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 $F_1(x) = \frac{1}{n} \sum\limits_{i \subset s_1} 1\{x_i \leq x\}$. 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.

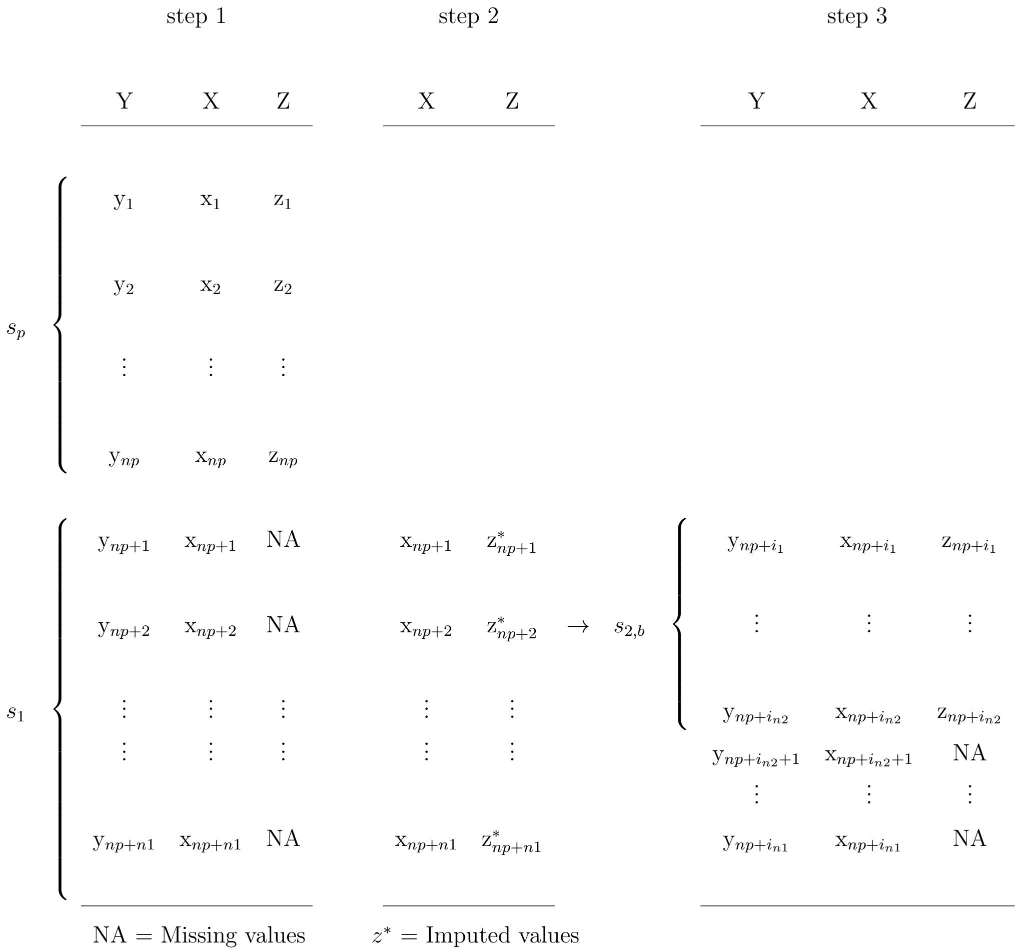
step 1
step 2
step 3
Y X Z
X Z
Y X Z
s_p
y_1 x_1 z_1
y_2 x_2 z_2
y_{np} x_{np} z_{np}
s_1
y_{np+1} x_{np+1} NA
y_{np+2} x_{np+2} NA
y_{np+n1} x_{np+n1} NA
x_{np+1} z^*_{np+1}
x_{np+2} z^*_{np+2}
x_{np+n1} z^*_{np+n1}
$\rightarrow$ $s_{2,b}$
y_{np+i_1} x_{np+i_1} z_{np+i_1}
$y_{np+i_{n2}}$ $x_{np+i_{n2}}$ $z_{np+i_{n2}}$
$y_{np+i_{n2}+1}$ $x_{np+i_{n2}+1}$ NA
$y_{np+i_{n1}}$ $x_{np+i_{n1}}$ NA
NA = Missing values
z^* = Imputed values

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

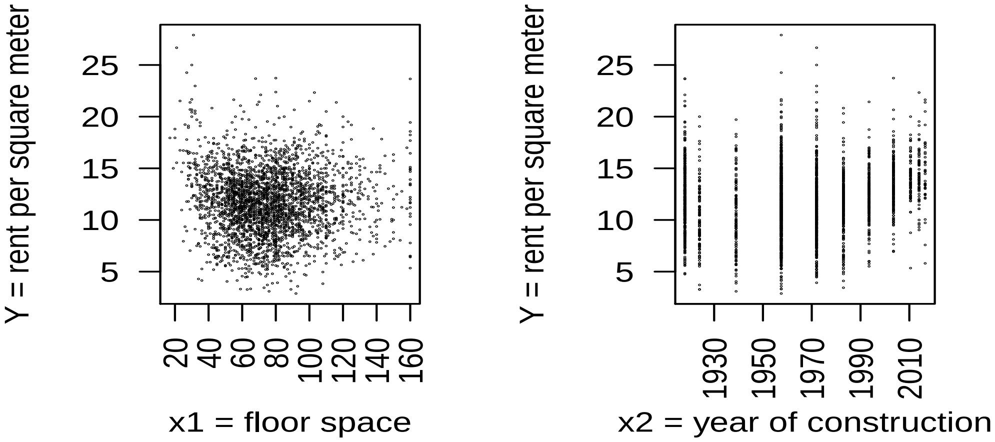

Fig. 2. Rent per square meter relation with floor space (left) and year of construction (right)

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|)
$1/\mathrm{x}_1$	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 $\mathrm{E}(var(\hat{\beta}))$, 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
$1/\mathrm{x}_1$	**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, Cambridge, 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