sciendo

Journal of Official Statistics, Vol. 39, No. 4, 2023, pp. 459–487, http://dx.doi.org/10.2478/JOS-2023-0022

Block Weighted Least Squares Estimation for Nonlinear Cost-based Split Questionnaire Design

Yang Li[1], *Le Qi*[1], *Yichen Qin*[2], *Cunjie Lin*[1], *and Yuhong Yang*[3]

In this study, we advocate a two-stage framework to deal with the issues encountered in surveys with long questionnaires. In Stage I, we propose a split questionnaire design (SQD) developed by minimizing a quadratic cost function while achieving reliability constraints on estimates of means, which effectively reduces the survey cost, alleviates the burden on the respondents, and potentially improves data quality. In Stage II, we develop a block weighted least squares (BWLS) estimator of linear regression coefficients that can be used with data obtained from the SQD obtained in Stage I. Numerical studies comparing existing methods strongly favor the proposed estimator in terms of prediction and estimation accuracy. Using the European Social Survey (ESS) data, we demonstrate that the proposed SQD can substantially reduce the survey cost and the number of questions answered by each respondent, and the proposed estimator is much more interpretable and efficient than present alternatives for the SQD data.

Key words: Block weighted least squares estimation; block-wise missing data; nonlinear cost function; split questionnaire design; large-scale survey.

1. Introduction

The European Social Survey (ESS) is an academically driven survey that has been conducted across Europe every two years since 2001. It measures moral, religious, social, economic, political attitudes and behavior patterns of various populations (Schnaudt et al. 2014). A large number of studies have been conducted based on ESS data. For example, Best and Wolf (2013) studied the attitudes toward homosexuality using a linear regression model; Davidov et al. (2018) performed a test for measurement invariance across all countries with ESS data; Vonneilich et al. (2019) discussed the effects of different aspects of social relations on educational inequalities using a linear mixed effects model. Although ESS data have generated the interest of many researchers, its collection takes a long time and requires great efforts. For example, in 2018, over 40,000 respondents were sampled from 31 countries in Europe (http://www.europeansocialsurvey.org/). Such a large-scale, hour-long, and face-to-face interview includes more than 500 questions, which requires much time and resources for respondents to complete. Furthermore, a long questionnaire often means a heavy burden for respondents, which usually results in low

[1] Renmin University of China, Center for Applied Statistics and School of Statistics, att: Cunjie Lin, 59 Zhongguancun St, Beijing 100872, China. Emails: yang.li@ruc.edu.cn, qle1999@163.com and lincunjie@ruc.edu.cn
[2] University of Cincinnati, Department of Operations, Business Analytics, and Information Systems Cincinnati, Ohio, U.S.A. Email: qinyn@ucmail.uc.edu
[3] University of Minnesota, School of Statistics Minneapolis, U.S.A. Email: yangx374@umn.edu
Acknowledgments: Li's work and Lin's work was supported by the MOE Project of Key Research Institute of Humanities and Social Sciences (22JJD910001).

© Statistics Sweden

response rates, high sampling errors, and measurement errors (Early 2016; Peytchev and Peytcheva 2017; Liu and Wronski 2018).

The National Assessment of Educational Progress (NAEP) in the USA is another large-scale nationally representative survey to measure student achievements, such as what American students know and can do in various subject areas. More than 150,000 students are included in the sample and each student answers many questions as part of the assessment (Neidorf and Sheehan 2014). The details of sampling in NAEP can be found in Rust and Johnson (1992). Because NAEP is a large-scale educational assessment, it also faces the aforementioned challenges and a balanced incomplete block design is used. Typically, the design splits the pool of items into a set of blocks and the split may depend on such practical issues as the wish to offer students blocks with motivating combinations of items or to match blocks across sub-questionnaires with respect to the time needed to complete them. Then the sub-questionnaires are assigned to students in the lowest unit (usually school classes) to minimize the cluster effects involved in sampling a hierarchically structured population.

Throughout these large-scale surveys, we see that one possible approach to manage survey cost and reduce respondent burden is to use a split questionnaire design (SQD), which divides a long questionnaire into short parts or modules, with each respondent filling only one of the sub-questionnaires that include one or more modules. Along the line of SQD, Raghunathan and Grizzle (1995) proposed a survey design via matrix sampling and the resulting data were analyzed using a multiple imputation method. Adigüzel and Wedel (2008) developed an optimal design by minimizing the Kullbak-Leibler distance between the distributions of the complete questionnaire data and the split questionnaire data, and illustrated the design's efficiency in synthetic and empirical web survey data. In cognitive development research, Rhemtulla and Little (2012) described two planned missing data designs, namely the multi-form design and the two-method measurement design, which were efficient in reducing participant fatigue and improving data quality. Additionally, Andreadis and Kartsounidou (2020) presented evidence that splitting a long questionnaire into short parts contributed positively to the response rates. With the increasing interest in SQD, the optimal design obtained by minimising a proper cost function for the survey, subject to specified reliability constraints, is an important aspect for more research. For example, Chipperfield and Steel (2009, 2011) and Ioannidis et al. (2016) introduced a linear cost function, which facilitates the search for the optimal design. However, the linear form of cost function may still be insufficient to accommodate a potentially complex relationship between the number of questions and the cost of the survey.

Besides questionnaire design, estimation of regression models with SQD data is another open problem. Compared with the analysis based on a complete data set, SQD data are more challenging to analyze due to the block-wise missing structure. Although likelihood-based methods (Little and Schluchter 1985; Skinner and Coker 1996; Chipperfield et al. 2018; Chipperfield and Steel 2012) are widely used in handling missing data, they often require strong assumptions on the distribution of missing data (Dziura et al. 2013). Poor performance was observed when the distribution assumptions were violated (Lesperance and Kalbfleisch 1992). For a linear regression model or other models of interest, another possible technique is to impute the missing values before estimation (Little 1992; Cai et al. 2010; Mazumder et al. 2010; Yuan and Bentler 2010; Yang and Kim 2017). However, it is

difficult to recover the missing blocks with (multiple) imputation due to extensive missing blocks in SQD data. To address the defects of these traditional techniques for missing data, several methods were proposed that can manage multi-source/multi-modality data or fragmentary data with high missing rates. Xiang et al. (2014) presented a unified "bi-level" learning model for complete multi-source data and extended it to incomplete data, which can be generalized to other applications with block-wise missing data sources. Fang et al. (2019) developed a model averaging method for fragmentary data with missing completely at random, which fit candidate models using a vailable covariate data and combined them according to some criterion. Yu et al. (2020) proposed a direct sparse regression procedure by using covariance from multi-modality data (DISCOM) to find the optimal sparse linear prediction of a continuous response variable. However, none of these methods are designed for the block-wise missing data from a SQD. The SQD data differs from the fragmentary data in its "orderly" missing structure in the sense that there are only a few missing patterns despite the high missing rate. Thus, in addition to a well-designed split questionnaire, a good estimator for regression models based on the SQD data is also needed beyond the existing methods.

In this study, we consider a two-stage framework with respect to the aforementioned issues encountered in a large-scale survey with long questionnaires. We first focus on the design of a split questionnaire to reduce the cost of the survey while ensuring precision requirements on the estimates of the sample means of the question items, which are usually the first priority for most surveys. We then consider estimation of the linear regression model based on data obtained using a SQD. The two stages are not isolated, which means that the SQD needs to cope with both the survey cost and the subsequent estimation. Specifically, in our approach, we need a small sample size of the complete cases data obtained with full questionnaire in Stage I for correcting the possible bias of the estimate based on the SQD data in Stage II. Moreover, instead of assuming that the cost increases linearly with the number of questions, a quadratic cost function is introduced, which may be more practical in various situations. In Stage II, we propose a block weighted least squares (BWLS) estimation approach, which makes full use of all available data and has an explicit expression that is easy to implement. In particular, the proposed estimator has shown satisfactory performance in improving both estimation and prediction accuracies. Overall, this work provides a practical and useful design as well as an estimation approach for large-scale surveys in a two-stage framework.

The article is organized as follows. In Section 2, we propose a framework consisting of a SQD and an estimation of linear regression model based on the SQD data. Several numerical simulations are conducted in Section 3. In Section 4, we illustrate the performance of the proposed framework based on the ESS data in 2018, followed by a discussion in Section 5.

