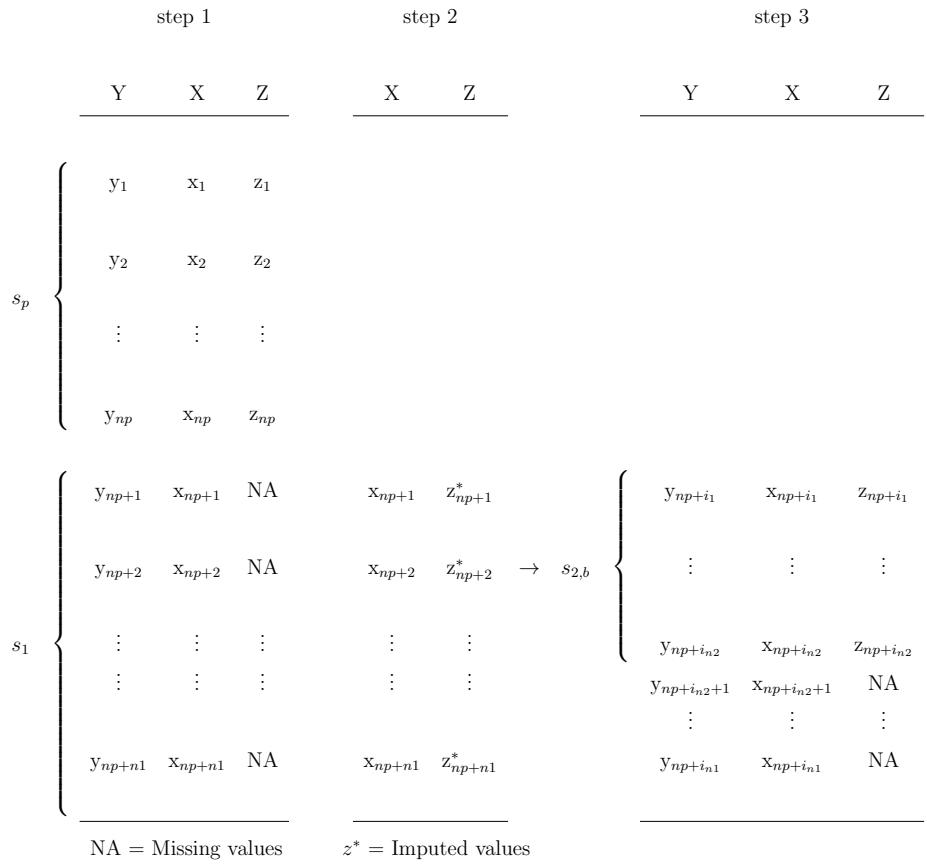

Fig. 1. Sketch of sampling procedure

3. Simulation and Example

3.1. Simulation

We run a simulation study to demonstrate the performance of our sampling scheme. To do so we simulate data from the model

$$Y = x\beta_x + z\beta_z + \varepsilon$$

where $\varepsilon \sim N(0, 1.5)$ and $z = (z_1, z_2, z_3, z_4)$ is a vector of binary covariates which are correlated with vector $x = (x_1, x_2)$. We generate 10000 values as super-population. The parameters values are $\beta_x \in \{2.1, 1.58\}$, $\beta_z \in$

$\{1.33, 0.90, -1.38, 0.82\}$, $\beta = (\beta_x, \beta_z)$ and $n_2 \in \{300, 600\}$. We simulate six dimensional multivariate normal data $(x_1, x_2, \tilde{z_1}, \tilde{z_2}, \tilde{z_3}, \tilde{z_4})$ such that the marginal distributions of $x_1 \sim N_1(20, 5)$, $x_2 \sim N_2(80, 10)$ and the entire vector has the correlation structure

$$R = \begin{pmatrix} 1 & & & & & \\ 0.45 & 1 & & & & \\ 0.55 & 0.50 & 1 & & & \\ 0.45 & 0.50 & 0.30 & 1 & & \\ 0.45 & 0.50 & 0.19 & 0.25 & 1 & \\ 0.45 & 0.45 & 0.40 & 0.21 & 0.19 & 1 \end{pmatrix}$$

In the next step we dichotomize $\tilde{z_1}$, $\tilde{z_2}$, $\tilde{z_3}$ and $\tilde{z_4}$ such that

$$p(z_1 = 1) = p(\tilde{z_1} \leq \mu_1) = 0.2$$

where μ_1 is an arbitrary threshold value of $\tilde{z_1}$ and probabilities for z_2, z_3, z_4 are 0.3, 0.4 and 0.5 respectively. We use the R package `Binnor` (Demirtas, Amatya and Doganay, 2014).

To apply the sampling scheme to simulated data, we select a simple random sample s_1 of size $n_1 = 3000$ from a super-population and observe a response variable Y and covarites x whereas, covariates z are missing. We consider this sample as sample s_1. Sample s_p is drawn accordingly with $n_p = 3000$ from the same super-population and observe a response variable Y, covarites x and z.

In order to select a sample $s_{2,b}$, we impute the missing z values for the first phase sample by combing s_1 with s_p and chosen 1000 second phase sample of size $n_2 = 300$. Apply formula (3) on each imputed sample and select $s_{2,b}$. Then model (1) is fitted on this sample and we compare the performance of our proposal with a simple random sample s_2 of size $n_2 = 300$ chosen from s_1. The simulation is repeated 200 times leading to 200 samples of $s_{2,b}$ and s_2. The results of $est.var(\hat{\beta})$ (estimated variance) and $E(var(\hat{\beta}))$ (average of variance) of simulated model coefficients are given in Table 1 and calculated as

$$est.var(\hat{\beta}) = \frac{1}{m}\sum_{i}^{m}(\hat{\beta}_i - \beta)^2 \quad \text{and} \quad E(var(\hat{\beta})) = \frac{1}{m}\sum_{i}^{m} var(\hat{\beta}_i)$$

Table 1. Estimated and average variance of $\hat{\beta}$ for simulated data

	$n_2 = 300$				$n_2 = 600$			
	$est.var(\hat{\beta})$		$E(var(\hat{\beta}))$		$est.var(\hat{\beta})$		$E(var(\hat{\beta}))$	
Covariates	Prop.Me	SRS	Prop.Me	SRS	Prop.Me	SRS	Prop.Me	SRS
x_1	0.0035	**0.0034**	**0.0030**	0.0036	0.0017	0.0017	0.0015	**0.0013**
x_2	0.0016	0.0016	0.0025	**0.0021**	0.0008	0.0008	0.0014	0.0014
z_1	**0.0612**	0.0667	**0.0555**	0.0737	**0.0311**	0.0329	**0.0304**	0.0307
z_2	**0.0548**	0.0572	0.0753	**0.0722**	**0.0274**	0.0281	**0.0417**	0.0473
z_3	**0.0472**	0.0476	**0.0501**	0.0508	**0.0232**	0.0236	**0.0312**	0.0329
z_4	0.0532	**0.0524**	**0.0478**	0.0508	0.0261	**0.0259**	**0.0193**	0.0257

Smallest values when compared Prop.Me with SRS are denoted with bold

where m is number of simulations, β are the true values for our simulated model and $var(\hat{\beta})$ is the model based estimated variance derived from the OLS formula in equation (Var). In our results, "Prop.Me" describes our proposed method and "SRS" shows the standard simple random sample results. It can be seen in Table 1 that under our sampling procedure most of coefficients give less $est.var(\hat{\beta})$ and $E(var(\hat{\beta}))$ amounts compared to the simple random sample. To see the effect of sample size, we increase the second phase sample of size n_2 from 300 to 600. The results are remained in favour of our proposed sampling procedure.

3.2. Rent Data Example

Now we apply our sampling scheme to a real data example. We consider the two rent surveys for the years 2015 and 2017. We label the 2015 data as previous survey and 2017 as current survey. We have data on the rent per square meter (in Euros) for 3024 apartments available for current survey. Besides the floor space and the year of construction we aim to record in the second phase sample the following indicator variables describing the facilities of an apartment: $z_1 = 1$ if the apartment lies in an average residential location, $z_2 = 1$ if the apartment has an open kitchen, $z_3 = 1$ if the apartment has not an upmarket kitchen, $z_4 = 1$ if the apartment lies in an apartment type building, $z_5 = 1$ if there is under floor heating, $z_6 = 1$ if the apartment has the standard central heating, $z_7 = 1$ if the apartment has a good bathroom

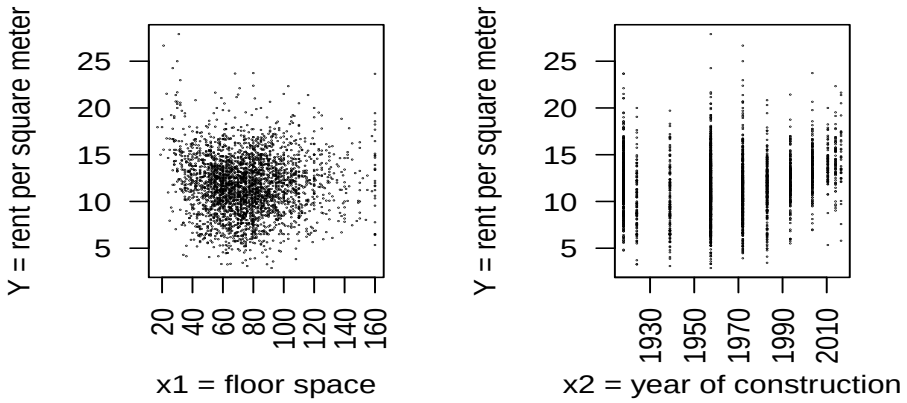

Fig. 2. Rent per square meter relation with floor space (left) and year of construction (right)

equipment, $z_8 = 1$ if the apartment has new floor, $z_9 = 1$ if the apartment has bad floor, $z_{10} = 1$ if the apartment has good floor and $z_{11} = 1$ if the apartment is located in a back premises. In our data we have all variables observed but we pretend now, that measurements $z_1, \ldots, z_{11}$ are missing in sample s_1 and need to be protocolled with sample s_2. The same variables have been recorded in previous survey which contains 3065 apartments data.

The effects of floor space (x_1) and year of construction (x_2) are non-linearly related to the response variable rent per square meter (y) as shown in Figure 2, so we use inverse transformation for x_1 and add a quadratic polynomial additionally for x_2 (see Fahrmeir et al., 2013, Chapter 2). We use the following regression model

$$Y = \frac{1}{x_1}\beta_1 + x_2\beta_2 + x_2^2\beta_3 + z\beta_z + \varepsilon, \tag{4}$$

where ε is a zero mean residual and $z\beta_z$ is a linear predictor from covariates z as described above. The estimates of the complete data (survey 2017) for model (4) are shown in Table 2. The numbers in the table show, for instance that the rent per square meter decrease by 1.2008 for average residential location for z_1.

Table 2. Estimates for rent data

Covariates	Estimate	Std. Error	t value	Pr(>|t|)
$1/x_1$	116.9433	8.4051	13.9134	0.0000
x_2	-1.7018	0.2395	-7.1066	0.0000
x_2^2	0.0004	0.0001	7.1130	0.0000
z_1	-1.2008	0.0967	-12.4146	0.0000
z_2	0.7361	0.1534	4.7987	0.0000
z_3	-1.1764	0.1098	-10.7149	0.0000
z_4	-1.0176	0.1310	-7.7651	0.0000
z_5	1.3634	0.1890	7.2120	0.0000
z_6	0.4154	0.1226	3.3867	0.0007
z_7	1.3857	0.2383	5.8145	0.0000
z_8	1.2091	0.1530	7.9044	0.0000
z_9	-1.0417	0.1761	-5.9155	0.0000
z_{10}	1.2204	0.1467	8.3189	0.0000
z_{11}	0.4932	0.1845	2.6733	0.0076

To measure the performance of the proposed method we consider 3024 apartments available for current survey as a first phase sample (this is a random sample drawn from the population of all apartments in the city or community) and impute the entries on the z covariates for first phase. We select the second phase sample of size $n_2 = 350$ from phase one sample s_1 using our method discussed in Section 2. We repeat this step 1000 times leading to 1000 samples of the second phase sample. For each sample we calculate (3) and select $s_{2,b}$ which maximizes (3) for the imputed sample. We repeat the whole process 100 times leading to 100 $s_{2,b}$ and s_2 samples. We calculate regression estimator variance for model (4) for both sampling methods and the results of their average estimation variation are compared. We can see in Table 3 that our proposed method gives smaller $est.var(\hat{\beta})$ and E($var(\hat{\beta})$), for the rent data example we calculated $est.var(\hat{\beta})$ as

$$est.var(\hat{\beta}) = \frac{1}{m}\sum_{i}^{m}(\hat{\beta}_i - \tilde{\beta})^2$$

where $\tilde{\beta}$ is the estimated values when fitting the model to the 3024 apartments of first phase which are given in second column of Table 2. The analysis on the rent data example is repeated by increasing the second phase sample size to $n_2 = 700$. The results are given in Table 3. We can seen that our proposed

Table 3. Estimated and average variance of $\hat{\beta}$ for rent data

	$n_2 = 350$				$n_2 = 700$			
	$est.var(\hat{\beta})$		$E(var(\hat{\beta}))$		$est.var(\hat{\beta})$		$E(var(\hat{\beta}))$	
Covariates	Prop.Me	SRS	Prop.Me	SRS	Prop.Me	SRS	Prop.Me	SRS
$1/x_1$	**1019.4392**	1090.9109	**652.1014**	664.3837	432.1901	**331.3845**	314.9245	**308.7452**
x_2	0.5348	**0.3575**	**0.4868**	0.5094	**0.1718**	0.2032	**0.2407**	0.2509
x_2^2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
z_1	**0.0702**	0.0769	**0.0803**	0.0828	**0.0249**	0.0287	**0.0398**	0.0409
z_2	**0.1984**	0.2407	**0.1916**	0.2166	**0.0586**	0.0740	**0.0932**	0.1027
z_3	**0.0954**	0.1044	**0.1022**	0.1087	0.0392	**0.0355**	**0.0505**	0.0527
z_4	0.1467	**0.1396**	**0.1529**	0.1548	0.0618	**0.0587**	**0.0746**	0.0754
z_5	0.2640	**0.2502**	**0.2913**	0.3396	0.1145	**0.0652**	**0.1458**	0.1555
z_6	**0.1135**	0.1270	**0.1344**	0.1355	**0.0639**	0.0708	**0.0658**	0.0659
z_7	**0.4034**	0.6750	**0.4407**	0.5232	**0.1737**	0.1930	**0.2219**	0.2612
z_8	**0.2288**	0.2387	**0.2064**	0.2128	**0.0887**	0.1156	0.1024	**0.1015**
z_9	**0.1743**	0.2370	**0.2656**	0.2726	0.0964	**0.0937**	**0.1318**	0.1366
z_{10}	**0.1072**	0.1334	**0.1781**	0.1900	**0.0596**	0.0697	**0.0899**	0.0950
z_{11}	0.3270	**0.3253**	**0.3002**	0.3203	0.1329	**0.1096**	**0.1436**	0.1515

Smallest values when compared Prop.Me with SRS are denoted with bold

sampling procedure give better results as compared to SRS similarly as for $n_2 = 350$

4. Discussion

The motivation of our research comes from a survey on rent for the apartments which is regularly conducted in all the large cities in Germany. The results of this survey are used as an official instrument to control the rent of the apartments. In our real data example, we used data of the rent of the apartments in Munich. The collection of this data through long questionnaire is expensive and time consuming. This suggests to use the two phase sampling. As Kauermann et al. (2020) described the first phase sample can be drawn from residents' registration office of Munich. We proposed that the second phase sample can be selected by the method of imputations. The missing values in the first phase sample can be imputed using previous time survey data and finding norm of design matrix from imputed sample to obtain minimum variance of the regression coefficients.

The proposed sample selection procedure is easy to apply in practice. It is shown in simulation and in a real data example that the idea of using information available in previous survey with first phase data can be more helpful to obtain the second phase sample which provides a simulation based lower variance of $\hat{\beta}$ as compared to s_2 (which is a standard simple random sample). Our proposed sampling scheme can be used for the efficient selection of simulation based second phase sample, if information from previous studies is available.

References

Breidenbach, P., Eilers, L., Fries, J., 2019. Rent control and rental prices: High expectations, High effectiveness? German Council of Economic Experts. Working Paper 07/2018.

Demirtas, H., Amatya, A., Doganay, B. 2014. Binnor: An R package for concurrent generation of binary and normal data. Communications in Statistics-Simulation and Computation. 43 (3), 569-579.

Duncan, G. J., Kalton, G. 1987. Issues of design and analysis of surveys across time. International Statistical Review. 55 (1), 97-117.

Fahrmeir, L., Kneib, T., Lang, S., Marx, B., 2013. Regression-Models, Methods and Applications. Springer.

Fitzenberger, B., Fuchs, B., 2017. The residency discount for rents in Germany and the tenancy law reform act 2001: Evidence from quantile regressions. German Economic Review. 18 (2), 212-236.

Fuller, W.A., 1990. Analysis of repeated survey. Survey Methodology. 16 (2), 167-180.