2. Method

2.1. Stage I: Nonlinear Cost-Based SQD

Consider a complete questionnaire containing p questions/items. The data collected for each question corresponds to one variable, which is assumed to be either continuous or

dichotomous. Suppose the p questions can be divided into m modules, which are presented by disjoined sets $M_1, ..., M_m \subseteq [p] = \{1, 2, ..., p\}$. Each module M_i corresponds to a particular aspect of the research and the number of questions in module M_i is denoted as p_i and $p = \sum_{i=1}^{m} p_i$. We assign these modules to different sub-questionnaires $Q_j, j = 1, 2, \cdots, q$. In particular, Q_1 is a full questionnaire containing all modules. Here, we use a structure matrix $\mathbf{A} = (a_{ij})_{m \times q}$ to describe the assignment, where $a_{ij} = 1$ indicates the module M_i appears in sub-questionnaire Q_j, and $a_{ij} = 0$ otherwise. Note that each module can be assigned to multiple sub-questionnaires. Such a structure matrix $\mathbf{A}$ is desired for the purpose of improving data quality and reducing survey cost.

To quantify the survey cost, we introduce the total cost function C,

$$C = \sum_{j=1}^{q} C_j = \sum_{j=1}^{q} (C_j^f + n_j \alpha_j), \tag{1}$$

where C_j is the cost of the sub-questionnaire Q_j, which consists of the fixed cost C_j^f and the varying cost $n_j \alpha_j$. The number of respondents filling the sub-questionnaire Q_j is n_j and Q_j contains $L_j = \sum_{i=1}^{m} a_{ij} p_i$ questions. Usually, C_j^f is incurred regardless of sample size, and it includes, but is not limited to, the upfront cost of implementing the survey, such as training interviewers, which is fixed for a given sub-questionnaire Q_j. The unit cost α_j is expressed as a function of L_j. Here, we use a quadratic function to describe the relationship between α_j and L_j:

$$\alpha_j = B_2 L_j^2 + B_1 L_j + B_0, j = 0, 1, ..., q,$$

where B_0, B_1, and B_2 are the coefficients, and we assume that $B_2 > 0$. If $B_1 \geq 0$, the cost function is nondecreasing with respect to L_j; If $B_1 < 0$ and $-\frac{B_1}{2B_2} \geq 2$, α_j first decreases and then increases with L_j. Here we adopt the quadratic form instead of linear form because we would like to take the difficulty in recruiting respondents into consideration. As the number of questions L_j increases, it becomes more challenging to find qualified respondents, hence increases our unit cost due to the reduced number of respondents. It is possible that the unit cost decreases at a small number of questions. This is because it is relatively easy to recruit respondents with a questionnaire of a reasonable small length. We admit the true unit cost function can be complex and even mathematically intractable. The quadratic form is only an approximation to the reality. But such a modification allows us to offer a more realistic and flexible solution to the issue. In the quadratic cost function, B_0 is the "baseline cost" when $L_j = 0$, and $\max(-B_1/2B_2, 1)$ is the critical point, beyond which the unit cost will increase with the number of questions. Generally, these parameters can be determined according to practical considerations or operational data. Figure 1 demonstrates an example of α_j with $B_2 = 0.1, B_1 = -1, B_0 = 10$. Note that when $B_2 = 0$, the cost function corresponds to the linear cost function (Ioannidis et al. 2016).

As mentioned above, the total cost C depends on the design, namely the structure matrix $\mathbf{A}$ as well as the sample size n_j of sub-questionnaire Q_j. We use $C(\mathbf{A}, \boldsymbol{n}_\mathbf{A})$ to stress the relationship between C and the design, where $\boldsymbol{n}_\mathbf{A} = (n_1, ..., n_q)^\top$ is the vector of sample sizes. In this study, the goal of the SQD is to find an optimized matrix $\mathbf{A}$ and the corresponding sample sizes $\boldsymbol{n}_\mathbf{A}$ with the goal of minimizing the total cost $C(\mathbf{A}, \boldsymbol{n}_\mathbf{A})$. In addition, we impose certain constraints on the minimum sample size for each module to ensure the estimation precision for each item mean μ. Let $\hat{\mu}_l$ be the sample mean of the l-th item. Then $n_i^* = \max_{l=1,\cdots,p_i} (\frac{z_{1-\alpha}^2 \hat{\sigma}_l^2}{e_l^2})$ is the minimum

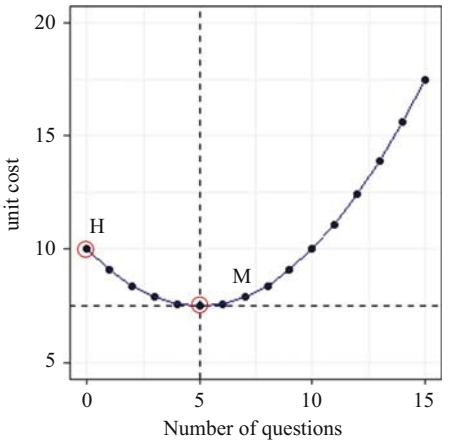


Fig. 1. The curve of cost function.

sample size for the i-th module, ensuring the estimation precision of $\hat{\mu}_l$ to achieve the absolute error of at most e_l. Here e_l is a prespecified value for precision of estimation, and $z_{1-\alpha}$ is the $1-\alpha$ quantile of the standard normal distribution, $\hat{\sigma}_l^2$ is sample variance of the l-th response item in the i-th module, estimated by the historical data. A general form of the design is given in Table 1, where the first m rows and q columns with 0 or 1 denote the structure matrix $\boldsymbol{A}$. Then the optimal design can be determined by solving the optimization problem:

$$\begin{aligned} &\min_{\mathbf{A}, n_{\mathbf{A}}} \; C(\mathbf{A}, \boldsymbol{n}_{\mathbf{A}}), \\ &s.t. \quad \mathbf{A}\boldsymbol{n}_{\mathbf{A}} > \boldsymbol{n}^*, \end{aligned} \tag{2}$$

where $\boldsymbol{n}^* = (n_1^*, ..., n_m^*)^\top$.

For the optimization problem (2), we use the simulated annealing algorithm (Ioannidis et al. 2016) to obtain the structure matrix $\mathbf{A}$ and the sample sizes $\boldsymbol{n}_{\mathbf{A}}$. Different from Ioannidis et al. (2016), we require a small sample size of the full questionnaire for correcting the bias of the estimate of the linear regression model based on the SQD data in Stage II. Therefore, we fix $a_{i1} = 1$ for all $i = 1, ..., m$, and require $n_1 > p$. After obtaining

Table 1. The general form of SQD.

	Q_1	Q_2	. . .	Qq	n_i^*
M_1	1	1	. . .	1	n_1^*
M_2	1	0	. . .	0	n_2^*
⋮	⋮	⋮	⋮	⋮	⋮
M_m	1	1	. . .	0	n_m^*
n_j	n_1	n_2	. . .	n_q	-
L_j	L_1	L_2	. . .	L_q	-

the optimal design of the questionnaire, we randomly assigned different sub-questionnaires to different interviewees. The collected data is of planned missing type since we are aware of which modules are included in each sub-questionnaire before data collection.

2.2. *Stage II: Block Weighted Least Squares Estimation*

Based on the nonlinear cost SQD in Stage I, we obtained the survey data with block-wise missing structure. As shown in Figure 2, the response variable $\boldsymbol{Y}$ is fully observed, while the design matrix $\boldsymbol{X}$ is block-wise missing as n_j interviewees only fill out L_j questions in the j-th sub-questionnaire Q_j for $j > 1$. Besides the item means, we are interested in the estimation of the linear regression model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim N(0, \sigma^2 \boldsymbol{I}_n).$$

To this end, we propose a block weighted least squares procedure (BWLS) to make full use of the block-wise missing data.

Specifically, we let $r_{il}, l = 1, 2, ..., p$ be the missing indicator such that $r_{il} = 1$ means the i-th respondent answers the l-th question, and $r_{il} = 0$ means missing. Let $\Omega_j, j = 1, ..., q$, be the set of indices of questions in the j-th sub-questionnaire Q_j and $\Omega_1 = [p]$. Denote the total sample size as $n = \sum_{j=1}^{q} n_j$. For $1 \le j \le q$, let $\mathcal{S}_j = \{i : r_{il} = 1 \text{ for } l \in \Omega_j, 1 \le i \le n\}$ denote the set of respondents who have answered at least the questions in Ω_j. Hence $\mathcal{S}_1$ is the set of complete cases. In addition, we define $\mathcal{S}_j^* = \{i : r_{il} = 1 \text{ for } l \in \Omega_j, r_{il} = 0 \text{ for } l \notin \Omega_j, 1 \le i \le n\}$ as the set of respondents who have answered exactly the questions in Ω_j, that is, the sub-questionnaire Q_j. Hence we have $\mathcal{S}_j^* \subseteq \mathcal{S}_j$ for $1 \le j \le q$. Let $|\mathcal{S}|$ denote the cardinality of a set $\mathcal{S}$, then $n_j = |\mathcal{S}_j^*| \le |\mathcal{S}_j|$ for $1 \le j \le q$. Furthermore, define $\boldsymbol{Y}_{(j)} \in R^{|\mathcal{S}_j|}, \boldsymbol{X}_{(j)} \in R^{|\mathcal{S}_j| \times |\Omega j|}$, where $R^{|\mathcal{S}_j|}$ denotes the $|\mathcal{S}_j|$-dimensional real space, and also define $\mathcal{D}_j = \{\boldsymbol{Y}_{(j)}, \boldsymbol{X}_{(j)}\}$ for $j = 1, ..., q$. Assume $\tilde{\boldsymbol{X}}_{(j)} = (1, \boldsymbol{X}_{(j)})$, where 1 is a column of 1 with size of n_j. Based on the complete cases $\mathcal{D}_1$, we can get the complete case (CC) least squares estimator

$$\hat{\boldsymbol{\beta}}_{cc} = \left(\tilde{\boldsymbol{X}}_{(1)}^{\top} \tilde{\boldsymbol{X}}_{(1)}\right)^{-1} \tilde{\boldsymbol{X}}_{(1)}^{\top} \boldsymbol{Y}_{(1)}.$$

As discussed earlier, n_1 is usually set to be not too large compared to n_j for $j = 2, ..., q$ due to budget and respondent load considerations. Thus, the efficiency of $\hat{\boldsymbol{\beta}}_{cc}$ is typically

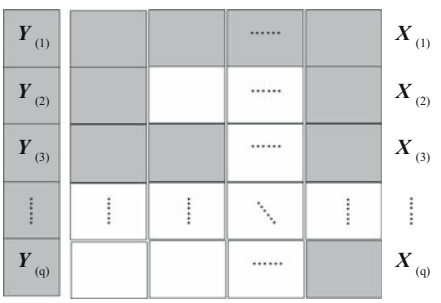


Fig. 2. An example for data structure.

unsatisfactory. Hence we propose an improved least squares estimator, with which we calculate the weights in the BWLS estimator.

Note that the covariance matrix of $\hat{\beta}_{cc}$ is $\sigma^2(\tilde{\boldsymbol{X}}_{(1)}^\top \tilde{\boldsymbol{X}}_{(1)})^{-1}$, and the sample covariance matrix $\boldsymbol{S}_{xx} = \tilde{\boldsymbol{X}}_{(1)}^\top \tilde{\boldsymbol{X}}_{(1)}/n_1$ is invertible but may be numerically unstable since n_1 is not large and p/n_1 is close to 1. Such a setting usually results in a dramatically increased estimation error, hence we use a well-conditioned covariance matrix to replace $\boldsymbol{S}_{xx}$ (Ledoit and Wolf 2004). Specifically, we consider an optimal linear shrinkage estimator for the covariance matrix, which has the form

$$\hat{\boldsymbol{\Sigma}}_{xx} = \hat{a}_1 \boldsymbol{I}_{p+1} + \hat{a}_2 \boldsymbol{S}_{xx},$$

where $\boldsymbol{I}_{p+1}$ is a $(p+1)\times(p+1)$ identity matrix, $\hat{a}_1$ and $\hat{a}_2$ are the coefficients determined by the optimization problem (4) of Ledoit and Wolf (2004). Specifically, $\hat{a}_1 = \frac{\alpha_n b_n^2}{d_n^2}$ and $\hat{a}_2 = \frac{a_n^2}{d_n^2}$, where $\alpha_n = tr(\boldsymbol{S}_{xx})/(p+1)$, $d_n^2 = \|\boldsymbol{S}_{xx} - \alpha_n \boldsymbol{I}_{p+1}\|_n^2$, $b_n^2 = \min\{d_n^2, \frac{1}{n_1^2}\sum_{i=1}^{n_1} \|\tilde{\boldsymbol{X}}_{(1),i} \tilde{\boldsymbol{X}}_{(1),i}^{\mathrm{T}} - \boldsymbol{S}_{xx}\|_n^2\}$, $a_n^2 = d_n^2 - b_n^2$, and $\tilde{\boldsymbol{X}}_{(1),i}$ is the i-th column of $\tilde{\boldsymbol{X}}_{(1)}$. Note that $\hat{\boldsymbol{\Sigma}}_{xx}$ is similar in spirit to ridge regression (Hoerl and Kennard 1970). Then the improved complete case (ICC) estimator is

$$\hat{\boldsymbol{\beta}}_{Icc} = \hat{\boldsymbol{\Sigma}}_{xx}^{-1} \boldsymbol{S}_{xy},$$

where $\boldsymbol{S}_{xy} = \tilde{\boldsymbol{X}}_{(1)}^\top \boldsymbol{Y}_{(1)}$. Although the well-conditioned covariance matrix gives $\hat{\boldsymbol{\beta}}_{Icc}$ better performance, it still ignores the data from other sub-questionnaires and needs further improvement.

To make full use of the data from all other sub-questionnaires, we use the weighted least squares estimation. Based on the CC data $\mathcal{D}_1$ from the full questionnaire Q_1, we can get an unbiased estimator $\hat{\boldsymbol{\beta}}_{cc}$. Meanwhile, the incomplete data $\mathcal{D}_j, j = 2, ..., q$, yield possibly biased estimators $\hat{\boldsymbol{\beta}}^{(j)} = (\tilde{\boldsymbol{X}}_{(j)}^\top \tilde{\boldsymbol{X}}_{(j)})^{-1} \tilde{\boldsymbol{X}}_{(j)}^\top \boldsymbol{Y}_{(j)}$. Then we define $\hat{\boldsymbol{\beta}}_{sub}^{(j)} = \Pi_j^\top (\tilde{\boldsymbol{X}}_{(j)}^\top \tilde{\boldsymbol{X}}_{(j)})^{-1} \tilde{\boldsymbol{X}}_{(j)}^\top \boldsymbol{Y}_{(j)}$ where Π_j is the $L_j \times p$ projection matrix such that $\Pi_j \boldsymbol{\beta} = \boldsymbol{\beta}_{(j)}$, and $\boldsymbol{\beta}_{(j)} \in R^{|\Omega_j|}$. Since $\hat{\boldsymbol{\beta}}_{sub}^{(j)}$ have different biases for different Q_js, we propose to treat samples from Q_js differently in estimating the linear regression model. Specifically, we assign different weights to $\mathcal{D}_j$s for balancing the information from various sub-questionnaires. The samples $\mathcal{S}_j$ should be assigned large weight if the sub-questionnaire Q_j provides important information in the sense that it yields a smaller residual sum of squares. With this in mind, the weight is set to be the reciprocal of the average of residual sum of squares. The j-th weight for samples $\mathcal{S}_j^*$ given by

$$w_j = \frac{|\mathcal{S}_j|}{\sum_{i\in\mathcal{S}_j} \left(Y_i - \tilde{\boldsymbol{X}}_{(j)} \Pi_j \hat{\boldsymbol{\beta}}_{Icc}\right)^2},$$