Haslett, S.J., 1986. Time series methods and repeated sample surveys. Ph.D. Thesis. Victoria University of Wellington. http://researcharchive.vuw.ac.nz/handle/10063/971

Horn, R.A., Johnson, C.R., 2013. Matrix Analysis. Second edition, Cam-

bridge, England: Cambridge University Press.

Imbriano, P., 2018. Methods for improving efficiency of planned missing data designs. Ph.D. Thesis. The University of Michigan.
https://deepblue.lib.umich.edu/handle/2027.42/144155

Ismail, M. A., Auda, H.A., Elzafrany, Y.A., 2018. On time series analysis for repeated surveys. Journal of Statistical Theory and Applications. 17 (4), 587-596.

Kauermann, G., Windmann, M., Münnich, R., 2020. Data collection for rent indexes: Overview and classification from the perspective of statistics. AStA Economic and Social Statistics Archive. 14 (2), 145–162.
https://doi.org/10.1007/s11943-020-00272-x

Kott, P.S., 1994. Regression analysis of repeated survey data (with available software). American Statistical Association, Proceedings of the Survey Research Methods Section. 116-123.

Quality, L., Tille, Y., 2008. Variance estimation of changes in repeated surveys and its application to the Swiss survey of value added. Survey Methodology. 34 (2), 173-181.

Scott, A. J., Smith, T.M.F., 1974. Analysis of repeated surveys using time series methods. American Statistical Association. 69 (347), 674-678.

Steel, D., McLaren, C., 2008. Design and analysis of repeated surveys. Centre for Statistical and Survey Methodology. University of Wollongong. Working Paper Series, 11-08.
https://ro.uow.edu.au/cssmwp/10/

Steinberg, D., 2005. Computation of matrix norms with applications to robust optimization. Research thesis. Technion - Israel University of Technology.

Thomschke, L., 2019. Regional impact of the German rent brake. German Economic Review. 20 (4), 892–912.
https://doi.org/10.1111/geer.12195

Van Buuren, S., Groothuis-Oudshoorn, G., 2011. mice: Multivariate imputation by chained equations in R. Journal of Statistical Software. 45 (3), 1-67.

Yuan, S.F., Yu, Y.B., Li, M.Z., Jiang, H., 2020. A direct method to Frobenius norm based matrix regression. International Journal of Computer Mathematics. 97 (9), 1767-1780.
https://doi.org/10.1080/00207160.2019.1668558

A.2:

Kauermann, G. and Ali, M. (2020). Semi-parametric Regression When Some (Expensive) Covariates Are Missing By Design. Accepted for publication in *Journal of Statistical Papers.*

Available online at: https://doi.org/10.1007/s00362-019-01152-5

Statistical Papers
https://doi.org/10.1007/s00362-019-01152-5

REGULAR ARTICLE

Semi-parametric regression when some (expensive) covariates are missing by design

Göran Kauermann[1] · **Mehboob Ali**[1]

Received: 22 January 2019 / Revised: 15 December 2019

Abstract

The paper deals with the scenario where some covariates are observed by design for a subset of the observations only. In the example treated in the paper this occurs with a two phase sampling scheme where in the first phase a relatively large sample is drawn to record a response variable Y and a set of (cheap) covariates x. In a second phase a smaller sample is drawn from the first phase sample where additional (usually expensive) covariates z are also recorded. The second phase can be drawn with unequal probability sampling, where the sampling weights depend on the observed Y and x. The overall intention is to fit a regression model of Y on both, x and z. Due to the design of the data collection we are faced with missing values for z for a majority of observations. We propose an approximate estimation approach using semi-parametric mean and variance regression of Y on x only and augment this fit with a full regression model of Y on x and z. The idea extends the approach of Little (1992) towards non-normal data and non-linear models. The proposed estimation is numerically rather simple and performs convincingly well in simulation studies compared to alternatives such as complete-case and multiple imputation analysis.

Keywords Semi-parametric model · Two phase sampling · Unequal probability · Missing data methods

1 Introduction

The work in this paper is stimulated by a practical problem which is not uncommon in regression analysis. We are interested in fitting a regression model for a continuous response variable Y which depends on a number of covariates, here labelled as x and z. To draw conclusion about a (finite) population we draw a random sample and observe

✉ Göran Kauermann
goeran.kauermann@lmu.de

[1] Department of Statistics, Ludwig-Maximilians-Universität München, Ludwigstrasse 33, D-80359 Munich, Germany

Published online: 01 January 2020

Y, x and z. While both, Y and x are easy to observe and record, the measurement of z is time and/or cost intensive. The example which motivated our research comes from a survey on rents for apartments regularly run in German cities. Such surveys are an official instrument in the German apartment rental market (see e.g. Fahrmeir et al. 1998 or Fitzenberger and Fuchs 2017). In this case we have Y as net rent per square meter of an apartment of floor size x while z are additional covariates describing facilities and equipment of the apartment. Quantities Y and x can be easily obtained, e.g. from a telephone survey or through a data base query. In contrast, the data collection of z is time consuming and is typically pursued by (expensive) personal interviews. This suggests to use a two phase sampling strategy. In the first phase a large random sample is drawn from the population (i.e. the apartments in a city or community) and variables Y and x are recorded. In a second phase a smaller random sample out of the first sample is drawn to record covariates z. The second phase sample can be drawn with inclusion probabilities dependent on Y and x, for instance one may select the sampling probability to depend on x (e.g. large apartments) or on y (e.g. expensive apartments) or on x and y (e.g. expensive apartments when adjusted for floor size). The aim is to fit a regression model for Y given both covariate x and z, which is then used to predict the rent for an apartment with given floor size x and facilities z. Overall a high predication accuracy of the resulting fitted model is our ultimate goal.

The problem of missing covariates is generally reviewed e.g. in Meng (2000), Toutenburg and Nittner (2002), Ibrahim et al. (2005) or Horton and Kleinman (2007). The situation where one (or more) covariates are always observed and other covariates are missing except for a smaller subsample is previously discussed for instance in Zhang and Rockette (2005), Mcleish and Struthers (2006) or Zhao et al. (2009) or more recently in Lumley (2017). Typically a likelihood based approach is employed which requires the specification of a joint distribution for covariates x and z. Assuming that Y, x and z follow a joint normal distribution simplifies the likelihood, which is shown in the classical paper by Little (1992) or even earlier by Anderson (1957). Missing covariates resulting form a two-phase sampling strategies are also treated in Mandallaz et al. (2013). A more general likelihood based setting is discussed in Lawless et al. (1999), Little and Rubin (2002) or Ibrahim et al. (2005) by decomposing the joint distribution of Y, x and z to

$$f(y, x, z) = f(y|x, z)g(z|x)h(x)$$

where $f(.)$ contains the regression model of interest and $g(.)$ and $h(.)$ model the covariate distribution which might depend on additional nuisance parameters. For observations where z is observed we have $f(y_i|x_i, z_i)$ as likelihood contribution while for observation where z remains unobserved the likelihood contribution equals $\int f(y_i|x_i, z_i)g(z_i|x_i)dz_i$. In fact, this requires to specify a distribution or prediction model among the covariates which can be difficult and cumbersome, in particular if z is multi-dimensional like in our example.

A second strand to tackle missing data is to make use of multiple imputation. Imputation can thereby rely on pure prediction models like tree based models. We refer to Donders et al. (2006) for a review or Carpenter and Kenward (2013) for a general discussion. In this paper we employ available and implemented imputation

routine as alternative and show in simulations and the example that the semi-parametric approach outperforms imputation in the data constellation considered in this paper.

While most of the above articles deal with linear regression we here apply semi-parametric regression, i.e. we assume that the effect of x is a smooth function $m(x)$ while z is modelled parametrically, that is we assume $E(Y|x, z) = m(x) + z\beta_z$. This is commonly known as partial linear model and estimation can be carried out with e.g. kernel smoothing or penalized-spline smoothing. For kernel smoothing we refer exemplary to Liang et al. (2004), Liang (2008), Wang (2009) or Qin et al. (2012) who discuss unbiased estimation in case of missing covariates. Kernel smoothing can be numerically cumbersome which is why we pursue penalized spline smoothing in this paper. This smoothing technique has become very popular over the last two decades. Originally proposed in O'Sullivan (1986) and Eilers and Marx (1996), the book by (Ruppert et al. 2003) and available software (see Wood 2017) made the smoothing technique to become a common standard as shown in the review paper by Ruppert et al. (2009). Our proposal is to use non-parametric estimation for both, the regression of Y on x and z resulting from the second phase sample but also mean and variance regression of Y on x only, using the first phase sample. In this view, we follow the original idea of Little (1992) and extending it towards non-parametric regression and categorical covariates z. Penalized splines in the framework of missing data have also been discussed by Little and An (2004) and Zhang and Little (2009) making use of propensity scores. Simulation studies demonstrate that making use of all collected data leads to smaller forecasting errors, i.e. if we make full use of the first phase sample with data on Y and x and the second phase sample with data on Y, x and z, we can reduce the forecasting error compared to alternative methods dealing with missing covariates.

The paper is organized as follows. In Sect. 2, we provide the introduction of our two phase sampling scheme and propose our method. In Sect. 3, we give simulation studies and compare our proposed method with the complete data case, default imputation and Classification and Regression Trees (CART) imputation methods under different sampling schemes. We also compare the methods on a real data example and use different sample sizes for the second phase to vary missing data percentage. Section 4 discusses our findings.

2 Two phase sampled data

2.1 Simple random sample

We are interested in the regression (as well as prediction) of Y given the covariates x and z. Response Y is assumed to come from the additive semi-parametric regression model

$$Y = \beta_0 + m(x) + z\beta_z + \varepsilon \tag{1}$$

where ε is a zero mean (homoscedastic) residual, $m(.)$ is a smooth but otherwise unspecified function and $z\beta_z$ is a linear predictor built from covariates z. In the example

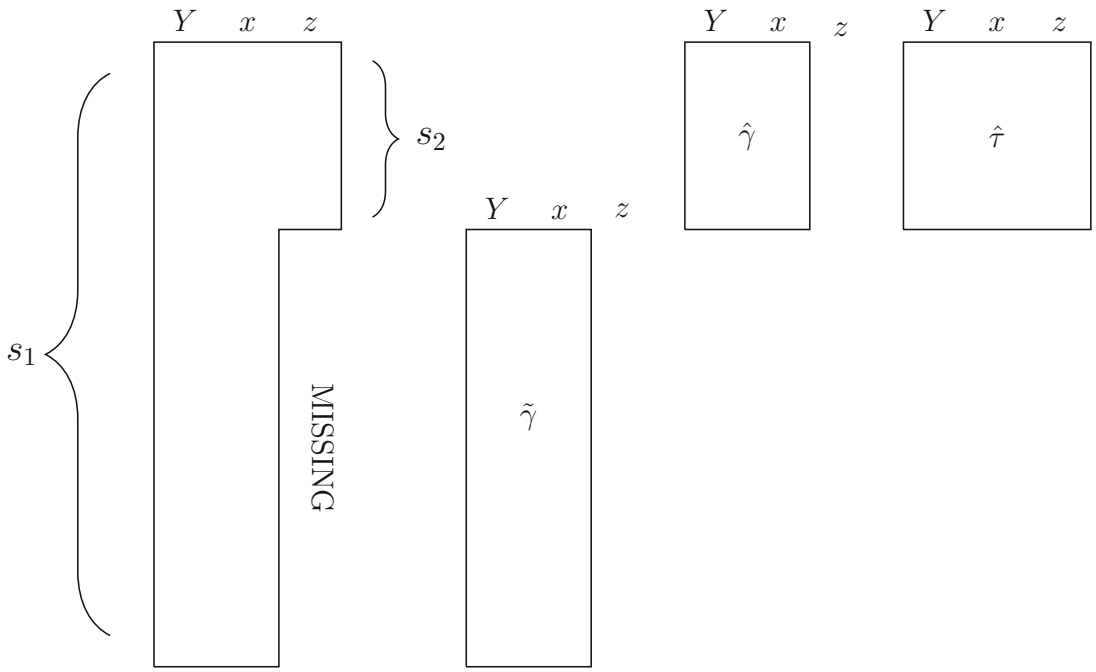

Fig. 1 Missing data pattern (left) and notation of estimates (right)

which motivated this work we have x as continuous covariate (floor space) while z is a vector of discrete valued quantities (apartment facilities) influencing Y (rent of apartment) in a semi-parametric form. For identifiability reasons we additionally assume that $m(0) = 0$ where different identifiability constraints are possible as well (see Wood 2017).

The data at hand have a clear missing pattern since we observe Y and x for the (large) first phase sample of size n_1. For the smaller second phase sample drawn from the first sample we observe the remaining covariates z. Note that the data structure can be sketched as shown in Fig. 1 (left side). To be specific, from the population $(Y_i, x_i, z_i) : i = 1, ..., N$ we draw the random sample $s_1 \subset \{1, ..., N\}$ with $|s_1| = n_1$ and obtain the data $(Y_j, x_j) : j \in s_1$. The second phase random sample $s_2 \subset s_1$ with $|s_2| = n_2 < n_1$ leads to information $(Y_k, x_k, z_k) : k \in s_2$. The question is now how to make use of the available data. Our intention is thereby to obtain a model with small prediction error.

Following Little (1992) we can decompose the distribution of Y and z given x as

$$f(y, z|x) = f(y|x) f(z|y, x). \tag{2}$$

We assume that this decomposition applies to the parameterization in that some parameter θ decomposes uniquely to γ and τ, that is for some invertible function $h(.)$ we have $\theta = h(\gamma, \tau)$. In this case formula (2) becomes

$$f(y, z|x; \theta) = f(y|x; \gamma) f(z|y, x; \tau). \tag{3}$$

This holds exactly if we assume, for instance, joint normality of Y and z given x. Our argument subsequently will be that even if the parameter decomposition does not

hold exactly it may still be used approximately. Nonetheless, the general estimation principle becomes clear with (3) and equals the proposal of Little (1992). Following the likelihood principle we can use information from the data in $s_1 \setminus s_2$ to provide information about parameter γ while sample s_2 is used to estimate τ, that is the second component in (3). We notate the estimates based on sample $s_1 \setminus s_2$ with a tilde notation; while estimates based on the second sample s_2 are written with a hat notation. The final estimate is notated with hat and tilde so that we can rewrite the estimated version of (3) as

$$\hat{\tilde{f}}(y, z|x) = \tilde{f}(y|x)\hat{f}(z|y, x). \tag{4}$$

where $\hat{\tilde{f}}(y, z|x) = f(y, z|x; \hat{\tilde{\theta}})$ and $\hat{\tilde{\theta}} = h(\tilde{\gamma}, \hat{\tau})$ and obvious definitions for the right hand side of (4). We are interested in the conditional distribution of Y given x and z so that we need to condition on z. This is achieved through

$$\hat{\tilde{f}}(y|x, z) := \frac{\hat{\tilde{f}}(y, z|x)}{\hat{f}(z|x)} = \tilde{f}(y|x)\frac{\hat{f}(z|y, x)}{\hat{f}(z|x)}. \tag{5}$$

In order to estimate the later component in (5) we retransform the ratio to

$$\frac{\hat{f}(z|y, x)}{\hat{f}(z|x)} = \frac{\hat{f}(y, z|x)}{\hat{f}(z|x)\hat{f}(y|x)} = \frac{\hat{f}(y|z, x)}{\hat{f}(y|x)}, \tag{6}$$

where the hat notation indicates that all components are estimated from s_2.
To estimate the above quantities we make a strategic simplification of using a semi-parametric model for the mean and the variance. We start with the partial linear model (1) which we rewrite to

$$Y|x, z \sim N(\beta_0 + m(x) + z\beta_z, \sigma^2). \tag{7}$$

If we marginalize over z we obtain the mean and variance model

$$Y|x \sim (\beta_{01} + m_1(x), \sigma_1^2(x)). \tag{8}$$

Note that we do not assume a particular distribution for model (8) but just give the first two moments. Index 1 in the model notation above refers to the fact, that the model relies on Y and x only and hence can be fitted from the first sample. In contrast model (7) is a model for Y, x and z and hence needs sample s_2 to be fitted. If we assume linearity for $z\beta_z$ and postulate normality of z given x, then model (8) is in fact a normal distribution model obtained by marginalizing the joint normality of Y and z given x. This leads to Little (1992) with the extension of assuming a non-linear influence of x. To see this assume that conditional on x we have the joint normality

$$\begin{pmatrix} Y \\ z \end{pmatrix} |x \sim N\left[\begin{pmatrix} \beta_{01} + m_1(x) \\ \beta_{0z} + m_z(x) \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \\ \sigma_{yz}^2 & \sigma_{zz}^2 \end{pmatrix}\right].$$

Then (8) results from (7) with

$$z\beta_z = z\sigma_{yz}\sigma_{zz}^{-1} \quad \text{and} \quad m(x) = m_1(x) + \sigma_{yz}\sigma_{zz}^{-1}m_z(x)$$

and obvious definition for β_0. We generalize this step in so far, that we do not assume a joint normality of Y and z given x but use model (8) as a proxy for the marginal normal model resulting from (7). In our setting, z is categorical. In this case the marginal model resulting from (7) becomes a mixture of normal distributions which equals

$$\sum_l N(\beta_0 + m(x) + z_{(l)}\beta_z, \sigma^2)P(z = z_{(l)}|x)$$

with $z_{(l)}$ as the possible outcomes of z. This yields to the mean value

$$E(Y|x) = \sum\nolimits_l (\beta_0 + m(x) + z_{(l)}\beta_z)P(z = z_{(l)}|x)$$

which for a linear predictor $z\beta_z$ leads to

$$E(Y|x) = \beta_{01} + m_1(x)$$

where $m_1(x) = m(x) + E(z|x)\beta_z$ and $E(z|x) = \sum_l z_{(l)}P(z = z_{(l)}|x)$. As long as $P(z|x)$ is smooth in x, which is a reasonable assumption, we obtain $m_1(x)$ as smooth function in x.
The marginal variance of Y conditional on x results through

$$\begin{aligned} Var(Y|x) &= E_z[Var(Y|x,z)] + Var_z[E(Y|x,z)] \\ &= \sigma^2 + \beta^T Var(z|x)\beta =: \sigma_1^2(x) \end{aligned}$$

where $Var(z|x) = \sum_l (z_{(l)} - E(z|x))(z_{(l)} - E(z|x))^T P(z = z_{(l)}|x)$. This in turn justifies the semi-parametric mean and variance model (8) for Y given x.