which means we compute the weights for D_j with more available cases since $\mathcal{S}_j^* \subseteq \mathcal{S}_j$. It is noticed that the denominator in w_j is the prediction error using the items in Q_j and the estimates of the associated regression coefficients are obtained from the respondents that answer the full range of questions in Q_1. Intuitively, $\frac{1}{|\mathcal{S}_j|}\sum_{i\in\mathcal{S}_j} (Y_i - \tilde{\boldsymbol{X}}_{(j)} \Pi_j \hat{\boldsymbol{\beta}}_{Icc})^2$ measures the missing information of the sub-questionnaire Q_j, and we should lower the weight of these samples if the corresponding sub-questionnaire misses too much information. Based

on the weights, we can get the final estimator by minimizing the weighted loss function

$$L(\boldsymbol{\beta}) = \sum_{j=1}^{q} w_j \sum_{i \in \mathcal{S}_j^*} \left(Y_i - \tilde{\boldsymbol{X}}_i^\top \boldsymbol{\beta} \right)^2,$$

where $\tilde{X}$ is the imputed matrix of $\boldsymbol{X}$ and replacing the missing values with the mean of each variable and $\tilde{X}_j$ is the i-th row of $\tilde{X}$. In fact, the final BWLS estimator has the explicit formulation

$$\hat{\boldsymbol{\beta}} = \left(\tilde{\boldsymbol{X}}^\top \boldsymbol{W} \tilde{\boldsymbol{X}} \right)^{-1} \tilde{\boldsymbol{X}}^\top \boldsymbol{W} \boldsymbol{Y},$$

where $\boldsymbol{W} = \text{diag}(w_1 \boldsymbol{I}_{n_1}, ..., w_q \boldsymbol{I}_{n_q})$ is the weight matrix. When $\boldsymbol{W}$ is an identity matrix, the estimator degenerates to the OLS estimator using the mean imputed values of the missing data items. The final estimator $\hat{\beta}$ uses information from all respondents, and assigns different weights to samples according to the information loss from different sub-questionnaires, which can reduce the bias induced by the block-wise missing data while improving the CC estimator by reducing its variance. Strictly speaking, we need the assumption of missing completely at random (MCAR), since the weight involves the improved CC estimator $\hat{\beta}_{Icc}$, which may not be effective when the assumption is violated. But the numerical studies show that the BWLS estimator also has a satisfactory performance under missing at random.

In summary, in Stage I, we create a survey design by minimizing a nonlinear cost function under the constraints of the variances of item means. Based on the SQD data obtained by implementing the design in Stage I, we develop a BWLS estimator for linear regression in Stage II. We would like to note that, if the item means are the target, the best linear unbiased estimator (BLUE) for a SQD (Chipperfield and Steel 2009), combining different estimates in an optimal way, is a good choice and a more practical estimator proposed by Merkouris (2015) can also be used when the number of variables or the sizes of the samples are large.

3. Simulation

In this section, we conduct multiple sets of simulations to evaluate the performance of the proposed two-stage framework. First, we conduct a simulation based on the ESS data set to highlight the advantages of our nonlinear cost-based design. Second, we compare the BWLS estimator with alternatives based on simulated block-wise missing data and the SQD data obtained in Subsection 3.1.

3.1. Simulation for Design

In this subsection, we demonstrate the merit of the proposed nonlinear cost-based design using the ESS data set, and more information about this data set will be given later in Section Here, the response of interest is how the respondent is satisfied with the way democracy works in their country, and it is measured using an 11-point Likert scale ranging from zero to ten. Among over 500 questions, we included the 87 questions with the largest correlations with the response into the following process of SQD. These questions are divided into four modules $M_1, ..., M_4$ according to their different focuses.

For the design of the questionnaire, we consider the complete questionnaire as well as the following SQD methods:

(a). "Proposed": The nonlinear cost-based design proposed in this article with $B_2 = 1$, $B_1 = -10$ and $B_0 = 40$ and the fixed design cost $C_j^f, j = 1, ..., q$ is set to be zero.

(b). "Linear": The integrated survey design proposed in Ioannidis et al. (2016), where the matrix $\boldsymbol{A}$ is obtained based on a linear cost function. Specifically, the unit cost α_j is defined as

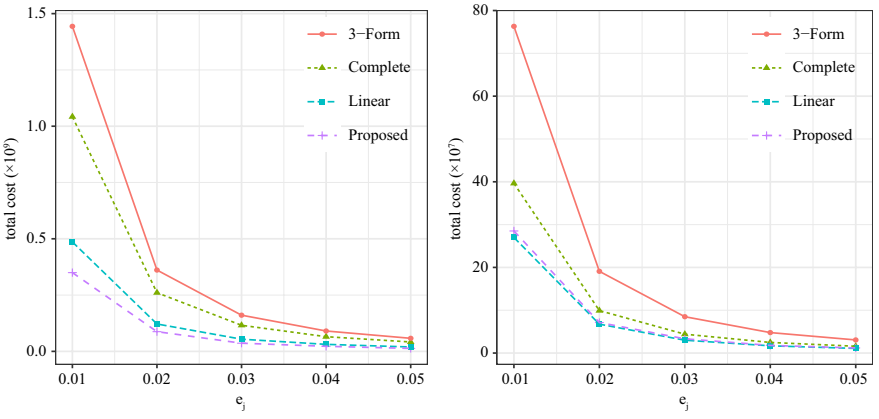
$$\alpha_j = B_1' L_j + B_0', \quad j = 0, 1, ..., q,$$

where B_1' represents for the unit cost per question and B_0' is the "baseline cost" when $L_j = 0$. The total cost $C(\mathbf{A}, \boldsymbol{n}_\mathbf{A})$ is the same as Equation (1). For fair comparison, we set $B_0 = B_0' = 40$ to ensure that the designs under linear cost function and nonlinear cost function share the same "baseline cost" when $L_j = 0$. For B_1', we first identify a possible range of it by considering that α_js under linear and non-linear cost function are equal for the number of questions $L_j = 87$. We find that the unit cost is the same as the proposed nonlinear design when $B_1' = 77$, that is $\alpha_j = 6739$ for $L_j = 87$ under both designs. Within the range of B_1', we search the optimal value from (0, 77] so that the total linear cost is minimum given the absolute error $e_l = 0.03$, and we obtained $B_1' = 29$.

(c). "3-Form": The three-form design studied in Rhemtulla and Little (2012).

Here, we consider two situations where the true unit cost function is assumed to be quadratic with $\alpha_j - L_j^2 - 10L_j + 40$ in Situation 1 and linear with $\alpha_j = 29L_j + 40$ in Situation 2, respectively. Considering that only 29 variables are included in the subsequent linear regression analysis, we set $n_1 = 39$ in the proposed design. For both "Linear" and "3-Form", the sample size of the full questionnaire Q_1 is set to be the same as the proposed design. For all designs, the first module M_1 collects information on demographic characteristics, and it appears in all sub-questionnaires. In the "3-Form" design, each sub-questionnaire is composed of the first module M_1 as well as two of the other three modules. We compare the total cost and the average number of questions answered per respondent with different absolute errors.

The results of total cost are shown in Figure 3. The total cost decreases with the increase of the allowed e_l for all designs in both situations. When the true unit cost is quadratic, the proposed design yields the lowest total cost among the four methods across all the different e_ls. The gap of the total cost between the "Linear" and the proposed design decreases with the increase of the allowed e_l. Given e_l, the total cost of "3-Form" design is larger than that of the complete questionnaire, primarily because the sample size of each sub-questionnaire must meet the most stringent requirement of modules contained in "3-Form" design, which leads to substantial extra cost. Specifically, the first module, M_1, appears in all sub-questionnaires and its required sample size is the largest among all modules for a given e_l, denoted as n_{max}. Ignoring the fixed cost C_j^f for simplicity, the total cost of the complete questionnaire is $n_{max}\alpha_1$, where α_1 is the unit cost for respondents answering the full-questionnaire with $L_1 = 87$ questions. However, in "3-Form" design, the sample size of each sub-questionnaire is also n_{max} since each sub-questionnaire includes module M_1. Then the three sub-questionnaires of sample size n_{max} and a full-questionnaire of sample size n_1 yield a total cost, $n_1\alpha_1 + n_{max}(\alpha_2 + \alpha_3 + \alpha_4)$. Here, we fix



(a) The true cost function is quadratic. (b) The true cost function is linear.

Fig. 3. The comparasion of total cost for Situation 1 (a) and Situation 2 (b).

$n_1 = 39$ for the full questionnaire to ensure the feasibility of BWLS estimator. Apparently, we have $L_2 + L_3 + L_4 > L_1$, and this is why the total cost of "3-Form" design is larger than that of the complete questionnaire.

To quantify the respondent burden, the average number of questions answered per respondent defined as $\bar{L} = \sum_{j=1}^{q} L_j n_j / \sum_{j=1}^{q} n_j$ are calculated for different methods with $e_l = 0.03$, which are "Proposed": 37.97, "Linear": 58.80, "3-Form": 62.58, and "Complete": 87, respectively. It is clear that the proposed method yields the minimum number of questions answered per respondent. We also calculate $\bar{L}$ when $e_l = 0.01, 0.02, 0.04$, and 0.05. The average number of questions answered per respondent for the complete design is always 87, which has been reduced by 53% with the proposed method on average. Moreover, the "3- Form" and "Linear" designs reduce the average number of questions by 28% and 32% on average, respectively. Overall, the proposed non-linear cost-based design outperforms the alternatives in terms of both total cost and the average number of questions answered per respondent.

For a demonstration, setting the absolute error to be 0.05, the resulting design with the proposed method is shown in Table 2. The column n^{**} is the actual sample size of module M_i based on the proposed design.

Table 2. The structure matrix.

	Q_1	Q_2	Q_3	Q_4	n_i^*	n_i^{**}
Module 1	1	1	1	1	6186	6186
Module 2	1	0	1	0	4095	4095
Module 3	1	1	0	1	2063	2130
Module 4	1	1	1	0	4329	4329
n_j	39	234	4056	1857	-	-
L_j	87	46	79	16	-	-

3.2. Simulation for Estimators

In this subsection, we compare the BWLS estimator with alternatives, including the CC estimator $\hat{\beta}_{CC}$, the improved CC estimator (I-CC) $\hat{\beta}_{ICC}$, a model average estimator (Fang et al. 2019, MA-CV), and an imputation-based estimator termed SI-LS here, which replaces missing values in a large-scale matrix using the SOFT-IMPUTE algorithm proposed in Mazumder et al. (2010) and performs a least squares estimation for $\boldsymbol{\beta}$ based on the single imputed data set.

Simulation 1. We generate data from the linear regression model

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon}$ is generated from $N(0, \sigma^2)$. The covariates $\boldsymbol{X}$ are generated from $N(0, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is the corresponding $p \times p$ covariance matrix, with the structure as

$$\Sigma = \begin{bmatrix} \Sigma_{in} & \cdots & \Sigma_{off} \\ \vdots & \ddots & \vdots \\ \Sigma_{off} & \cdots & \Sigma_{in} \end{bmatrix}.$$

The non-diagonal and diagonal elements in $\boldsymbol{\Sigma}_{\text{in}}$ are λ_1 and one, respectively. Additionally, $\boldsymbol{\Sigma}_{off}$ is a constant matrix with element λ_2. Here, λ_1 is the correlation between questions within the same module, while λ_2 is the correlation between questions from different modules. We consider two cases of correlation structures: $\lambda_1 = 0.3$, $\lambda_2 = 0.1$ in Case 1, and $\lambda_1 = 0.5$, $\lambda_2 = 0.2$ in Case 2. The variance σ^2 takes different values such that $R^2 = Var(\boldsymbol{X}^\top \boldsymbol{\beta})/\{Var(\boldsymbol{X}^\top \boldsymbol{\beta}) + \sigma^2\} = 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1$. We consider two settings for co-efficients with different sparsities. Let $\boldsymbol{\beta} = (\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}, ..., \boldsymbol{\beta}^{(m)})$, where $\boldsymbol{\beta}^{(i)}, i = 1, ..., m$ is the coefficients of the variables in i-th module. In Setting I, $\boldsymbol{\beta}^{(1)} = (0.5, -0.5, \ 0.5, 1)$, $\boldsymbol{\beta}^{(2)} = (0.5, -0.5, -0.3, 0.5)$, $\boldsymbol{\beta}^{(3)} = (0, 0, 0, 0)$, $\boldsymbol{\beta}^{(4)} = (0, 0, 0, 0)$, $\boldsymbol{\beta}^{(5)} = (-0.5, 0.5, 1, 0.3)$, $\boldsymbol{\beta}^{(6)} = (1, 0.3, 0.3, 1)$, $\boldsymbol{\beta}^{(7)} = (-1, 1, 0.3, 0.3)$, $\boldsymbol{\beta}^{(8)} = (1, 0.5, 0.5, 0.5)$, $\boldsymbol{\beta}^{(9)} = (-0.5, 0.5, 0.3, -0.3)$ and $\boldsymbol{\beta}^{(10)} = (0.3, -0.3, 1, 0.3)$, thus M_3 and M_4 are inactive modules. In Setting II, we consider sparsity within modules, $\boldsymbol{\beta}^{(1)} = (-1, -1, 0, \ 1,)$, $\boldsymbol{\beta}^{(2)} = (0, 1, -1, 0.3)$, $\boldsymbol{\beta}^{(3)} = (0, 1, -0.5, -1)$, $\boldsymbol{\beta}^{(4)} = (0.3, 0.3, 0.3, 0)$, $\boldsymbol{\beta}^{(5)} = (0.3, \ 0.5, 1, 0)$, $\boldsymbol{\beta}^{(6)} = (-1, 0, 0, 1)$, $\boldsymbol{\beta}^{(7)} = (0.3, 0, 0, 0.3)$, $\boldsymbol{\beta}^{(8)} = (0, 0, 0.5, -0.3)$, $\boldsymbol{\beta}^{(9)} = (0.3, 0.5, 1, -1)$, and $\boldsymbol{\beta}^{(10)} = (-0.3, 1, 0.5, -0.3)$. In both settings for coefficients, we set $\beta_0 = 1$. We further consider multiple scenarios with different dimensions of covariates, number of modules, and sample sizes:

S1. $p = 20, m = 5, n_1 = 30, n_2 = ... = n_5 = 200$, with Setting I for $\boldsymbol{\beta} = (\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}, ..., \boldsymbol{\beta}^{(5)})$.

S2. $p = 40, m = 5, n_1 = 45, n_2 = ... = n_{10} = 200$, with Setting I for $\boldsymbol{\beta}$ (here, we merge $\boldsymbol{\beta}^{(i)}, i = 1, 2, ..., 10$ in to 5 modules pairwise).

S3. $p = 40, m = 10, n_1 = 45, n_2 = ... = n_{10} = 200$, with Setting II for $\boldsymbol{\beta} = (\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}, ..., \boldsymbol{\beta}^{(10)})$.

S4. $p = 40, m = 10, n_1 = 45, n_2 = ... = n_{10} = 200$, with Setting I for $\boldsymbol{\beta} = (\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}, ..., \boldsymbol{\beta}^{(10)})$.

S5. $p = 40, m = 10, n_1 = 45, n_2 = ... = n_{10} = 500$, with Setting I for $\boldsymbol{\beta} = (\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}, ..., \boldsymbol{\beta}^{(10)})$.

For each scenario, there are 2^m sub-questionnaires, of which we randomly select m sub-questionnaires. For all the settings, $B = 500$ replicates are simulated. The performances of different estimation methods are evaluated in terms of prediction error (PE) and mean square error (MSE), defined as

$$\text{PE} = \frac{1}{Bn_{test}} \sum_{b=1}^{B} \sum_{i=1}^{n_{test}} \left(\hat{Y}_i^{(b)} - E(Y_i) \right)^2, \quad \text{MSE} = \frac{1}{B(p+1)} \sum_{b=1}^{B} \sum_{l=0}^{p} \left(\hat{\beta}_i^{(b)} - \beta_l \right)^2,$$

respectively. Here $\hat{\beta}_i^{(b)}$ is the estimate of β_l, $\hat{Y}_i^{(b)}$ is the predicted value for Y_i in the b-th replicate and $E(Y_i) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p$. n_{test} is the sample size of the testing data set, which is set to be 200 in this simulation. Since the MSE can be decomposed into MSE = Variance + Bias^2, we have

$$\frac{1}{B(p+1)} \sum_{b=1}^{B} \sum_{l=0}^{p} (\hat{\beta}_l^{(b)} - \beta_l)^2 = \frac{1}{p+1} \sum_{l=0}^{p} \frac{1}{B} \sum_{b=1}^{B} (\hat{\beta}_l^{(b)} - \hat{\overline{\beta}}_l)^2 + \frac{1}{p+1} \sum_{l=0}^{p} (\hat{\overline{\beta}}_l - \beta_l)^2,$$

where $\hat{\overline{\beta}}_l = \frac{1}{B} \sum_{b=1}^{B} \hat{\beta}_l^{(b)}$. For Case 1, the results are shown in Figures 4, 5 and 6. The results of Case 2 are shown in Figures 8, 9 and 10 in Appendix (Section 6).