Using model (7) and the mean and variance relation in (8), we propose to fit the following models:

1. Fit the heteroscedastic model (8) with sample $s_1 \setminus s_2$ and allow that the variance depends on x. The corresponding estimates are denoted with $\tilde{\mu}_1(x) = \tilde{\beta}_{01} + \tilde{m}_1(x)$ and $\tilde{\sigma}_1^2(x)$. This yields estimate $\tilde{f}(y|x)$ in (5).
2. Fit model (8) with sample s_2 and allow that the variance depends on x. The corresponding estimates are denoted with $\hat{\mu}_1(x) = \hat{\beta}_{01} + \hat{m}_1(x)$ and $\hat{\sigma}_1^2(x)$. This yields estimate $\hat{f}(y|x)$ in (6).
3. Finally, estimate model (7) with sample s_2 which gives estimate $\hat{f}(y|x,z)$ in (6).

We now approximate (8) through a heteroscedastic normal distribution that is we replace (8) by

$$Y|x \overset{a}{\sim} N(\beta_{01} + m_1(x), \sigma_1^2(x)). \tag{9}$$

That is the estimate $\hat{\tilde{f}}(y|x,z)$ in (5) by making use of (6) results to a heteroscedastic normal distribution with moments

$$\hat{\tilde{\sigma}}^2(x) = \frac{1}{\hat{\sigma}^{-2} + \tilde{\sigma}_1^{-2}(x) - \hat{\sigma}_1^{-2}(x)} \quad \text{and} \quad \hat{\tilde{\mu}}(x) = \hat{\tilde{\sigma}}^2(x)\left\{\frac{\hat{\mu}}{\hat{\sigma}^2} + \frac{\tilde{\mu}_1(x)}{\tilde{\sigma}_1^2(x)} - \frac{\hat{\mu}_1(x)}{\hat{\sigma}_1^2(x)}\right\}.$$

The above estimates can then be combined to provide the final fit for the original regression model (1) through $\hat{\tilde{\beta}}_0 + \hat{\tilde{m}}(x) + z\hat{\tilde{\beta}}_z(x)$, where

$$\begin{aligned}
\hat{\tilde{\beta}}_0 &= \hat{\tilde{\sigma}}^2(x)\left\{\frac{\hat{\beta}_0}{\hat{\sigma}^2} + \frac{\tilde{\beta}_{01}}{\tilde{\sigma}_1^2(x)} - \frac{\hat{\beta}_{01}}{\hat{\sigma}_1^2(x)}\right\},\\
\hat{\tilde{m}}(x) &= \hat{\tilde{\sigma}}^2(x)\left\{\frac{\hat{m}(x)}{\hat{\sigma}^2} + \frac{\tilde{m}_1(x)}{\tilde{\sigma}_1^2(x)} - \frac{\hat{m}_1(x)}{\hat{\sigma}_1^2(x)}\right\},\\
\hat{\tilde{\beta}}_z(x) &= \frac{\hat{\tilde{\sigma}}^2(x)}{\hat{\sigma}^2}\hat{\beta}_z.
\end{aligned}$$

Note that with the heteroscedastic model (8) we have induced an interaction between x and z. We simplify this by defining the average coefficient estimate

$$\bar{\hat{\tilde{\beta}}}_z = \frac{1}{n}\sum_i^n \hat{\tilde{\beta}}_z(x_i).$$

If $\hat{\sigma}_1^2(x) \approx \tilde{\sigma}_1^2(x)$, that is if sample $s_1 \setminus s_2$ and s_2 lead to the same residual variance estimate in regression model (8), then $\bar{\hat{\tilde{\beta}}}_z \approx \hat{\beta}_z$, i.e. the estimate for coefficient β_z remains unchanged. Note also that if s_2 is a simple random sample of s_1 we have both $\hat{\sigma}_1^2(x)$ and $\tilde{\sigma}_1^2(x)$ being consistent estimates of $\sigma_1^2(x)$ in model (8) so that $\bar{\hat{\tilde{\beta}}}_z$ is approximately unbiased since $\hat{\sigma}_1^2(x) \approx \tilde{\sigma}_1^2(x)$ for sample sizes of s_2 and s_1 increasing. Similarly, since $\tilde{m}_1(x)$ and $\hat{m}_1(x)$ are consistent estimates of $m_1(x)$ we get that $\hat{\tilde{m}}(x)$ is consistent. If sample s_2 is drawn with unequal inclusion probabilities it is necessary to fit the models based on sample s_2 using weighted regression. This applies in particular if an informative design is used for sample s_2. We will demonstrate this subsequently.

The proposed models can easily be fitted in practice. In fact simple fitting routines like the `gamlss()` function (see Stasinopoulos et al. 2017) in R can be used to fit the models, as will be shown below. Hence, the proposed estimation is very practical and as we see in the simulations below, it can outperform alternative routines like multiple imputation. We sketch some principle ideas penalized spline smoothing in the Appendix A.

We are later primarily interested in prediction and will use model (1) with estimates derived above to obtain an appropriate predictor. For the proposed method this means we replace $m(x)$ and $z\beta_z$ by their estimates $\hat{\tilde{m}}(x)$ and $\hat{\tilde{\beta}}_z(x)$ which yields for new values x_0 and z_0 say, the prediction value $\hat{\tilde{Y}}_0 = \hat{\tilde{\beta}}_0 + \hat{\tilde{m}}(x_0) + z\hat{\tilde{\beta}}_z(x)$. This is compared to

Y_0, the true observed value, leading to the prediction error

$$E\left\{(Y_0 - \hat{\tilde{Y}}_0)^2 | x_0, z_0\right\}.$$

Apparently, the prediction error depends on both, the data used for fitting as well the new observation Y_0. To get an estimate for the prediction error we use cross-validation. That is we will use parts of the data for fitting the model (training data) and parts of the data for prediction (test data).

2.2 Unequal probability sample

Simple random sampling does not necessarily lead to efficient estimates. In fact assigning unequal sampling probability to the units in the population can reduce the estimator variability. We refer to Hanif and Brewer (1980) or Thompson (2012) for a review about unequal probability sampling. We can make use of these ideas apply an informative sampling design based on the available data from the first sample. That is, we draw the second sample s_2 such that the inclusion probabilities depend on x_i and y_i from the first sample. We make use of the idea in two ways: First, we use covariate dependent sampling probabilities in which the selection of large values of x (large apartments with larger floor space) have higher probability and accordingly units will small x (smaller apartments) have lower probability. To assign such proportional probability for the second phase sample we use sampling probabilities like

$$p = \exp(1 + 0.25\tilde{x})/[1 + \exp(1 + 0.25\tilde{x})] \tag{10}$$

where $\tilde{x}$ is a standardizes version of covariate x drawn in the first phase sample (i.e. $\tilde{x} = \frac{x - mean(x)}{sd(x)}$). Secondly, a residual dependent sampling procedure is used in which residuals from the regression model of Y and x from the first phase sample are used. The motivation behind this approach is that large absolute residuals indicate an apartment with high or low facility standards expressed in z. Hence, for these apartments the knowledge of z is informative. To select the residual dependent sample, we assign the sampling probability

$$p = \exp(\text{abs}(\varepsilon))/[1 + \exp(\text{abs}(\varepsilon))] \tag{11}$$

where abs(ε) is the absolute value of the residual from first phase sample that is $\varepsilon = Y - \tilde{\beta}_{01} - \tilde{m}_1(x)$. Note that in both unequal probabilities sampling schemes, the missing mechanism is missing at random (MAR) because missing covariates z dependent on the fully observed data x and response variable Y. We refer to Mitra and Reiter (2016), Yang and Kim (2016) or Zhang et al. (2016) for a review of covariate dependent missing mechanisms generated by a response model. The second phase sample for unequal probability sampling can now be drawn with unequal probability sampling schemes as proposed in Tille (1996, 2006) and Deville and Tille (1998).
For estimation based on unequal probability sampled data we use sampling weights for fitting the models above.

2.3 Weighted estimation

We denote with δ_i the observation indicator for sample s_2, that is $\delta_i = 1$ for $i \in s_2$ and $\delta_i = 0$ for $i \in s_1 \setminus s_2$. Using the sampling probabilities (10) and (11), respectively, we have for each individual in sample s_1 the probability p_i, say. The likelihood for fitting model (7) takes the form $\sum_{i \in s_2} \delta_i l_i(\theta)$ where $\theta = (u, \beta, \sigma^2)$ and

$$l_i(\theta) = -\frac{1}{2} log(\sigma^2) - \frac{1}{2}\{y_i - \eta_i(u, \beta)\}^2/\sigma^2.$$

The mean structure $\eta_i(.)$ results through replacing $m(x)$ by some spline bases representation $B(x)u$ so that $m(x) + z\beta_z$ in (7) results through $\eta_i(u, \beta) = B(x_i)u + z_i\beta_z$. We refer to the Appendix A for details. It is obvious that the unweighted likelihood is biased and needs to be replaced by the weighted likelihood

$$\sum_{i \in s_1} \frac{\delta_i}{p_i} l_i(\theta). \tag{12}$$

This is the common setting to achieve unbiased estimates in regression if data are missing at random, that is

$$P(\delta_i = 1 | y_i, x_i, z_i) = P(\delta_i = 1 | y_i, x_i), \tag{13}$$

which holds by construction of sample s_2 (see Robins et al. 1994, 1995; see also Ibrahim et al. 2005). Apparently, we pursue penalized estimation, as sketched in the Appendix A, but the penalty itself is uneffected by any missing data patterns so that we can just add the usual penalty to the weighted likelihood. In the same way we also obtain a weighted penalized likelihood for fitting model (9) based on the sample s_2. In practice this is easily accommodated by including weights in the `gamlss(.)` procedure.

In the same way we need to weight the estimates based on the data from sample $s_1 \setminus s_2$. Defining with $\tilde{l}_i(\gamma)$ the likelihood in the marginal model (9) we get

$$\tilde{l}_i(\gamma) = -\frac{1}{2} log(\sigma_i^2(u_\sigma)) - \frac{1}{2}\{y_i - \eta_i(u)\}^2/\sigma_i^2(u_\sigma)$$

where $\eta_i(u) = B(x_i)u$ and $\sigma_i^2(u_\sigma) = \exp(B_\sigma(x_i)u_\sigma)$. We use the notation of the Appendix A and set $B(.)$ and $B_\sigma(.)$ as spline bases functions with corresponding coefficient vectors u and u_σ, respectively. The weighted likelihood function to estimate $\gamma = (u^t, u_\sigma^t)$ then results through

$$\sum_{i \in s_1} \frac{1 - \delta_i}{1 - p_i} \tilde{l}_i(\gamma).$$

This approach extends the proposed weighting scheme of Robins et al. (1994, 1995) for the estimation of the marginal model for y and x using sample $s_1 \setminus s_2$.

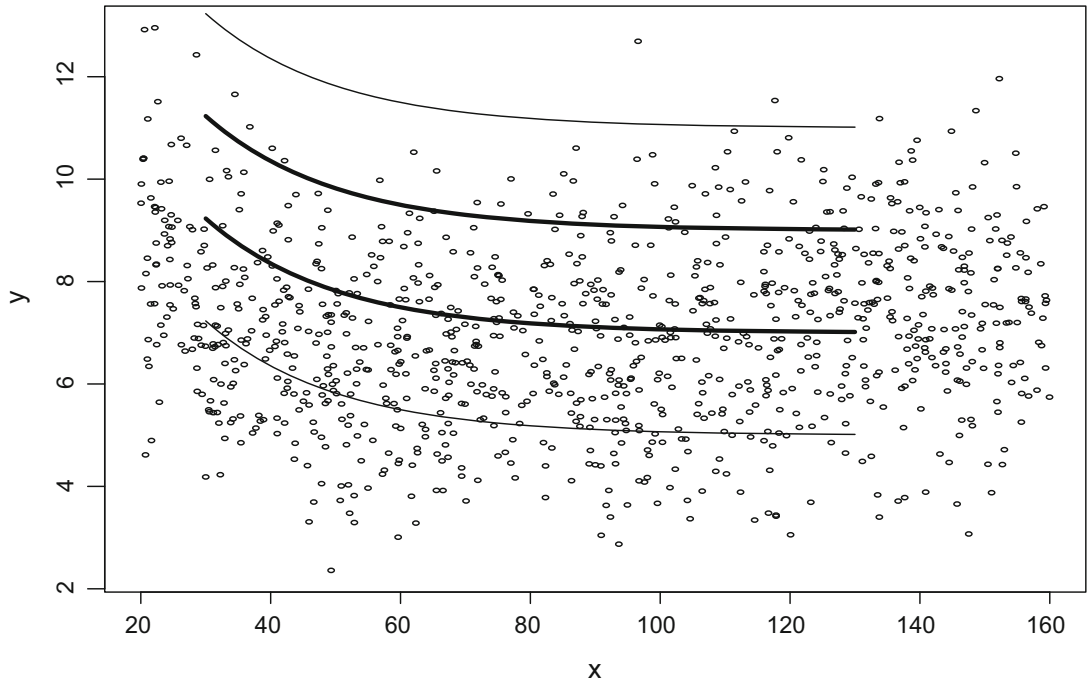

Fig. 2 Typical data constellation for simulation with different levels of $z_1\beta_1 + z_2\beta_2 + z_3\beta_3$ shown by the thickness of the line

3 Simulation and example

3.1 Simulation

We run a simulation study to demonstrate the performance of our approach in comparison to imputation alternatives. To do so we simulate (population) data from the model

$$Y = \beta_0 + m(x) + z\beta_z + \varepsilon \tag{14}$$

where $\varepsilon \sim \mathrm{N}(0, 1.5)$ and $z = (z_1, z_2, z_3)$ is a vector of binary covariates which are correlated with x. For the functional form $m(x)$ we use the response function shown in Fig. 2 for different values of z_1, z_2 and z_3. We generate 5000 values as population. The first phase sample size n_1 is 3000 and $\beta_z \in \{1, 1, 1\}$. The second phase sample size is $n_2 \in \{200, 400\}$.

For the simulation study, covariates z_j are drawn in two steps. First, covariate x is generated from a uniform distribution with parameters (20, 160). Secondly, z_j is generated as $z_j \sim \text{Bernoulli}(\pi_j)$ with

$$\pi_j = \exp(\alpha_0 + x\alpha_x)/[1 + \exp(\alpha_0 + x\alpha_x)], \quad j = 1, 2, 3$$

where $\alpha_0 = 0$ and $\alpha_x = 1$.

The first phase sample is selected by simple random sampling (SRS) and we used three sampling schemes in the second phase: (a) equal probability sampling (simple random sampling without replacement), (b) covariate dependent probability sampling

Table 1 Median of mean squared prediction error for simulated data

$n_2 = 200$						$n_2 = 400$					
Error term	Sampling	Prop.Me	Imp.R	Imp. CART	CC	Prop.Me	Imp.R	Imp. CART	CC	S1.Yx	All.Data
ε^*	a	2.354	2.372	2.379	2.392	2.333	2.336	2.346	2.350	2.916	2.315
	b	2.365	2.370	2.383	2.398	2.339	2.344	2.342	2.359	2.916	2.315
	c	2.344	2.357	2.378	2.389	2.335	2.339	2.349	2.353	2.916	2.315
ε^{**}	a	1.656	1.670	1.675	1.680	1.636	1.648	1.651	1.648	2.227	1.624
	b	1.659	1.671	1.674	1.683	1.643	1.650	1.647	1.653	2.227	1.624
	c	1.648	1.655	1.671	1.684	1.639	1.644	1.647	1.648	2.227	1.624

a: Equal probability
b: Tille covariate dependent
c: Tille residual dependent
ε^*: $\varepsilon \sim N(0, 1.5)$
ε^{**}: $\varepsilon \sim N(0, 1.5(0.9 + (1/x))^2)$

as given in (10) and (c) residual dependent probability sampling as in (11). For the selection of both unequal probabilities sampling (i.e. covariate and residual dependent sampling of (10) and (11)), we use the inclusionprobabilities( ) and UPtille( ) functions from `sampling` package in R (Tille and Matei 2016). Note that the variables Y and x are given for all sampled data (i.e. n_1 and n_2), but vector z is available only for the second sample s_2. We repeated the simulations 100 times.

For the three simulation scenarios (SRS, Tille covariate, Tille residual) of the second phase sampling schemes, we focus on the mean squared prediction errors of the fitted model in the population data (i.e. out of sample prediction) and compare these with three alternatives. First, we make use of a complete case analysis, that is we use the data from the second phase sample only. Complete case analysis is a commonly used method in missing data structures (see Hayati et al. 2015). Secondly, we make use of two (multiple) imputation techniques using the R package `mice`, (see van Buuren and Groothuis-Oudshoorn 2011). First, the default setting of mice is used which means mice assigns multiple imputation methods according to the type of variable. In our case the default mice setting for binary variables is logistic regression. Classification and regression trees are used for multiple imputation in the second method. We use the default number of imputations for both imputation methods in mice.

We displays our findings in graphs and tables. In our results, we used the following abbreviation: "Prop.Me" describes our proposed method, "Imp.R" gives imputation for regular/default mice setting, "Imp.CART" indicates imputation for CART mice, "CC" stands for complete case analysis, "S1.Yx" shows results by fitting a regression of Y on x only using the first data sample, that is only use model (8) for fitting and prediction and finally "All.Data" is for the hypothetical case where all information on x and z is recorded in the first sample. Note that we are fitting the "S1.Yx" model on sample s_1 and do not include covariates z. In contrast, in "All.Data" case we are including covariates z. We look at the mean squared prediction error.

Median values of the mean squared prediction error of 100 simulations are listed in Table 1 (first three rows). We also calculate the ratio of the prediction error by

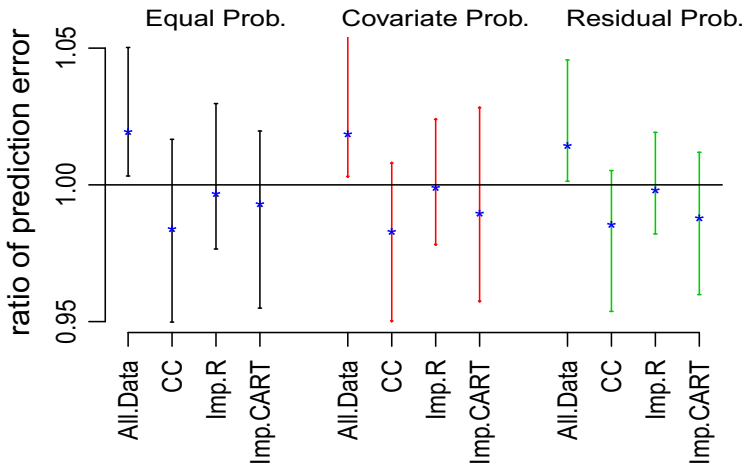

Fig. 3 $n_2 = 200$: ratio of mean squared prediction error for simulated data. Second phase sample selection with equal, Tille covariate and Tille residual dependent probability sampling with $\varepsilon \sim N(0, 1.5)$

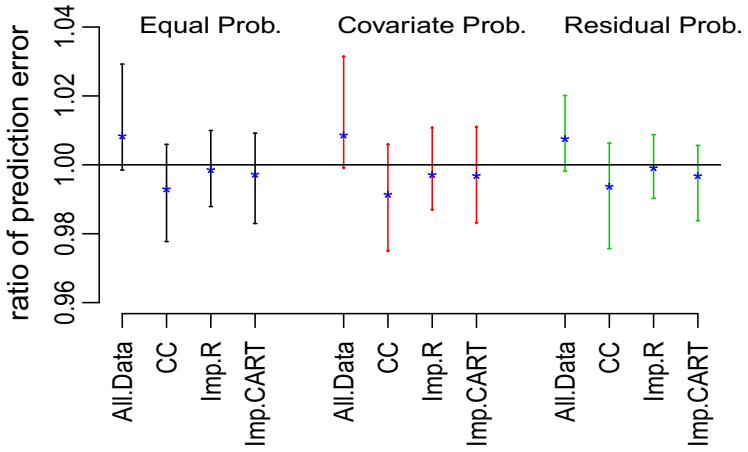

Fig. 4 $n_2 = 400$: Ratio of mean squared prediction error for simulated data. Second phase sample selection with equal, Tille covariate and Tille residual dependent probability sampling with $\varepsilon \sim N(0, 1.5)$

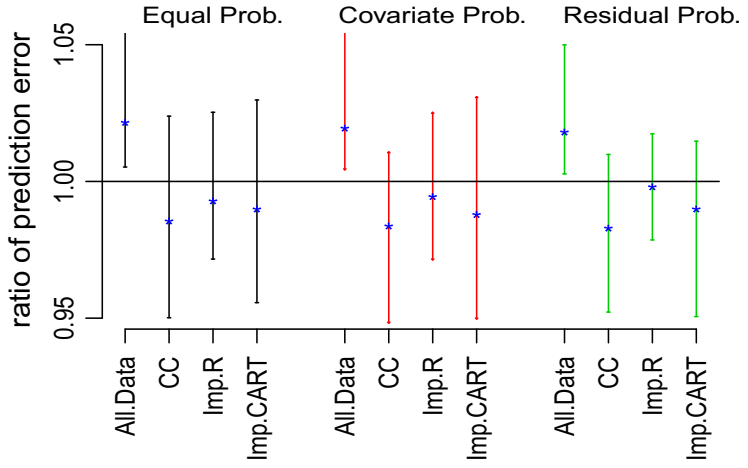

Fig. 5 $n_2 = 200$: ratio of mean squared prediction error for simulated data. Second phase sample selection with equal, Tille covariate and Tille residual dependent probability sampling with $\varepsilon \sim N(0, 1.5(0.9 + (1/x))^2)$

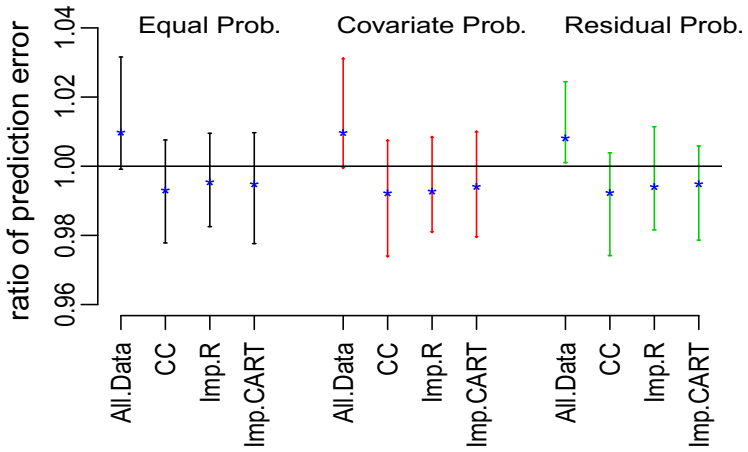

Fig. 6 $n_2 = 400$: ratio of mean squared prediction error for simulated data. Second phase sample selection with equal, Tille covariate and Tille residual dependent probability sampling with $\varepsilon \sim N(0, 1.5(0.9 + (1/x))^2)$

dividing the mean squared prediction error of the proposed method in each simulation by the corresponding mean squared prediction error resulting from the competing routines. This is shown in Fig. 3 where we did not include S1.Yx due to its weak performance. The horizontal line indicates the value 1. Note that values below 1 speak in favor of our proposal. The vertical bars include the inner 90 percent range of the ratios for all simulations. The cross mark on bars represent the median values of the ratios of prediction error. We observe that our proposed method on average provides the minimum the prediction error compared to the complete case and the multiple imputation methods, respectively. Naturally the prediction error ratio using all data cases is smaller since it assumes that missing z values are known. The same behavior is seen if we increase the second phase sample size to $n_2 = 400$, see Fig. 4.

We now replace $\varepsilon \sim N(0, 1.5)$ in model (14) with $\varepsilon \sim N(0, 1.5(0.9 + (1/x))^2$, where x is the continuous covariate generated from the uniform distribution described above. All other settings remain unchanged. The median values are reported in Table 1 (last three rows) and ratio of the prediction error are shown in Figs. 5 and 6 for second phase samples of size $n_2 = 200$ and $n_2 = 400$, respectively. The overall results remains unchanged under heteroskedasticity response model (14). To compare the performance of the proposed method with alternatives we also reported the bias and the estimated variance exemplary for the regression coefficient $\hat{\beta}_2$ in Table 2. An analogous performance is observable for $\hat{\beta}_1$ and $\hat{\beta}_3$, which is therefore not reported here. The bias and the estimated variance are calculated as

$$\text{bias}(\hat{\beta}_2) = E_m(\hat{\beta}_2) - \beta_2 \quad \text{and} \quad \text{est.var}(\hat{\beta}_2) = \frac{1}{m}\sum_{i}^{m}(\hat{\beta}_{2i} - \beta_2)^2$$

where $E_m(\hat{\beta}_2)$ is average value of m=100 fitted regression coefficient $\hat{\beta}_2$ and β_2 is corresponding true value of covariates z used in the simulation model (14). We can see that our proposed routine provides a low bias and as well as a small variance. In fact, the variance of our method and the complete case analysis show very similar results

Table 2 Bias and estimated variance for regression coefficient $\hat{\beta}_2$ for simulated data

			$n_2 = 200$				$n_2 = 400$			
	Error Term	Sampling	Prop.Me	Imp.R	Imp.CART	CC	Prop.Me	Imp.R	Imp.CART	CC
bias($\hat{\beta}_2$)	ε^*	a	−0.037	−0.090	−0.159	−0.064	−0.055	−0.103	−0.136	−0.063
		b	−0.001	−0.082	−0.136	−0.024	−0.046	−0.103	−0.133	−0.052
		c	−0.072	−0.141	−0.220	−0.092	−0.060	−0.092	−0.134	−0.067
	ε^{**}	a	−0.026	−0.094	−0.147	−0.053	−0.044	−0.107	−0.143	−0.052
		b	0.002	−0.088	−0.175	−0.020	−0.037	−0.105	−0.134	−0.043
		c	−0.032	−0.120	−0.176	−0.051	−0.034	−0.081	−0.107	−0.039
est.var($\hat{\beta}_2$)	ε^*	a	0.054	0.071	0.085	0.054	0.027	0.031	0.055	0.031
		b	0.055	0.059	0.094	0.053	0.030	0.041	0.046	0.032
		c	0.045	0.064	0.110	0.053	0.027	0.033	0.043	0.030
	ε^{**}	a	0.037	0.054	0.055	0.038	0.019	0.025	0.039	0.022
		b	0.038	0.044	0.091	0.037	0.021	0.032	0.039	0.022
		c	0.031	0.045	0.066	0.035	0.015	0.022	0.028	0.017

a: Equal probability
b: Tille covariate dependent
c: Tille residual dependent
ε^*: $\varepsilon \sim N(0, 1.5)$
ε^{**}: $\varepsilon \sim N(0, 1.5(0.9 + (1/x))^2)$

which outperform of the different multiple imputation methods. This also applies to the setting where we observe more than one continuous covariate, that is smoothing is carried out over two dimensions. We refer to the simulations provided in Appendix B.

3.2 Example

Finally we apply the routine to the real data example discussed in the introduction. We have data on the rent per square meter (in Euros) for 3024 apartments available. Besides the floor space we look at the following ten indicator variables describing the facilities of an apartment: $z_1 = 1$ if the apartment lies in an average residential location, $z_2 = 1$ if the apartment has an open kitchen/eat in kitchen, $z_3 = 1$ if the apartment has not an upmarket kitchen, $z_4 = 1$ if there is under floor heating, $z_5 = 1$ if the apartment has the standard central heating, $z_6 = 1$ if the apartment has a good bathroom equipment, $z_7 = 1$ if the apartment has new floor, $z_8 = 1$ if the apartment has bad floor, $z_9 = 1$ if the apartment has good floor and $z_{10} = 1$ if the apartment has back side building. The estimates for the complete data are shown in Fig. 7 and Table 3. The numbers in the table show, for instance that the rent per square meter decrease by 1.27 for average residential location.

Though the data result from a sample we consider them now as population model to measure out-of-sample performance of the routines. To be specific we draw a first phase sample of $n_1 = 2500$ apartments by simple random sampling from the 3024 apartments. We measure all subsequent "out of sample" mean squared prediction

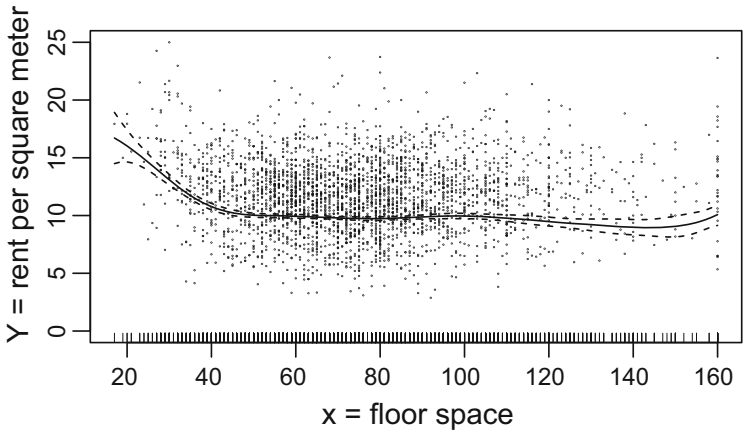

Fig. 7 Estimated effect of floor space for rent data

Table 3 Estimates for rent data

Covariates	Estimate	SE	t value	Pr(>|t|)
z1	− 1.27	0.10	−13.22	< 2e−16
z2	1.04	0.15	6.99	3.46e−12
z3	− 1.36	0.11	− 12.50	< 2e−16
z4	1.78	0.17	10.21	< 2e−16
z5	0.42	0.12	3.44	0.0006
z6	1.26	0.25	5.15	2.76e−07
z7	1.07	0.15	7.00	3.09e−12
z8	− 1.11	0.18	− 6.29	3.63e−10
z9	1.25	0.15	8.54	< 2e−16
z10	0.64	0.19	3.48	0.0005

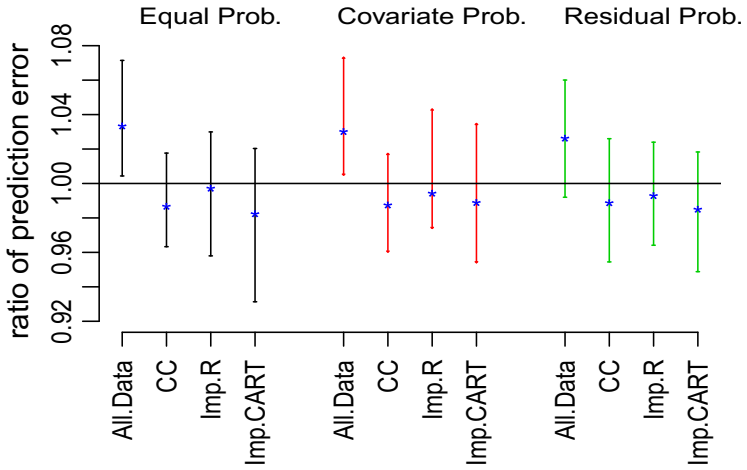

Fig. 8 $n_2 = 400$: ratio of mean squared prediction error for rent data. Second phase sample selection with equal, Tille covariate and Tille residual dependent probability sampling

errors based on the 3024 - 2500 = 524 apartments not in the first sample. That is, we divide the data into a "training" and "test" data set. We select the second phase sample of size 400 using the three different ways as discussed in simulation study above.