Figures 4 and 5 show the comparison of the MSE and PE among different methods. We purposely truncated the figure for ease of comparison, so results with large values do not appear in certain figures. We can see the estimation and prediction accuracies gradually improve with the increase of the signal-to-noise ratio R^2. In most of the considered scenarios, the MSE of the BWLS estimator is the smallest among the five methods except for large R^2. Additionally, a larger sample size of sub-questionnaire n_i leads to a lower MSE of the proposed estimator, as is expected. The result of PE also shows the superiority of BWLS. The MSE and PE of all five estimators are increasing as the number of variables increases.

From Figure 6, we find that the variance of the BWLS estimator is significantly smaller than that of the alternatives. This is mainly due to the full utilization of the samples at hand. Moreover, when the correlation between sub-questionnaires is much smaller, the bias of the BWLS estimator is also small. Overall, I-CC outperforms the CC estimator, but is inferior to MA-CV. In addition, the performance of SI-LS is barely satisfactory with regard to MSE and PE, especially when compared with BWLS.

The results of Case 2 are displayed in the Appendix (Section 6). Figures 8, 9 and 10 demonstrate the comparison of MSE and PE among the different estimators and the decomposition of MSE. In Case 2, we can draw similar conclusions as Case 1. However, not surprisingly, BWLS performs worse in MSE and PE as λ_2 increases and the bias of BWLS is getting larger with the increasing correlation between modules. It suggests that the correlation between modules should be lower than that within modules when dividing the complete questionnaire into shorter parts.

Additionally, results about the comparison between BWLS, mean imputation, and multiple imputation (MI) are shown in Appendix (Section 6), where mean imputation replaces the missing values with the item means and multiple imputation is conducted with

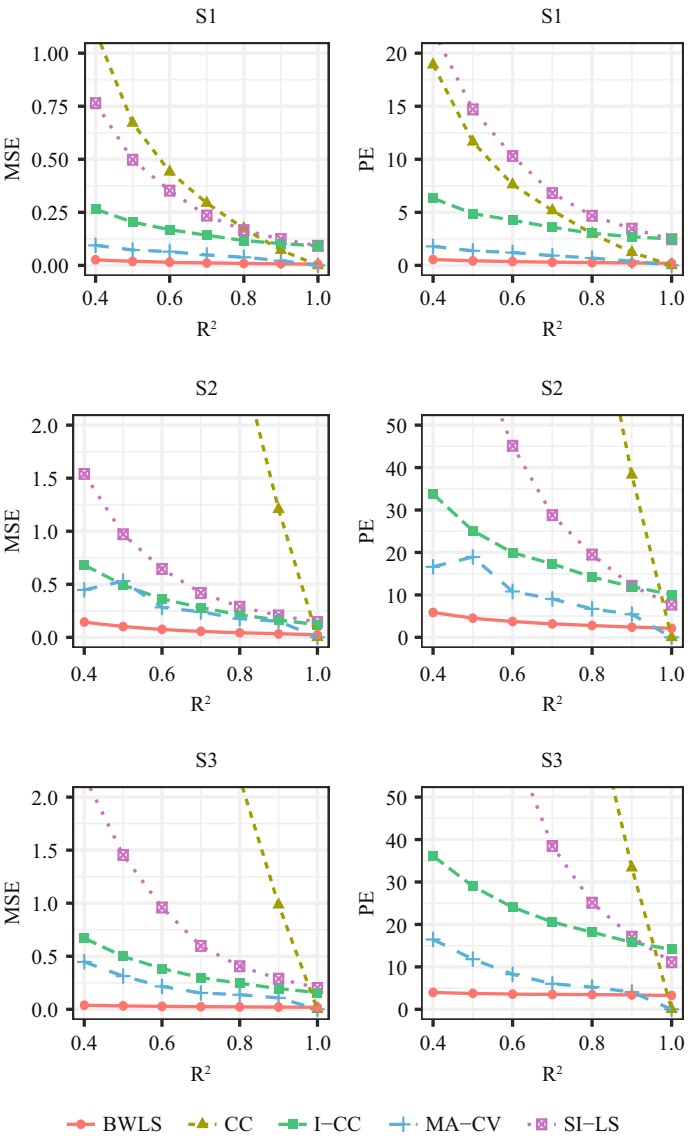


Fig. 4. The comparison of MSE and PE for different estimators in S1–S3 under case 1.

the MICE package in R. From the results, we can see that BWLS outperforms the mean imputation in most settings, especially in terms of PE. The superiority of BWLS in MSE is not that obvious, and this is mainly because the estimated weights in BWLS increase the variance of the estimator while the weights in mean imputation are constantly equal to 1. Besides, BWLS performs better than MI in terms of MSE for almost all scenarios we considered here. And BWLS also shows its superiority in terms of PE in most scenarios.

Simulation 2. Based on the design given in Table 2 with ESS data, we conduct a real data-based simulation to compare the proposed estimator BWLS with the alternatives as in

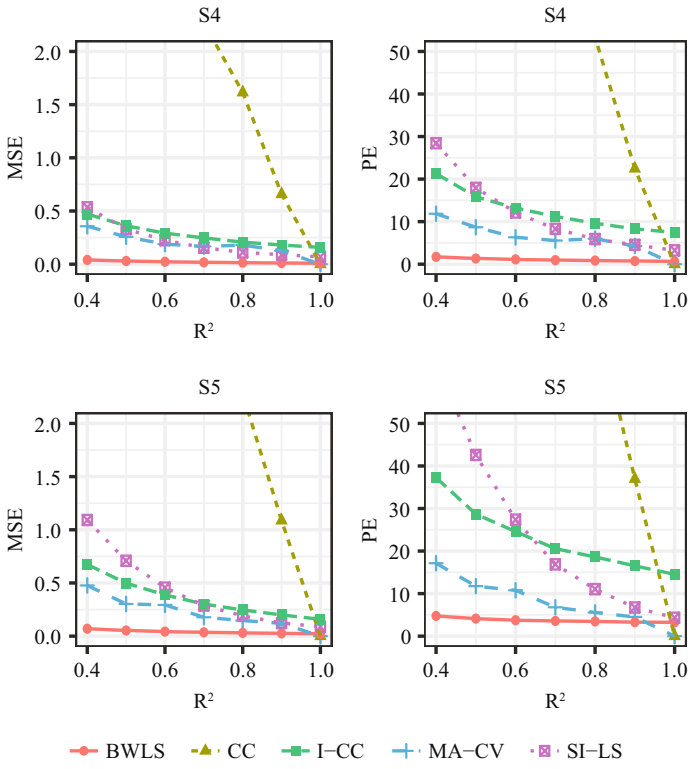


Fig. 5. The comparison of MSE and PE for different estimators in S4–S5 under case 1.

Simulation 1. We regress the response variable on each variable using the full data and select the top 29 variables with the smallest p-value for subsequent analysis. The selected variables are listed in Table 3. Based on the observations of these 29 variables, we generate Y from the linear regression model

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_{29} X_{29} + \epsilon,$$

with $\epsilon \sim N(0, \sigma^2)$. The variance of ϵ, σ^2 takes different values such that R^2 ranging from 0.4 to 1, and $\boldsymbol{\beta} = (1, -1, -1, 0, 1, 0, 1, -1, 0.3, 0, 1, -0.5, -1, 0.3, 0.3, 0.3, 0, 0.3, 0.5, 1, 0, -1, 0, 0, 1, 0.3, 0, -1, 0.3, 0.5, 0)^\top$. The estimation results based on 500 replications for BWLS, CC, I-CC, MA-CV, and SI-LS are presented in Figure 11 of Appendix (Section 6). We can draw similar conclusions as in Simulation 1 that BWLS has the best prediction and estimation performance.