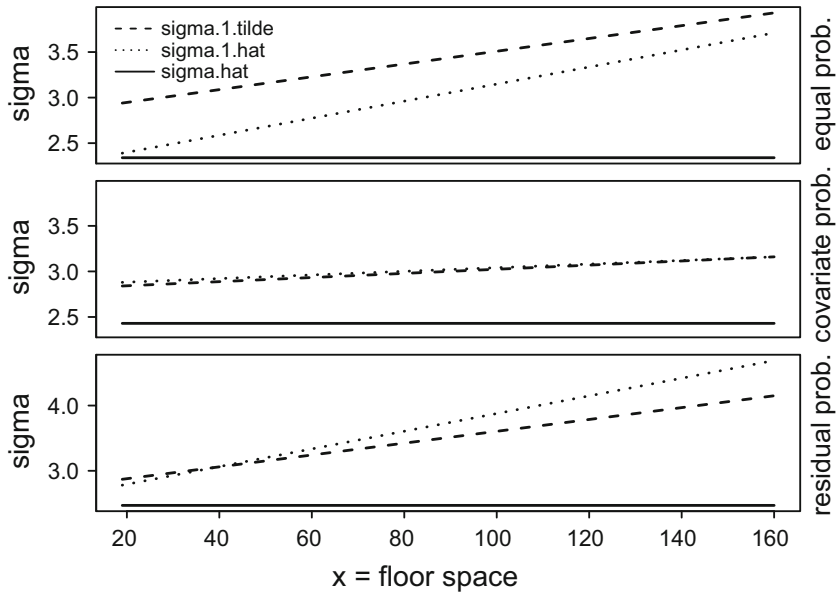

Fig. 9 $n_2 = 400$: estimated variances for rent data. Second phase sample selection with equal, Tille covariate and Tille residual dependent probability sampling

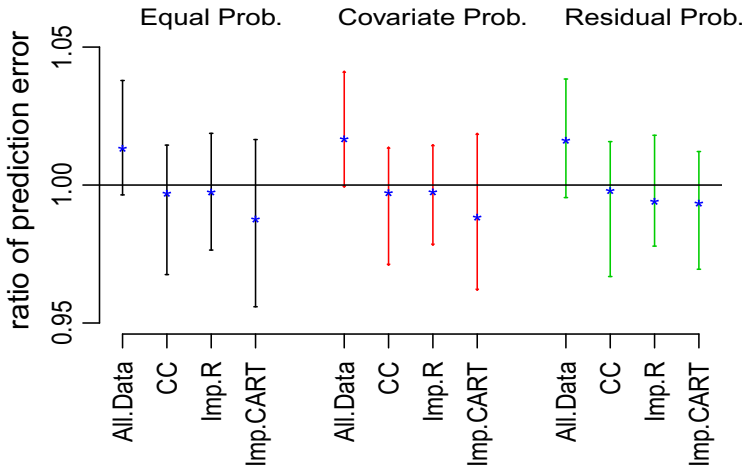

Fig. 10 $n_2 = 700$: ratio of mean squared prediction error for rent data. Second phase sample selection with equal, Tille covariate and Tille residual dependent probability sampling

The apartment size is given for all data, but entries on the ten covariates describing the facilities of the apartment are only available for 400 apartments (sample s_2). We repeat this step 50 times leading to 50 samples of the first and second phase sample. For each sample we fit the model using the proposed method from above and compare this with the fit based on the complete data (i.e. the second phase sample of size 400) and two multiple imputations using mice. For completeness we also calculate the prediction error using all data, i.e. omitting the missing pattern induced on the original data set of first phase. As visualized in Fig. 8, where we use SRS, Tille covariate and Tille residual dependent sampling schemes for second phase sample selection, that the proposed method performs better or at least comparable to the alternatives. To demonstrate the effect of the heteroscedasticity induced with model (8) we plot the estimated variances for a single sample in Fig. 9. Estimates $\tilde{\sigma}_1(x)$ and $\hat{\sigma}_1(x)$ are very

Table 4 Median of mean squared prediction error for rent data

$n_2 = 400$					$n_2 = 700$					
Sampling	Prop.Me	Imp.R	Imp.CART	CC	Prop.Me	Imp.R	Imp.CART	CC	S1.Yx	All.Data
a	6.614	6.666	6.691	6.635	6.489	6.563	6.565	6.520	9.346	6.378
b	6.585	6.623	6.648	6.724	6.485	6.519	6.570	6.493	9.346	6.378
c	6.533	6.636	6.662	6.606	6.472	6.487	6.600	6.554	9.346	6.378

a: Equal probability
b: Tille covariate dependent
c: Tille residual dependent

close to each other, these are estimated from $s_1 \setminus s_2$ and s_2 sample data respectively. The solid line shows the estimated variance $\hat{\sigma}$ for complete case. The fitted heteroscedastic curves are shown as dotted and dashed line. Heteroscedasticity is quite apparent.

The analysis on the rent data example is repeated by increasing the second phase sample size to $n_2 = 700$. We find similar results shown in Fig. 10. Median values of mean squared prediction error of 50 simulations for the rent data example for all cases can be seen in Table 4.

4 Discussion

Usually multiple imputation is not necessarily the first choice to deal with missing data in a constellation when a large number of covariates is missing by design. Our approach provides an alternative to multiple imputation which leads to low mean squared prediction error. This holds even when a high percentage of the observations have missing covariates. The approach also allows to apply two phase sampling, that is draw units with unequal probabilities in the second phase based on a regression applied to the first phase data.

The proposed procedure is rather easily applicable without requiring imputation routines. It is shown in simulations and in a real data example that the method outperforms standard multiple imputation routines in terms of mean squared prediction error. We also find that the idea of using covariate and residual dependent sampling in the second phase sample can be more useful as simple random sampling in order to get a lower prediction error.

Our focus in this paper was on the prediction error and we did not tackle questions like model selection or variance estimation. The first can in principle be approached with tools like AIC (Akaike Information Criterion) or BIC (Bayesian Information Criterion). To do so one can fit model (7) for different subsets of the categorical covariates z. Since this is a simple model validation, both AIC and BIC are easy to calculate. Variance estimation becomes more challenging based on theoretical grounds, also because parameters are replaced by function such as $\hat{\tilde{\beta}}(x)$. It is therefore more advisable to rely on bootstrapping as proposed in Saegusa (2015), see also Saegusa (2014).

Acknowledgements Mehboob Ali acknowledges financial support provided by Punjab Higher Education Commission for finishing his dissertation at LMU Munich.

Appendix A: Penalized spline smoothing

Penalized spline smoothing is a very general and numerically stable routine for fitting smooth functions. We refer to Ruppert et al. (2003, 2009) for an excessive discussion of the field. Subsequently we sketch the basic ideas. The main principle is to replace the smooth function $m(x)$ in model (7) and the smooth functions $m_1(x)$ and $\sigma_1^2(x)$ in model (8) by spline bases representation. That is we make $m(x) = B(x)u$ and $m_1(x) = B(x)u_1$, $\sigma_i^2(x) = \exp(B_\sigma(x)u_\sigma)$, where $B(x)$ is spline basis and so is $B_\sigma(x)$ and in principle we can set $B(x) = B_\sigma(x)$. A convenient setting is to use a B-spline basis (see Boor 1972), which is constructed from piece-wise polynomial functions, tied together in a continuous (and where necessary differentable) way. This makes the whole model parametric where the spline coefficients u in model (7) and the coefficients u_1 and u_σ are the parameters which need to be estimated. Given that $B(x)$ is chosen as high dimensional basis we find the coefficient vectors to be high dimensional as well. Estimation will induce large estimation variability which is why Eilers and Marx (1996) proposed to impose a penalization on u, e.g. neighboring coefficients should not differ very much. Such penalization can be written as quadratic form $\lambda u^t Du$ for an appropriately chosen penalty matrix D. This leads to the penalized likelihood

$$l(\theta) - \frac{1}{2}\lambda u^t Du \tag{A1}$$

where θ is the parameter vector of the model that does also contain the coefficient vector u. Parameter λ plays the role of the smoothing parameter and increasing λ will lead to a more penalized fit. Comprehending the latter component in (A1) as log prior leads to a Bayesian framework so that

$$\begin{aligned} u &\sim N(0, \lambda^{-1} D^-) \\ y|u &\sim \exp(l(\theta)) \end{aligned}$$

where D^- stands for the (generalized) inverse of D. Now λ plays the role of a hyper parameter which can be estimated using empirical Bayes ideas. We refer to Wand (2003) for details in this direction.

Appendix B: Multivariate metrical variables

We repeat the simulation for bivariate x and simulate data from the model

$$Y = \beta_0 + m(x_1) + v(x_2) + z\beta_z + \varepsilon \tag{B1}$$

where $\varepsilon \sim N(0, \sigma)$ and $z = (z_1, z_2, z_3)$ is a vector of binary covariates which are correlated with $x = (x_1, x_2)$. For the functional forms $m(x_1)$ and $v(x_2)$ we use the same response functions as shown in Fig. 2 for different values of z_1, z_2 and z_3 for univariate x. The population size, the first and second phase sample size and the true

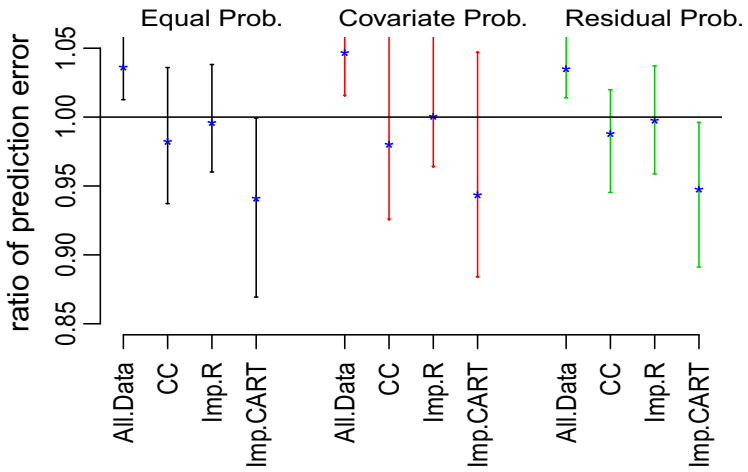

Fig. 11 $n_2 = 200$: ratio of mean squared prediction error for simulated data. Second phase sample selection with equal, Tille covariate and Tille residual dependent probability sampling with case 1

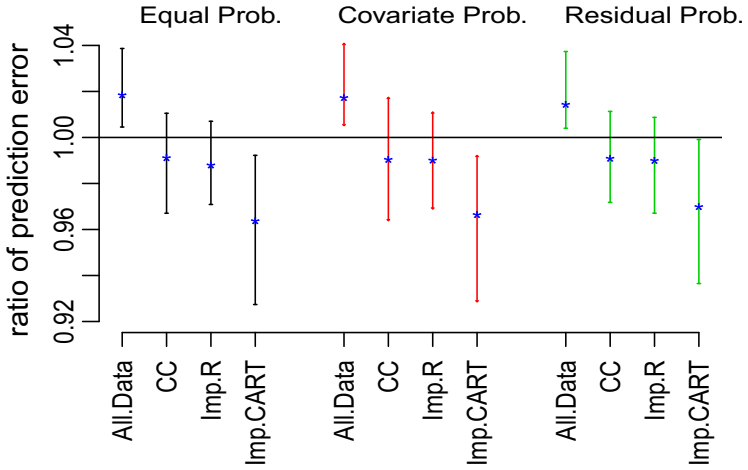

Fig. 12 $n_2 = 400$: ratio of mean squared prediction error for simulated data. Second phase sample selection with equal, Tille covariate and Tille residual dependent probability sampling with case 1

β_z values for covariates z remain unchanged as for model (14). We consider two cases here. In the first case, the response variable Y and the covariates x_1, x_2 are observed in first phase s_1 while covariates z are missing and observed in s_2 only. In the second case, we observe the values of a response variable Y and the covariate x_1 while x_2 and z are missing in first phase and observed in second phase sample only. We use $\varepsilon \sim N(0, 1)$ and $\varepsilon \sim N(0, 1.5)$ for model (B1) for the case 1 and 2, respectively. The covariates x_1 and x_2 are generated independently from a uniform distribution with parameters $x_1 \sim (20, 160)$ and $x_2 \sim (25, 100)$ for case 1 and $x_1 \sim (20, 160)$ and $x_2 \sim (5, 20)$ for case 2. The covariates z are generated from a Bernoulli distribution using both variables x_1 and x_2 in response functions similar as in the univariate case. The ratio of the prediction errors are shown in Figs. 11 and 12 for case 1, and 13 and 14 for case 2 and the median values of mean squared prediction error for both cases are given in Table 5. The overall interpretation remains unchanged. The bias and the estimated variance of the regression coefficient $\hat{\beta}_2$ are given in Table 6. The results are similar to those for univariate x as discussed in Sect. 3.1.

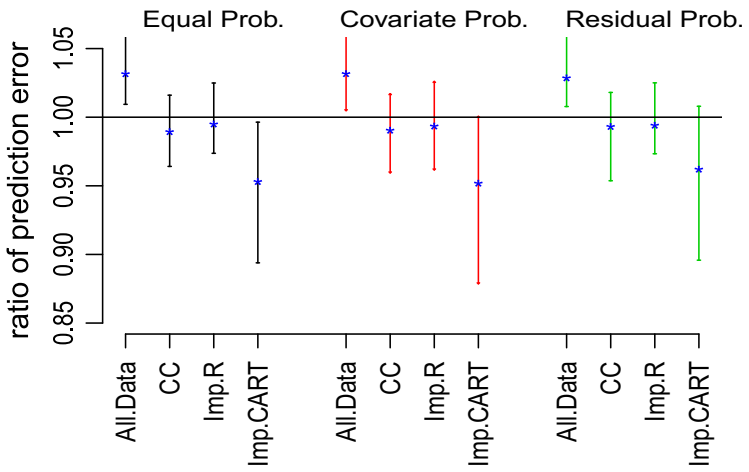

Fig. 13 $n_2 = 200$: ratio of mean squared prediction error for simulated data. Second phase sample selection with equal, Tille covariate and Tille residual dependent probability sampling with case 2

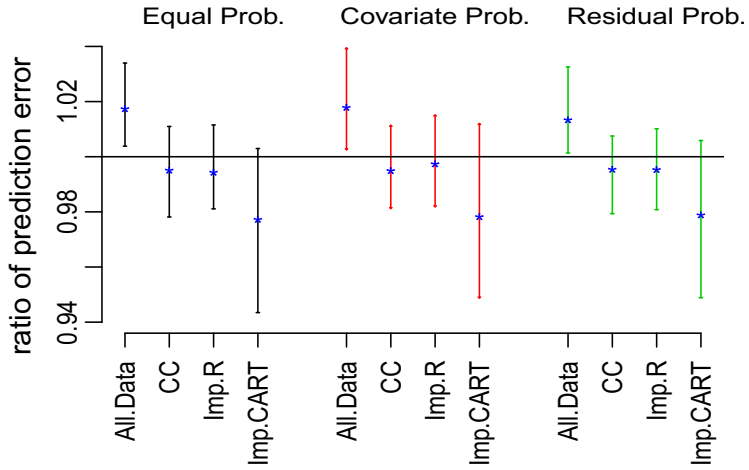

Fig. 14 $n_2 = 400$: ratio of mean squared prediction error for simulated data. Second phase sample selection with equal, Tille covariate and Tille residual dependent probability sampling with case 2

Table 5 Median of mean squared prediction error for simulated data (for bivariate x)

		$n_2 = 200$				$n_2 = 400$					
	Sampling	Prop.Me	Imp.R	Imp.CART	CC	Prop.Me	Imp.R	Imp.CART	CC	S1.Yx	All.Data
case 1	a	1.044	1.052	1.120	1.063	1.028	1.040	1.064	1.044	1.536	1.010
	b	1.056	1.053	1.118	1.073	1.029	1.037	1.067	1.036	1.536	1.010
	c	1.050	1.053	1.103	1.059	1.026	1.035	1.058	1.036	1.536	1.010
case 2	a	2.349	2.364	2.458	2.373	2.317	2.332	2.361	2.331	2.786	2.271
	b	2.342	2.352	2.469	2.368	2.310	2.322	2.365	2.317	2.786	2.271
	c	2.330	2.338	2.421	2.344	2.299	2.312	2.349	2.315	2.786	2.271

a: Equal probability
b: Tille covariate dependent
c: Tille residual dependent

Table 6 Bias and estimated variance for regression coefficient $\hat{\beta}_2$ for simulated data (for bivariate x)

			$n_2 = 200$				$n_2 = 400$			
		Sampling	Prop.Me	Imp.R	Imp.CART	CC	Prop.Me	Imp.R	Imp.CART	CC
bias($\hat{\beta}_2$)	case 1	a	−0.004	−0.148	−0.450	−0.036	−0.027	−0.140	−0.329	−0.044
		b	0.029	−0.104	−0.415	−0.016	−0.024	−0.123	−0.324	−0.048
		c	−0.006	−0.134	−0.436	−0.028	−0.029	−0.132	−0.307	−0.041
	case 2	a	−0.042	−0.136	−0.431	−0.060	−0.044	−0.105	−0.308	−0.048
		b	−0.020	−0.098	−0.481	−0.042	−0.025	−0.083	−0.305	−0.029
		c	0.036	−0.109	−0.394	−0.048	−0.045	−0.081	−0.289	−0.047
est.var($\hat{\beta}_2$)	case 1	a	0.025	0.047	0.240	0.028	0.014	0.035	0.124	0.015
		b	0.035	0.040	0.214	0.032	0.012	0.025	0.121	0.014
		c	0.024	0.049	0.227	0.027	0.013	0.028	0.109	0.016
	case 2	a	0.060	0.085	0.260	0.065	0.037	0.053	0.151	0.039
		b	0.065	0.077	0.307	0.066	0.034	0.043	0.138	0.038
		c	0.054	0.074	0.213	0.056	0.024	0.033	0.118	0.025

a: Equal probability
b: Tille covariate dependent
c: Tille residual dependent

References

Anderson TW (1957) Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. J Am Stat Assoc 52(278):200–203

Boor CD (1972) On calculating with B-splines. J Approx Theory 6(1):50–62

Carpenter JR, Kenward M (2013) Multiple imputation and its applications, 1st edn. Wiley, Chichester

Deville JC, Tille Y (1998) Unequal probability sampling without replacement through a splitting method. Biometrika 85(1):89–101

Donders ART, van der Heijden GJMG, Stijnen T, Moons KGM (2006) Review: a gentle introduction to imputation of missing values. J Clin Epidemiol 59(10):1087–1091

Eilers PHC, Marx BD (1996) Flexible smoothing with B-splines and penalties. Stat Sci 11(2):89–121

Fahrmeir L, Gieger C, Klinger A (1998) Econometrics in theory and practice. Physica-Verlag, Heidelberg

Fitzenberger B, Fuchs B (2017) The residency discount for rents in Germany and the tenancy law reform act 2001: evidence from quantile regressions. German Econ Rev 18(2):212–236

Hanif M, Brewer KRW (1980) Sampling with unequal probabilities without replacement: a review. Int Stat Rev 48(3):317–335

Hayati RP, Lee KJ, Simpson JA (2015) The rise of multiple imputation: a review of the reporting and implementation of the method in medical research. Med Res Methodol 15 30

Horton NJ, Kleinman KP (2007) Much ado about nothing: a comparison of missing data methods and software to fit incomplete data regression models. Am Stat 61(1):79–90

Ibrahim JG, Chen MH, Lipsitz SR, Herring AH (2005) Missing data methods for generalizes linear models: a comparative review. J Am Stat Assoc 100(469):332–346

Lawless JF, Kalbeisch JD, Wild CJ (1999) Semiparametric methods for response selective and missing data problems in regression. J R Stat Soc 61(2):413–438

Liang H (2008) Generalized partially linear models with missing covariates. J Multivar Anal 99(5):880–895

Liang H, Wang S, Robins JM, Carroll RJ (2004) Estimation in partially linear models with missing covariates. J Am Stat Assoc 99(466):357–367

Little RJA (1992) Regression with missing X's: a review. J Am Stat Assoc 87(420):1227–1237

Little R, An H (2004) Robust likelihood-based analysis of multivariate data with missing values. Stat Sin 14(3):949–968

Little RJA, Rubin DB (2002) Statistical analysis with missing data, 2nd edn. Wiley, New York
Lumley T (2017) Robustness of semiparametric efficiency in nearly-true models for two-phase samples. arXiv:1707.05924
Mandallaz D, Breschan J, Hill A (2013) New regression estimators in forest inventory with two phase sampling and partially exhaustive information: a design based monte carlo approach with applications to small area estimation. Can J For Res 43(11):1023–1031
Mcleish DL, Struthers CA (2006) Estimation of regression parameters in missing data problems. Can J Stat 34(2):233–259
Meng XL (2000) Missing data: dial m for ??? J Am Stat Assoc 95(452):1325–1330
Mitra R, Reiter JP (2016) A comparison of two methods of estimating propensity scores after multiple imputation. Stat Methods Med Res 25(1):188–204
O'Sullivan F (1986) A statistical perspective on ill-posed inverse problems. Stat Sci 1(4):502–518
Qin G, Zhu Z, Fung WK (2012) Robust estimation of the generalised partial linear model with missing covariates. J Nonparametric Stat 24(2):517–530
Robins JM, Rotnitzky A, Zhao LP (1994) Estimation of regression coefficients when some regressors are not always observed. J Am Stat Assoc 89(427):846–866
Robins JM, Rotnitzky A, Zhao LP (1995) Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. J Am Stat Assoc 90(429):106–121
Ruppert D, Wand MP, Carroll RJ (2003) Semiparametric regression. Cambridge University Press, Cambridge
Ruppert D, Wand MP, Carroll RJ (2009) Semiparametric regression during 2003–2007. Electron J Stat 3:1193–1256
Saegusa T (2014) Bootstrapping two-phase sampling. e-print https://arxiv.org/abs/1406.5580v1
Saegusa T (2015) Variance estimation under two phase sampling. Scand J Stat 42(4):1078–1091
Stasinopoulos DM, Rigby RA, Heller GZ, Voudouris V, De Bastiani F (2017) Flexible regression and smoothing: using GAMLSS in R. Chapman and Hall/CRC, Boca Raton
Thompson SK (2012) Sampling, 3rd edn. Wiley, New York
Tille Y (1996) An elimination procedure of unequal probability sampling without replacement. Biometrika 83(1):238–241
Tille Y (2006) Sampling algorithms. Springer, New York
Tille Y, Matei A (2016) The R package sampling. The comprehensive R archive network. http://cran.r-project.org/
Toutenburg H, Nittner T (2002) Linear regression models with incomplete categorical covariates. Comput Stat 17:215–232
van Buuren S, Groothuis-Oudshoorn K (2011) mice: multivariate imputation by chained equations in R. J Stat Softw 45(3):1–67
Wand MP (2003) Smoothing and mixed models. Comput Stat 18(2):223–249
Wang QH (2009) Statistical estimation in partial linear models with covariate data missing at random. Ann Inst Stat Math 61(1):47–84
Wood SN (2017) Generalized additive models—an introduction with R, 2nd edn. CRC Press, Boca Raton
Yang S, Kim JK (2016) Fractional imputation in survey sampling: a comparative review. Stat Sci 31(3):415–432
Zhang G, Little R (2009) Extensions of the penalized spline of propensity prediction method of imputation. Biometrics 65(3):911–918
Zhang Z, Rockette HE (2005) On maximum likelihood estimation in parametric regression with missing covariates. J Stat Plan Inference 134(1):206–223
Zhang N, Chen H, Elliott M (2016) Nonrespondent subsample multiple imputation in two-phase sampling for nonresponse. J Off Stat 32(3):769–785
Zhao Y, Lawless JF, Mcleish DL (2009) Likelihood methods for regression models with expensive variables missing by design. Biom J 51(1):123–136

A.3:

Ali, M. and Kauermann, G. (2021b). A Split Questionnaire Survey Design in the Context of Statistical Matching. Accepted for publication in *Journal of Statistical Methods and Applications.*

Available online at: https://doi.org/10.1007/s10260-020-00554-2

Statistical Methods & Applications
https://doi.org/10.1007/s10260-020-00554-2

ORIGINAL PAPER

A split questionnaire survey design in the context of statistical matching

Mehboob Ali[1] · **Göran Kauermann**[1]

Accepted: 10 December 2020

Abstract

In this paper, we tackle the problem of splitting a long (potentially time consuming) questionnaire into two parts, where each participant only responds to a fraction of the questions, and all respondents obtain a common portion of questions. We propose a method that combines regression models to the two independent samples (questionnaires) in the survey. Each sample includes the common response variable Y and common covariate x, while two vectors of specific covariates z and w are recorded such that no single sampling unit has answered both z and w. This corresponds to the problem of statistical matching that we tackle under the assumption of conditional independence. In the statistical matching context, we use a macro approach to estimate parameters of a regression model. This means that we can estimate the joint distribution of all variables of interest with available data utilizing the assumption of conditional independence. We make use of this here by fitting three regression models with the same response variable for each model. Combining the three models allows us to obtain a prediction model with all covariates in common. We compare the performance of our proposed method in simulation studies as well as a real data example. Our method gives better results as compared to commonly used alternative methods. The proposed routine is easy to apply in practice and it neither requires the formulation of a model for the covariates itself nor an imputation model for the missing covariates vectors z and w.

Keywords Split questionnaire design · Statistical matching · Conditional independence · Rental guide survey

✉ Mehboob Ali
mehboob.ali@stat.uni-muenchen.de

Göran Kauermann
goeran.kauermann@stat.uni-muenchen.de

[1] Department of Statistics, Ludwig-Maximilians-University, Munich, Germany

Published online: 16 February 2021

1 Introduction

The aim of this paper is to limit the respondent's burden by reducing the length of a questionnaire. Since long questionnaires usually discourage potential respondents and, hence, lead to high non-response rates, split questionnaires occur as an alternative to reduce respondent's burden (Raghunathan and Grizzle 1995; Chipperfield and Steel 2009; Kamgar and Navvabpour 2017). When a questionnaire is (too) long, the participant's often lose interest in surveys, making the response quality low and less accurate (Peytchev and Peytcheva 2017). The response quality can be increased by using a split questionnaires design (SQD) (Stuart and Yu 2019). The SQD produces results similar to a full questionnaire with one drawback; the power is decreased due to a decreased number of observations for each variable (Raghunathan and Grizzle 1995).

Raghunathan and Grizzle (1995) propose a SQD, which is an extension of multiple matrix sampling, where items are assigned in such a way that all bivariate associations are estimable. This means that SQD selects two or more independent samples, with partial overlap in different components. The other commonly used SQD is a 3-form design that divides the survey questionnaire into four components (Graham et al. 2006). In this design, each respondent participates in some common portion and two of the three other components are only provided to segments of the respondents. Kim et al. (2016) discuss a SQD where they select a random sample, s, from the population of interest and then split this sample into two sub-samples s_a and s_b such that $s_a \cup s_b = s$ and $s_a \cap s_b = \phi$.

We use a SQD similar to Kim et al. (2016) to collect data, with the aim of fitting a regression model for a continuous response variable Y regressed on several covariates. The aim is to calculate a prediction model with preferably small prediction error based on the observed data from the SQD. Throughout the entire paper, x represents a continuous covariate while both z and w are denoted as vectors of categorical covariates. The questionnaire is divided into various components so that each participant needs to respond only a fraction of the total components. This means that some variables are not observed in some components by sampling design. We split the full questionnaire into two parts; that is, we draw two independent random samples s_a and s_b from the same population. The information about common variables Y and x is observed from both samples, while the information about the specific variables z is available in sample s_a only and w is recorded in sample s_b only. This means that there is no single sampling unit which provides information of both covariates z and w simultaneously. The resulting data structure is sketched in Table 1. The procedure can be easily extended to more than two groups, but for the sake of simplicity we here constrain the presentation to two groups only.

The integration of two (or more) independent samples or data sources in order to estimate a joint distribution of all the variables of interest is usually known as statistical matching (or data fusion) problem. There are two commonly used ways to pursue statistical matching: the macro approach and the micro approach. The macro approach is to estimate the joint distribution of (y, x, z, w) or any of its characteristics of interest. In this approach, we do not need to create a complete synthetic data set, but instead we make use of conditional independence (CI) assumptions (D'Orazio et al.

Table 1 A split questionnaire survey design in the context of statistical matching

	Common variables	Component 1	Component 2
Samples	Y, x	z	w
s_a	Observed	Observed	Missing
s_b	Observed	Missing	Observed

2006a; Endres 2019). The micro approach is to create complete but synthetic data of all variables of interest (D'Orazio et al. 2006b). If both approaches are used together, this is known as the mixed approach (D'Orazio et al. 2006a). In this paper, we pursue the macro approach to estimate the joint distribution of all the variables of interest.

Statistical matching can be considered as a special type of missing data problem, since there does not exist a single observation which simultaneously contains complete information of all the specific variables of interest (z and w). Since the missingness is induced by the sampling design, this is a missing completely at random constellation (Little and Rubin 2002). Rubin (1986) considers file concatenation as a missing data problem where some variables are not jointly observed. He studies the situation where variables of interest are present in two different surveys (i.e. information on some variables can be obtained from a specific survey whereas information on other variables can be observed from another survey) or it is not possible to observe complete information in one survey. The missing data are then imputed with the multiple imputation method, which does not depend on the assumption of CI. Moriarity and Scheuren (2001) assume CI and multivariate normality to develop a theoretical framework for statistical matching. Rässler's (2002) multiple imputation methods rely on an explicit Bayesian model. Rendall et al. (2013) use cross survey multiple imputation to combine the regression estimation where more covariates are observed in one survey but less observations than the other survey. They exclude those variables which do not jointly provide the information for the analysis model for the survey with a large set of covariates. Kim et al. (2016) use a fractional imputation approach to fill the missing values of specific variables which are not jointly observed.

One particular type of multiple imputation is multiple imputation by chained equations (mice), which is more flexible to cope with the missing data problem variable-by-variable. For example, Rässler (2004); Kaplan and McCarty (2013); and Kamgar et al. (2018) used the mice approach to impute the missing values in a statistical matching problem. Multiple imputation can also rely on pure prediction models like tree based models. Classification and regression trees (CART) can be used for mice. The other tree based model like random forest can also be used to impute the missing values with mice which is an extension of CART (Burgette and Reiter 2010) and more flexible than CART.

All the SQDs and the articles cited above mostly deal with missing values in the data with some kind of imputation. We propose a SQD in the context of statistical matching as shown in Table 1. To solve the identification problem, our proposal relies on the assumption of CI. This assumption helps us to yield an identifiable model with available data (D'Orazio et al. 2006a). The CI is a strong assumption and cannot be tested with

available data. The results of simulation studies and real data applications tend to be similar when we assume strong CI. However it is worth mentioning that in a data set where the CI cannot be easily made, the results shown in Sect. 3 via simulation study and real data application may not be the same. If the CI assumption does not hold true, this leads to serious bias in the resulting joint relationships among variables of interest. This problem can be solved using available auxiliary information (Singh et al. 1993; Vantaggi 2008). Kim and Park (2019) proposed a mixed statistical matching method under the CI assumption which does not depend on available auxiliary information.

The CI assumption is frequently used to integrate different sources of data. For example, Donatiello et al. (2016) use the CI assumption to combine the information of two surveys: Statistics on Income and Living Condition, and Household Budget Survey. Endres and Augustin (2016) propose a Bayesian approach using the CI assumption to analysis German Socio-Economic Panel Survey. They divide the survey data into two groups to see the performance of their proposed method. Endres and Augustin (2019) use log-linear Markov networks to integrate German Socio-Economic Panel Survey under the CI assumption. Kim and Park (2019) propose a mixed approach under the CI assumption using multinomial logistic regression models. The method can be used for both, the micro and macro approaches in statistical matching. Recently, Cutillo and Scanu (2020) propose a mixed approach under the CI assumption to study the relationship between two surveys: Statistics on Income and Living Condition, and Household Budget Survey. Available softwares which assume CI to analyze two or more independent data sources are, `mice`, Van Buuren and Groothuis-Oudshoorn (2011) and `StatMatch`, D'Orazio (2015). For general review about the CI assumption, we refer to D'Orazio et al. (2006a) and Doretti et al. (2018).

We use real data of a rental guide survey, which is an official instrument and regularly run in different cities of Germany to control the rent market (Fahrmeir et al. 2013 or Fitzenberger and Fuchs 2017). In our example, we consider the rental guide survey for Munich. Let Y be the rent (Euro per square meter) as a response variable, the continuous covariate x is the floor space (square meter) and z and w be the categorical covariates describing the quality and facilities of the apartment (see Appendix for a complete list of all the variables). The recording of covariates z and w is both time consuming and expensive, which is why we split the questionnaire into two separate components to collect information of covariates z or w, respectively. The goal is to fit a regression model of the form

$$Y = \beta_{0p} + x\beta_{xp} + z\beta_{zp} + w\beta_{wp} + \epsilon, \tag{1}$$

where β_{zp} is a column vector with the same dimension as row vector z and likewise, β_{wp} is column vector with matching dimension to row vector w. Note that ϵ is a residual from the location-scale family and Y is a continuous response variable, and the covariates x, z and w are as described above. The index p is used for the pooled estimates and will be described in the next section. We estimate this regression model with our proposal and the previously discussed alternative routines (mice regular, CART and random forest) and compare them based on the out of sample prediction error. Our intention is to show that our proposal has smaller prediction error than the alternatives.

The paper is structured as follows. In Sect. 2, we describe the CI assumption and our proposed method, and discuss how to fit the separate regression models without imputing the missing values of covariates z and w, which are not jointly observed. Simulation studies are carried out in Sect. 3, and we compare the performance of the proposed method and multiple imputations methods based on the prediction error of simulated data as well as the real data example. We also compare our proposed routine with alternatives methods based on the bias and root mean squared error of missing covariates. Our findings are discussed in Sect. 4.

2 Proposed method under conditional independence assumption

2.1 Conditional independence assumption

The most commonly used method for missing data is the complete case analysis, which excludes all missing observations from the analysis (Pigott 2001; Little and Rubin 2002). This method is not applicable in the statistical matching context where we do not have a single sampling unit with joint information of all the variables of interest. Due to not jointly having observed variables z and w as shown in Table 1, the joint distribution of (y, x, z, w) is not identifiable. We assume CI to overcome the problem of identification. Specifically, we assume that z and w are conditionally independent given Y and x and x only. That is we assume

$$z \perp\!\!\!\perp w|\{Y, x\} \text{ and } z \perp\!\!\!\perp w|x. \tag{2}$$

Note that, we are interested in estimating the conditional distribution of Y given x, z, w. Given (2), we can estimate all parts of our conditional distribution directly with available data. Following the chain rule, the joint distribution of all variables of interest can be factorized as

$$\begin{aligned} f(z, w, x, y) &= f(z|x, y, w) f(w|x, y) f(y|x) f(x), \\ &= f(z|x, y) f(w|x, y) f(y|x) f(x). \end{aligned} \tag{3}$$

If the CI assumption is true, the available data of both samples, s_a and s_b is sufficient to estimate (3) (Roszka 2015).