4. ESS Data Analysis

The European Social Survey (ESS) was established at the National Centre for Social Research in London in 2001. Since 2003, the ESS Headquarters have been located at City, University of London in the UK. The survey measures the attitudes, beliefs, and behavior patterns of various populations, covering more than 500 questions and reaching over 40,000 respondents from more than 30 countries all over Europe. It's desirable to improve its design since such a large-scale survey requires significant resources (e.g., time and

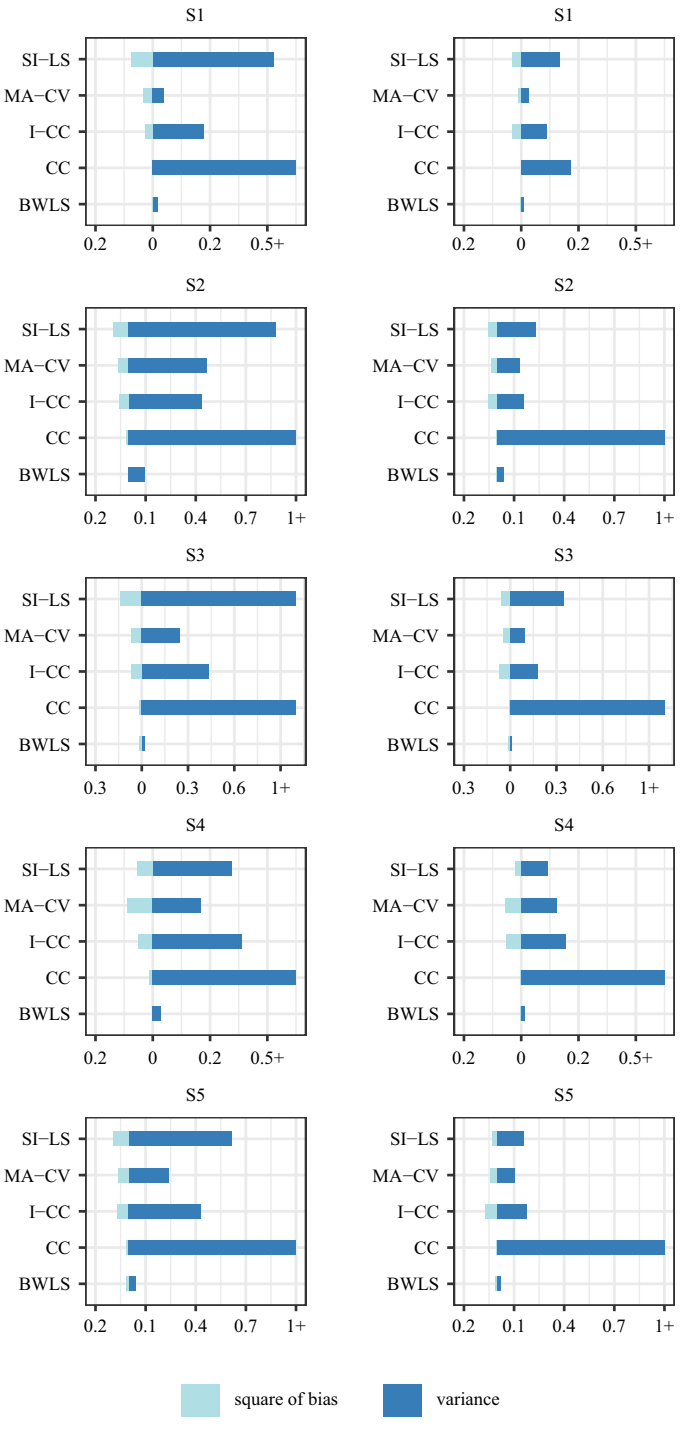


Fig. 6. The decomposition of MSE in case 1.

Table 3. The list of variables.

Module	Variable	Definition
	stfdem	How satisfied with the way democracy works in their country
	agea	Age of respondent, calculated
M_1	gndr	Gender
	eduyrs	Years of full-time education completed
	trstprl	Trust in country's parliament
	trstlgl	Trust in the legal system
	trstplc	Trust in the police
	trstprt	Trust in political parties
	trstun	Trust in the United Nations
	trstplt	Trust in politicians
	trstep	Trust in the European Parliament
M_2	stfeco	How satisfied with present state of the economy in their country
	stfedu	State of education in their country nowadays
	stfgov	How satisfied with the country's national government
	psppsgva	Political system allows people to have a say in what government does
	psppipla	Political system allows people to have an influence on politics
	frprtpl	Political system in their country ensures everyone a fair chance to participate in politics
	gvintcz	Government in their country takes into account the interests of all citizens
	poltran	Decisions in their country politics are transparent
	anvcld	Approves if a person chooses never to have children
	alvgptn	Approves if a person lives with partner they are not married to
M_3	acldnmr	Approves if a person has a child with a partner they are not married to
	aftjbyc	Approves if a person has full-time job while having children aged under three
	advcyc	Approves if a person gets divorced while having children aged under 12
	hinctnta	Household's total net income, all sources
	ifrjob	Compared with other people in their country, there is a fair chance of getting job I seek
	happy	How happy are you
M_4	recskill	Influence decision to recruit in their country: person's knowledge and skills
	ifredu	Compared to other people in their country, there is a fair chance of achieving the level of education I seek
	hincfel	Feeling about their household's income nowadays

money). Based on the proposed framework, we have finished the SQD in Subsection 3.1 and we use the SQD data obtained in Table 2 for the following regression analysis.

The ESS study concentrates on how the respondent is satisfied with the way democracy works in their country, and we choose *stfdem* as the response variable, which is measured

using an 11-point Likert scale ranging from zero to ten and treated as a continuous variable like Wu and Leung (2017). We apply the linear regression model to study the relationship between the response variable and the relevant factors in this section. As pointed out in Van Ham et al. (2017), economic outcomes and quality of governance should go a long way in explaining one's satisfaction with democracy. There are many other factors affecting popular satisfaction with the way democracy works according to some studies in political science, including the individual-level perceptions and aggregate measures based on expert and stakeholder surveys of corruption, fair and honest treatment by political officials, or impartiality. Based on the considerations above, we incorporate 29 covariates from four modules into the regression model (Table 3). Specifically, module M_1 is composed of three variables concerned with respondent background, that is age, gender, and years of education. Modules M_2, M_3, and M_4 consist of questions related to the attitude toward politics and society, marital relations and children, and the personal feelings about life, respectively. The correlation of these variables is shown in Figure 7. It illustrates the correlations between items within/between different modules. Specifically, the variables *trstprl*, *trstlgl*, ..., *poltran* are grouped into module M_2, which show obvious positive correlation. The variables *anvcld*, *alvgptn*, ..., *advcyc* are about the respondents' attitudes towards marriage and childbearing and make up the module M_3. Module M_4 are consist of variables *hinctnta*, *ifrjob*, ..., *hincfel*, which also show a positive correlation in Figure 7.