2.2 Proposed method

Define a simple random sample s of size n be drawn from the population of interest and divide this sample into two non-overlapping sub-samples s_a and s_b of sizes n_a and n_b respectively. For sample s_a we obtain the data $(Y_j, x_j, z_j) : j \in s_a$, and sample $s_b = s \setminus s_a$. This yields the information $(Y_k, x_k, w_k) : k \in s_b$. We are interested in fitting a regression model for Y regressed on the covariates x, z and w. The conditional

probability of Y given x, z and w equals

$$f(y|x,z,w) = \frac{f(z,w,x,y)}{f(z,w,x)}. \tag{4}$$

Applying the chain rule and the second CI assumption in (2) using (3) and (4) we obtain

$$\begin{aligned} f(y|x,z,w) &= \frac{f(z|x,y)f(w|x,y)f(y|x)f(x)}{f(z|x)f(w|x)f(x)} \\ &= \frac{f(z|x,y)}{f(z|x)}.\frac{f(w|x,y)}{f(w|x)}.f(y|x). \end{aligned} \tag{5}$$

To simplify (5), we follow Little (1992). The distribution of Y and z given x can be decomposed as

$$f(y,z|x) = f(y|x)f(z|y,x). \tag{6}$$

With (6) we can transform the first ratio in (5) into

$$\frac{f(z|x,y)}{f(z|x)} = \frac{f(y,z|x)}{f(z|x)f(y|x)} = \frac{f(y|x,z)}{f(y|x)}. \tag{7}$$

Rearranging the middle term in (5) in the same way using (6) and (7) leads to

$$f(y|x,z,w) := \frac{f(y|x,z)f(y|x,w)}{f(y|x)}. \tag{8}$$

We see that we can split the conditional distribution of Y into conditional distributions with only parts of the covariates given. These in turn can be estimated from the data obtained by the splitted design. For example, $f(y|x,z)$ can be fitted from s_a, $f(y|x,w)$ can be fitted from s_b, and $f(y|x)$ depends only on the common variables and can be fitted using the data from $s = s_a \cup s_b$.

One can, in principle, use any model to estimate the three densities. We here propose to work with simple ordinary least squared method. The estimated version of (8) is then

$$\hat{f}(y|x,z,w) := \frac{\hat{f}(y|x,z)\hat{f}(y|x,w)}{\hat{f}(y|x)}, \tag{9}$$

where the different components in (9) are fitted assuming

$$Y|x,z \sim N\left(\beta_{0a} + x\beta_{xa} + z\beta_{za}, \sigma_a^2\right), \tag{10}$$

$$Y|x,w \sim N\left(\beta_{0b} + x\beta_{xb} + w\beta_{wb}, \sigma_b^2\right), \tag{11}$$

$$Y|x \sim N\left(\beta_0 + x\beta_x, \sigma^2\right). \tag{12}$$

Note that models (10)–(12) hold jointly only in case of a multivariate normal distribution. We do not assume this generally but make use of (10)–(12) as approximation. Using model (9), we propose to:

1. fit model (10) with the data from s_a to get $\hat{f}(y|x, z)$ in (9),
2. fit model (11) with the data from s_b to get $\hat{f}(y|x, w)$ in (9) and
3. fit model (12) with the data from s to get $\hat{f}(y|x)$ in (9).

The corresponding estimates are

1. $\hat{\mu}_a = \hat{\beta}_{0a} + x\hat{\beta}_{xa} + z\hat{\beta}_{za}$ and $\hat{\sigma}_a^2$,
2. $\hat{\mu}_b = \hat{\beta}_{0b} + x\hat{\beta}_{xb} + w\hat{\beta}_{wb}$ and $\hat{\sigma}_b^2$,
3. $\hat{\mu} = \hat{\beta}_0 + x\hat{\beta}_x$ and $\hat{\sigma}^2$.

These estimates are calculated using the ordinary least squares method. The error term of each model from (10) to (12) are assumed to be independent and identically distributed with mean zero and constant variance. Using the normal densities from (10) to (12) we can write down the factorized density (9) as $Y|x, z, w \overset{\text{iid}}{\sim} N(\mu_p, \sigma_p^2)$, where

$$\hat{\sigma}_p^2 = \frac{1}{\hat{\sigma}_a^{-2} + \hat{\sigma}_b^{-2} - \hat{\sigma}^{-2}} \quad \text{and} \quad \hat{\mu}_p = \hat{\sigma}_p^2 \left\{ \frac{\hat{\mu}_a}{\hat{\sigma}_a^2} + \frac{\hat{\mu}_b}{\hat{\sigma}_b^2} - \frac{\hat{\mu}}{\hat{\sigma}^2} \right\}.$$

The index p indicates the pooled estimates. Note that the maximized log likelihood for model (10), (11) and (12) is

$$l(\hat{\beta}, \hat{\sigma}_\cdot^2) = -\frac{n}{2}\log(\hat{\sigma}_\cdot^2)$$

where $\hat{\sigma}_\cdot$ is the ML estimate in model (10), (11) or (12), respectively. Since model (12) is nested in (10) as well as in (11) we have

$$\begin{aligned} &-\frac{n}{2}\log(\hat{\sigma}_a^2) \geq -\frac{n}{2}\log(\hat{\sigma}_\cdot^2) \\ \Leftrightarrow\ &\frac{1}{\hat{\sigma}_a^2} \geq \frac{1}{\hat{\sigma}^2} \end{aligned}$$

and similarly

$$\frac{1}{\hat{\sigma}_b^2} \geq \frac{1}{\hat{\sigma}^2}$$

This proves that

$$\hat{\sigma}_p^2 = \frac{1}{\hat{\sigma}_a^{-2} + \hat{\sigma}_b^{-2} - \hat{\sigma}^{-2}} \geq 0,$$

where the equality holds if both $\hat{\beta}_{za} \equiv 0$ and $\hat{\beta}_{wb} \equiv 0$, which happens with probability 0.

The above estimates can then be combined to provide the final fit for the original regression model (1) through $\hat{\beta}_{0p} + x\hat{\beta}_{xp} + z\hat{\beta}_{zp} + w\hat{\beta}_{wp}$, where

$$\hat{\beta}_{0p} = \hat{\sigma}_p^2 \left\{ \frac{\hat{\beta}_{0a}}{\hat{\sigma}_a^2} + \frac{\hat{\beta}_{0b}}{\hat{\sigma}_b^2} - \frac{\hat{\beta}_0}{\hat{\sigma}^2} \right\}, \hat{\beta}_{xp} = \hat{\sigma}_p^2 \left\{ \frac{\hat{\beta}_{xa}}{\hat{\sigma}_a^2} + \frac{\hat{\beta}_{xb}}{\hat{\sigma}_b^2} - \frac{\hat{\beta}_x}{\hat{\sigma}^2} \right\},$$
$$\hat{\beta}_{zp} = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_a^2}\hat{\beta}_{za} \quad \text{and} \quad \hat{\beta}_{wp} = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_b^2}\hat{\beta}_{wb}.$$

We refer to Kauermann and Ali (2020) for general discussion in this direction for a two phase sampling scheme.

From the formula above we can also derive when the proposed approach is useful. We therefore consider the Kullback–Leibler (KL) divergence between the true density $f_{Y|x,z,w}$ and its approximation using the CI assumption, i.e. (8). Assuming the approximate models (10)–(12) we get

$$\begin{aligned} \text{KL} &= \int \log\{f_{Y|x,z,w}\} dF_{Y|x,z,w} - \int \log \frac{f_{Y|x,z} f_{Y|x,w}}{f_{Y|x}} dF_{Y|x,z,w} \\ &= -\frac{1}{2}\{\log(\sigma_Y^2) + \log(\sigma_a^2) + \log(\sigma_b^2) - \log(\sigma^2)\}, \end{aligned}$$

with $\sigma_Y^2 = \text{Var}(Y|x, z, w)$. If z and w are not the most important covariates, as in our example, we have that σ_Y^2 is moderately larger than σ_a^2, σ_b^2 and σ^2. In other words, the KL divergence is small. In contrast, if z (or w) would be the most important covariates, then $\sigma_Y^2 \ll \sigma_a^2$ (or $\sigma_Y^2 \ll \sigma_b^2$) leading to a high KL divergence. Hence, the approach suggested here is useful only if the most relevant covariates are always observed.

3 Simulation and example

3.1 Simulation

To show the performance of our proposed method, we generate data from the regression model (1), where $\varepsilon \sim N(0, \sigma_\epsilon^2)$, Y is a continuous response, x is a continuous covariate, and $z = (z_1, z_2)$ and $w = (w_1, w_2)$ are vectors of binary covariates which are correlated with x. In the simulation we vary the distribution of continuous covariate x and modify the size of the full (split) questionnaire(s). We compare the proposed routine with three alternatives. First, we make use of regular multiple imputation approach in the `mice` package in R, (see Van Buuren and Groothuis-Oudshoorn 2011). In our case the regular mice setting for binary variables is logistic regression. Secondly, we make use of classification and regression tree (CART). Lastly, we use random forest multiple imputations. We use the default number of imputations for all the imputation methods in the `mice` package. For the statistical matching problem, the mice assumes CI between specific variables given the common variables to impute the missing values of these specific variables (i.e. z and w).

We generate $N = 5000$ values as a super-population. The parameter values used in simulation are $\beta_{xp} \in \{1.32\}$, $\beta_{zp} \in \{1.15, -1.80\}$ and $\beta_{wp} \in \{1.98, 1.52\}$. The distributional assumptions for x are: $x \sim \{\text{uniform}(20, 160), \text{normal}(80, 18), \text{log-normal}(5.5, 0.5)\}$, $\sigma_\epsilon^2 \in \{1.2\}$. The sample sizes are $n \in \{500, 1000\}$ and $n_a = n_b = n/2$. Covariates z and w are drawn such that both of these binary covariates and x are correlated of order 0.4, respectively. For all the simulation scenarios we focus on the prediction error of the fitted model in the population data (i.e. out of sample prediction), and compare this with the three imputation alternatives. We use $N - n$ observation to calculate the out of sample prediction error.

For the presentation of our results, we use the following abbreviations: "FQ" stands for full questionnaire, where we hypothetically assume that the missing covariates z and w are fully observed in both samples, "Prop.Me" indicates our proposed method, "Imp.R" indicates imputation using the regular mice setting, "Imp.CART" indicates imputation for CART mice and "Imp.RF" indicates imputation for random forest mice.

To calculate a prediction error we use cross-validation. That is, we split the population into two parts: the sample of size $n = 500$ is used to fit the regression models according to the proposed method and the remaining $N - n$ values are used for prediction. Applying the proposed method and alternate routines, the median and standard deviation of mean squared prediction error of 120 simulations are produced in Table 2 (left side). Our method gives smaller average prediction error with low variation compared to alternative methods. We calculate the ratio of the prediction error by dividing the mean squared prediction error of the proposed method in each simulation by the corresponding mean squared prediction error resulting from each competing method. Figure 1 shows the median of the ratio of the prediction error of the alternative routines compare to our proposed approximate method. The horizontal line indicates the value 1, and values below this line speak in favor of our proposal. The vertical bars include the inner 90 percent range of the ratios for all simulations. The cross mark on bars represent the median values of the ratios of prediction error. A clear dominance of the approximate routine is apparent for nearly all simulation settings. Only for the simulation scenario where $x \sim$ normal with regular imputation routine shows a small difference.

We now repeat the simulation procedure with $n = 1000$. The results are shown in Table 2 (right side) and Fig. 2. Again the approximate routine outperforms the alternative routines. We also reported the bias and root mean squared error (RMSE) of missing covariates in Figs. 3 and 4, respectively, which are estimated as

$$\text{bias}(\hat{\beta}_{zp}) = E(\hat{\beta}_{zp}) - \beta_{zp} \quad \text{and} \quad \text{RMSE}(\hat{\beta}_{zp}) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(\hat{\beta}_{zpi} - \beta_{zp})^2}$$

where $E(\hat{\beta}_{zp}) = \frac{1}{m}\sum_{i=1}^{m}\hat{\beta}_{zpi}$ and β_{zp} is the corresponding true value used in the simulation model (1) and m is number of simulations. In the same way we estimate $\text{bias}(\hat{\beta}_{wp})$ and $\text{RMSE}(\hat{\beta}_{wp})$. In Fig. 3, the horizontal line indicates the value 0, and values close to this line show better results. The solid and the dashed lines are used for $n \in \{500, 1000\}$, respectively. Our proposed method produced very small amount of bias as compared to multiple imputation methods. In Fig. 4, we can see that our

Table 2 Median and standard deviation (SD) of mean squared prediction error for simulated data when $n = (500, 1000)$

	Distribution	$n_a = n_b = n/2 = 250$					$n_a = n_b = n/2 = 500$				
		Prop.Me	Imp.R	Imp.CART	Imp.RF	FQ	Prop.Me	Imp.R	Imp.CART	Imp.RF	FQ
Median	$x \sim$ uniform	1.608	2.009	2.013	2.154	1.463	1.584	2.000	2.002	2.142	1.456
	$x \sim$ log-normal	1.587	2.024	2.023	2.144	1.421	1.564	2.004	2.004	2.111	1.414
	$x \sim$ normal	1.579	1.620	1.751	2.063	1.420	1.553	1.566	1.644	2.016	1.415
SD	$x \sim$ uniform	0.046	0.087	0.090	0.093	0.016	0.038	0.076	0.081	0.080	0.017
	$x \sim$ log-normal	0.049	0.086	0.094	0.087	0.014	0.037	0.070	0.077	0.075	0.015
	$x \sim$ normal	0.048	0.053	0.079	0.093	0.015	0.032	0.040	0.051	0.067	0.015

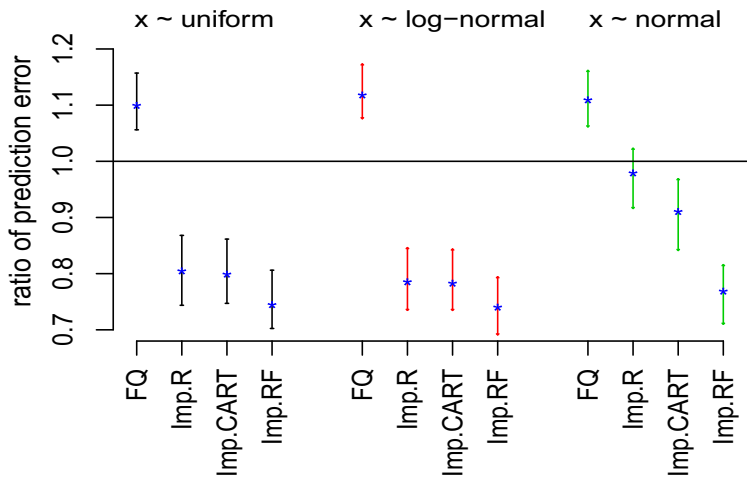

Fig. 1 $n_a = n_b = n/2 = 250$: Ratio of mean squared prediction error for simulated data

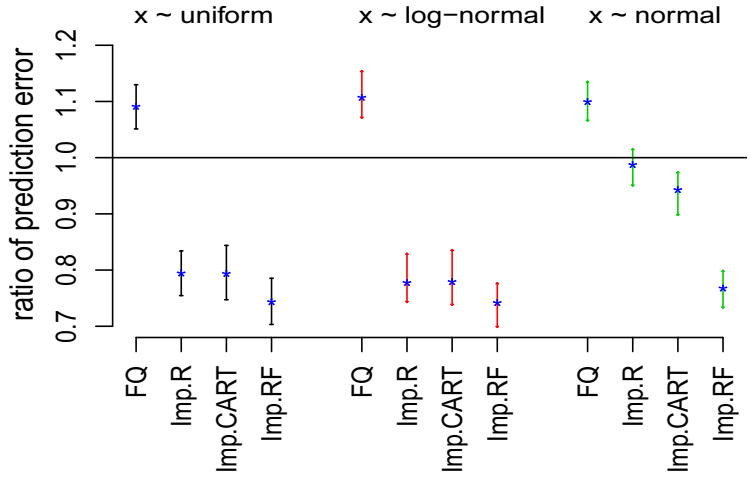

Fig. 2 $n_a = n_b = n/2 = 500$: Ratio of mean squared prediction error for simulated data

proposed routine outperforms the different multiple imputation methods based on RMSE of missing covariates. Our proposed method also shows better RMSE results as compared to alternatives methods when we increased sample size from 500 to 1000.

To demonstrate the effect of unequal sample sizes of a split questionnaire design, we divided $n = 500$ into two sub-samples where first and second samples contain $n_1 = 300$ and $n_2 = 200$ sampling units, respectively. The results of median and standard deviation of mean squared prediction error are shown in Table 3 and ratio of this prediction error in Fig. 5. The overall results remain unchanged.

3.2 Rent data example

In order to apply our proposed method to a real data example, we use the Munich rent survey data 2017 as our finite population. This population contains 3024 observations with the continuous response variable rent per square meter Y (in Euros), a continuous covariate the floor space x and vectors of binary covariates z and w describing the quality and the facilities of the apartment (see "Appendix" for a complete list of

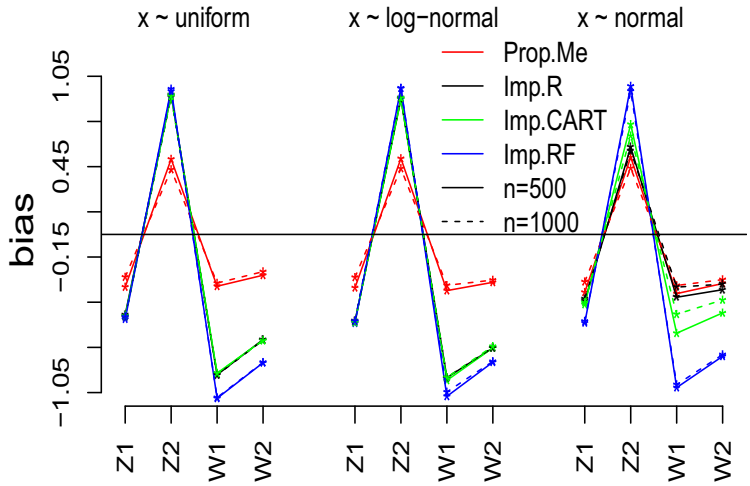

Fig. 3 $n_a = n_b = n/2 = (250, 500)$: Bias for missing covariates

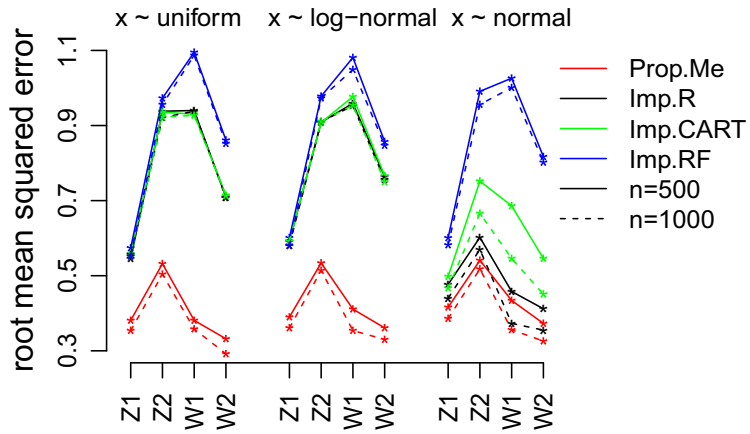

Fig. 4 $n_a = n_b = n/2 = (250, 500)$: Root mean squared error (RMSE) for missing covariates

Table 3 Median and standard deviation (SD) of mean squared prediction error for simulated data when $n = 500$, $n_a = 300$ and $n_b = 200$

	Distribution	Prop.Me	Imp.R	Imp.CART	Imp.RF	FQ
Median	$x \sim$ uniform	1.606	1.994	2.003	2.117	1.463
	$x \sim$ log-normal	1.596	2.019	2.021	2.104	1.421
	$x \sim$ normal	1.577	1.627	1.752	2.049	1.420
SD	$x \sim$ uniform	0.051	0.092	0.097	0.092	0.016
	$x \sim$ log-normal	0.051	0.096	0.103	0.096	0.014
	$x \sim$ normal	0.049	0.059	0.087	0.091	0.015

variables). The covariate x is non linearly related with response variable Y as shown in Fig. 6 (upper side), therefore we use a inverse transformation on this covariate like as $\frac{1}{x}$ (see Fahrmeir et al. 2013, Chap. 2). This transformation makes the regression model linear as shown in Fig. 6 (lower side) and these residuals are calculated based on entire population data.

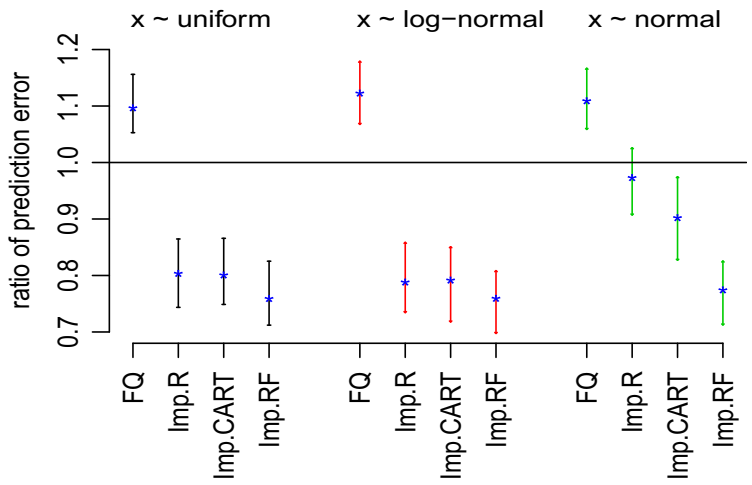

Fig. 5 $n = 500$, $n_a = 300$ and $n_b = 200$: Ratio of mean squared prediction error from simulated data for unequal sample sizes

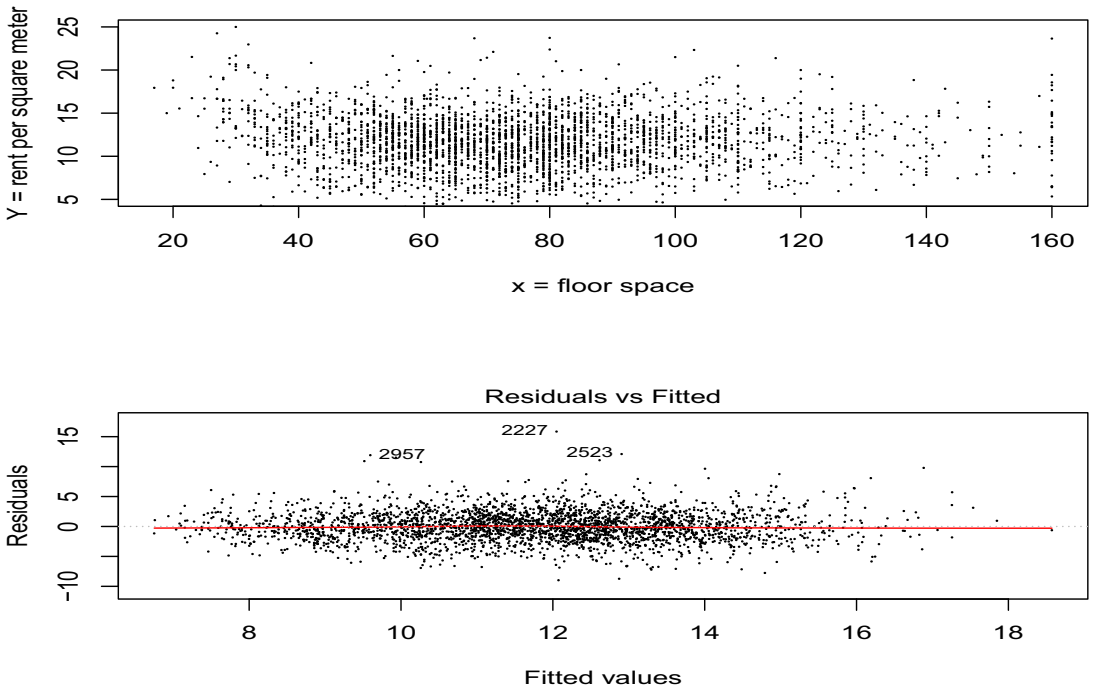

Fig. 6 Rent per square meter relation with floor space (upper side) and linear regression model residuals against fitted values after inverse transformation on floor space (x) (lower side)

To measure the performance of the proposed method, we draw a simple random sample without replacement of size $n = 500$ from the population and divide this sample into two equal non-overlapping parts (or draw two non-overlapping simple random samples of sizes $n_a = n_b = 250$ from the same population). The common variables Y and x are observed in both the samples, while the information of covariates z are observed in s_a only and w in s_b only. We repeat this process 50 times to obtain 50 simulated samples. We measure out of sample prediction error based on the $3024 - 500 = 2524$ apartments not included in the sample. The median and standard deviation of the mean squared prediction error for 50 samples are given in Table 4 (left side)

Table 4 Median and standard deviation (SD) of mean squared prediction error for rent data when $n = (500, 1000)$

	$n_a = n_b = n/2 = 250$					$n_a = n_b = n/2 = 500$				
	Prop.Me	Imp.R	Imp.CART	Imp.RF	FQ	Prop.Me	Imp.R	Imp.CART	Imp.RF	FQ
Median	6.512	6.543	6.628	6.533	6.337	6.394	6.463	6.477	6.429	6.279
SD	0.166	0.221	0.206	0.166	0.128	0.216	0.228	0.236	0.249	0.216

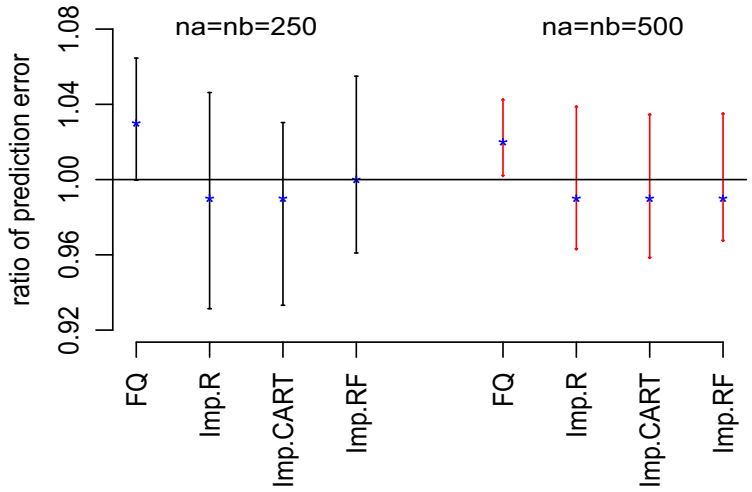

Fig. 7 $n_a = n_b = n/2 = (250, 500)$: Ratio of mean squared prediction error for rent data

and ratio of prediction error are shown in Fig. 7 (left side). To see the effect of sample size we increase the size of full questionnaire $n = 1000$ sampling units, similar to the simulation study. The results are shown in Table 4 and Fig. 7 (right side).

4 Discussion

In this paper, we propose a simple method to reduce the respondent's burden by splitting a long questionnaire and select two independent random samples in such a way that certain covariates are not jointly observed. We are interested in estimating the classical linear regression model Y given x, z, w to calculate the out of sample prediction error. This regression model fails when no single sampling unit has the information of specific covariates simultaneously, so we can not apply the complete case analysis on this data directly. To overcome this problem, we apply a CI assumption to factor the joint distribution into different sub factors. Through this factorization, we are able to estimate the classic linear regression model (1) with the available splitted data without having to use any imputation procedures.

We showed in simulations and a real data example that the proposed approach performs better with respect to prediction error and it does not require a specific imputation model for missing data. We assume the CI for our method, which is not testable with available data. If the specific variables are closely related to the common variables then this assumption is reasonable. In our rent data example, the specific covariates like the quality of kitchen in an apartment or bathroom equipment are likely to depend on the floor space of the apartment (common covariate x). Hence, the CI assumption is reasonable in our real data example.

Appendix: Variables list for rent data example

Common variables Y, x	Component 1 z	Component 2 w	Samples
Y = rent per square meter (in Euros), x = the floor space	z1 = 1 if the apartment does not have an upmarket kitchen, z2 = 1 if the apartment has an open kitchen, z3 = 1 if the apartment lies in an apartment type building, z4 = 1 if the apartment lies in an old building, z5 = 1 if the apartment is located in a back premises, z6 = 1 if apartment has standard central heating, z7 = 1 if the apartment has under floor heating	Missing	S_a
	Missing	w1 = 1 if the apartment has good bathroom equipment, w2 = 1 if the apartment lies in an average residential location, w3 = 1 if the apartment has a second rest room, w4 = 1 if the apartment has a new floor, w5 = 1 if the apartment has a bad floor, w6 = 1 if the apartment has a good floor, w7 = 1 if the apartment lies in a ground floor	S_b

References

Burgette LF, Reiter JP (2010) Multiple imputation for missing data via sequential regression trees. Am J Epidemiol 172(9):1070–1076

Chipperfield JO, Steel DG (2009) Design and estimation for split questionnaire surveys. J Offic Stat 25(2):227–244

Cutillo A, Scanu M (2020) A mixed approach for data fusion of HBS and SILC. J Soc Indic Res. https://doi.org/10.1007/s11205-020-02316-9

Donatiello G, D'Orazio M, Frattarola D, Rizzi A, Scanu M, Spaziani M (2016) The role of the conditional independence assumption in statistically matching income and consumption. Stat J IAOS 32:667–675

D'Orazio M (2015) Integration and imputation of survey data in R: the StatMatch package. J Rom Stat Rev 2:57–68

D'Orazio M, Di Zio M, Scanu M (2006a) Statistical matching: theory and practice. Wiley, New York

D'Orazio M, Di Zio M, Scanu M (2006b) Statistical matching for categorical data: displaying uncertainty and using logical constraints. J Offic Stat 22:137–157

Doretti M, Geneletti S, Stanghellini E (2018) Missing data: a unified taxonomy guided by conditional independence. Int Stat Rev 86(2):189–204

Endres E (2019) Statistical matching meets probabilistic graphical models: contributions to categorical data fusion. Ph.D. Dissertation. Ludwig-Maximilians-University Munich

Endres E, Augustin T (2016) Statistical matching of discrete data by Bayesian networks. Proc Eight Int Conf Probabil Graph Mod Proc Mach Learn Res 52:159–170

Endres E, Augustin T (2019) Utilizing log-linear Markov networks to integrate categorical data files, Technical Report 222. Department of Statistics, LMU Munich

Fahrmeir L, Kenib T, Lang S, Marx B (2013) Regression-models, methods and applications. Springer, Berlin

Fitzenberger B, Fuchs B (2017) The residency discount for rents in Germany and the tenancy law reform act 2001: evidence from quantile regressions. German Econ Rev 18(2):212–236

Graham JW, Taylor BJ, Olchowski AE, Cumsille PE (2006) Planned missing data designs in psychological research. Psychol Methods 11(4):323–343

Kamgar S, Navvabpour H (2017) An efficient method for estimating population parameters using split questionnaire design. J Stat Res Iran 14(1):77–99

Kamgar S, Meinfelder F, Münnich R (2018) Estimation within the new integrated system of household surveys in Germany. J Stat Pap 1–27

Kaplan D, McCarty AT (2013) Data fusion with international large scale assessments: a case study using the OECD PISA and TALIS surveys. Large-Scale Assess Educ. https://doi.org/10.1186/2196-0739-1-6

Kauermann G, Ali M (2020) Semi-parametric regression when some (expensive) covariates are missing by design. J Stat Pap 1–22. https://doi.org/10.1007/s00362-019-01152-5

Kim K, Park M (2019) Statistical micro matching using a multinomial logistic regression model for categorical data. Commun Stat Appl Methods 26(5):507–517

Kim JK, Berg E, Park T (2016) Statistical matching using fractional imputation. Surv Methodol 42(1):19–40

Little RJA (1992) Regression with missing X's: a review. J Am Stat Assoc 87(420):1227–1237

Little RJ, Rubin DB (2002) Statistical analysis with missing data, 2nd edn. Wiley, London. https://doi.org/10.1002/9781119013563

Moriarity C, Scheuren F (2001) Statistical matching: a paradigm for assessing the uncertainty in the procedure. J Offic Stat 17(3):407–422

Peytchev A, Peytcheva E (2017) Reduction of measurement error due to survey length: evaluation of the split questionnaire design approach. Surv Res Methods 11(4):361–368

Pigott TD (2001) A review of methods for missing data. Educ Res Eval 7(4):3535–3830

Raghunathan TE, Grizzle JE (1995) A split questionnaire survey design. J Am Stat Assoc 90(429):54–63

Rässler S (2002) Statistical matching: a frequentist theory, practical applications, and alternative bayesian approaches. Springer, New York. https://doi.org/10.1007/978-1-4613-0053-3

Rässler S (2004) Data fusion: identification problems, validity, and multiple imputation. Austrian J Stat 33:153–171

Rendall MS, Dastidar BG, Weden MM, Baker EH, Nazarov Z (2013) Multiple imputation for combined-survey estimation with incomplete regressors in one but not both surveys. Sociol Methods Res 42(4):483–530

Roszka W (2015) Some practical issues related to the integration of data from sample surveys. Statistika: Stat Econ J 95(1):60–75
Rubin DB (1986) Statistical matching using file concatenation with adjusted weights and multiple imputations. J Bus Econ Stat 4(1):87–94
Singh AC, Mantel H, Kinack M, Rowe G (1993) Statistical matching: use of auxiliary information as an alternative to the conditional independence assumption. Surv Methodol 19:59–79
Stuart M, Yu C (2019) A computationally efficient method for selecting a split questionnaire design. Creat Compon. https://lib.dr.iastate.edu/creativecomponents/252
Van Buuren S, Groothuis-Oudshoorn K (2011) mice: Multivariate imputation by chained equations in R. J Stat Softw 45(3):1–67
Vantaggi B (2008) Statistical matching of multiple sources: a look through coherence. Int J Approx Reason 49:701–711

.

Eidesstattliche Versicherung

(Siehe Promotionsordnung vom 12.07.11, § 8 Abs. 2 Pkt. 5)

Hiermit erkläre ich an Eidesstatt, dass die Dissertation von mir selbstständig, ohne unerlaubte Beihilfe angefertigt ist.

München, den 02.06.2021

(Mehboob Ali)