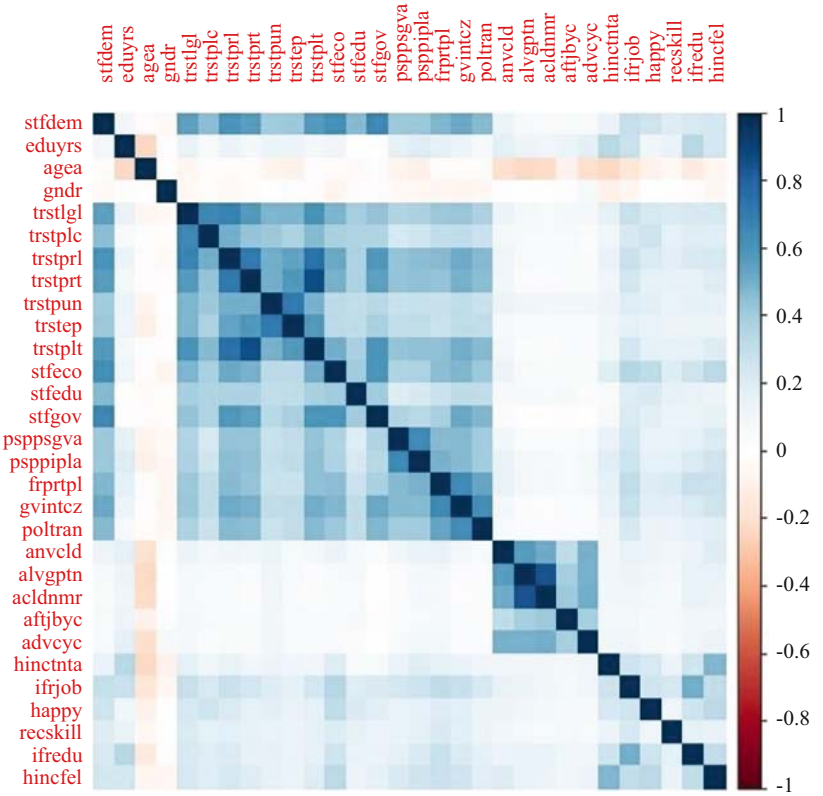


Fig. 7. The correlations between different variables.

With the proposed design, the assignment of modules in different sub-questionnaires can be obtained by minimizing the cost function as in simulation for design. The sample sizes of the four sub-questionnaires are shown in Table 2, and the total sample size is $n = 6{,}186$. After deleting the respondents with missing values, we finally obtain a complete data set containing 27,341 valid respondents, with 6,186 of them being used as the training data and the remaining 21,155 respondents being used for testing. For the training data set, the respondents are randomly assigned to one of the four sub-questionnaires. Based on the data with SQD in Table 2, we estimate the regression coefficients with BWLS, CC, I-CC, MA-CV, and SI-LS. The results based on 100 random assignments are shown in Table 4.

Table 4 shows the estimates of coefficients with training data as well as the prediction error (PE) of the testing data, with the standard deviation of the 100 replications contained in parentheses. In general, the PE of BWLS is the smallest among the five methods tested. The smaller standard error also indicates the superiority of BWLS. For interpreting the coefficients, we note that the response variable is likely to be driven by political systems' outputs and outcomes, such as economic performance (Van Ham et al. 2017). When the

Table 4. The estimates of coefficients and PE with 100 replications, where the averaged standard error is in parentheses.

Variable	CC	SI-LS	MA-CV	I-CC	BWLS
agea	0.003(0.142)	− 0.170(0.041)	0.016(0.029)	− 0.001(0.079)	0.011(0.007)
gndr	0.005(0.033)	− 0.007(0.002)	0.003(0.012)	0.002(0.019)	0.002(0.001)
eduyrs	0.031(1.156)	− 0.042(0.045)	− 0.073(0.248)	0.000(0.093)	− 0.082(0.048)
trstprl	0.127(0.362)	0.093(0.018)	0.063(0.089)	0.107(0.125)	0.092(0.018)
trstlgl	0.058(0.314)	0.013(0.018)	0.024(0.092)	0.081(0.123)	0.038(0.017)
trstplc	0.096(0.396)	0.065(0.020)	0.036(0.064)	0.070(0.143)	0.065(0.020)
trstprt	0.104(0.525)	− 0.003(0.024)	0.015(0.142)	0.065(0.133)	0.028(0.023)
trstun	− 0.007(0.412)	0.027(0.015)	0.017(0.093)	0.025(0.141)	0.034(0.015)
trstplt	0.034(0.389)	0.010(0.015)	0.002(0.086)	0.024(0.123)	0.014(0.016)
trstep	− 0.132(0.567)	− 0.015(0.022)	− 0.008(0.148)	0.015(0.128)	− 0.005(0.024)
stfeco	0.170(0.386)	0.170(0.016)	0.111(0.087)	0.161(0.141)	0.172(0.016)
stfedu	0.132(0.269)	0.089(0.016)	0.074(0.086)	0.112(0.108)	0.119(0.014)
stfgov	0.293(0.366)	0.257(0.019)	0.186(0.099)	0.251(0.124)	0.307(0.018)
psppsgva	0.163(0.819)	0.111(0.043)	0.086(0.222)	0.057(0.106)	0.098(0.040)
psppipla	− 0.060(0.923)	0.128(0.044)	− 0.007(0.284)	0.056(0.107)	0.073(0.039)
frprtpl	0.016(0.755)	0.175(0.039)	0.075(0.229)	0.081(0.118)	0.131(0.040)
gvintcz	0.181(0.812)	0.116(0.047)	0.081(0.151)	0.085(0.095)	0.108(0.049)
poltran	0.082(0.863)	0.024(0.037)	0.028(0.248)	0.066(0.107)	0.057(0.037)
anvcld	0.079(0.613)	0.074(0.051)	0.111(0.205)	0.024(0.132)	0.236(0.067)
alvgptn	0.182(1.277)	0.026(0.075)	0.055(0.427)	0.005(0.109)	0.095(0.103)
acldnmr	− 0.232(1.144)	− 0.105(0.083)	− 0.088(0.346)	− 0.001(0.096)	− 0.147(0.109)
aftjbyc	0.075(0.559)	− 0.023(0.045)	0.055(0.149)	0.007(0.126)	0.062(0.064)
advcyc	0.028(0.713)	− 0.056(0.054)	0.007(0.216)	0.008(0.120)	0.033(0.069)
hinctnta	− 0.004(0.201)	0.01(0.015)	− 0.007(0.044)	0.004(0.115)	− 0.008(0.012)
ifrjob	0.007(0.246)	0.004(0.013)	0.025(0.057)	0.015(0.122)	0.015(0.012)
happy	0.011(0.358)	0.039(0.019)	0.028(0.116)	0.037(0.123)	0.053(0.018)
recskill	0.054(0.503)	0.086(0.022)	0.063(0.145)	0.039(0.115)	0.069(0.021)
ifredu	0.032(0.290)	0.031(0.018)	0.014(0.086)	0.028(0.117)	0.011(0.012)
hincfel	0.053(0.754)	0.079(0.041)	0.064(0.175)	0.031(0.113)	0.042(0.037)
PE	12.398(6.704)	3.212(0.083)	4.150(2.102)	3.862(0.436)	2.783(0.039)

economy is doing well or at least people perceive it to be doing so, the response variable tends to increase (Van Ham et al. 2017), as the coefficients of variable *stfeco* are all positive under several estimation methods. In addition, the response variable also rises when the state apparatus, political institutions, and public officials are perceived to be transparent, impartial, and fair. Although we may expect the sign of most variables in module M_2 to be positive (Beichelt et al. 2014), we observe that CC, SI-LS, and MA-CV contain more estimated coefficients that are negative compared with I-CC and BWLS, which means that the estimators of BWLS are much more interpretable and in line with the economic significance. BWLS is likewise a more efficient estimator since it has the lowest standard deviation among the five methods.

5. Discussion

In this study, we propose a two-stage framework for a large-scale survey in order to improve its design and regression estimation. In Stage I, we create a survey design with a block-wise missing structure by minimizing a nonlinear cost function with constraint on the reliability of estimates of means. Reducing the cost and increasing the data quality can be achieved through the proposed design, as shown in our numerical studies. In Stage II, we put forward a BWLS estimator based on the SQD data obtained in Stage I under a linear regression model. Numerical studies have shown that the proposed design has better performance in terms of survey cost and respondent burden. Furthermore, the proposed estimator of BWLS leads to satisfactory estimation and prediction accuracies. In our ESS data analysis, the findings are consistent with other studies in political science.

Inspired by the two-stage framework proposed in this article, many studies can be performed in the future. First, various forms of function may be investigated to describe the total cost. Other factors such as response rates and information loss, can also be taken into account as criteria for a SQD. Also, there is an important constraint $n_1 > p$ in our framework. When it comes to the choice of n_1, we just give the guidance that n_1 is usually set to be not too large compared to n_j for $j = 2, ..., q$ due to budget and respondent load considerations.

Further research that examines how the proposed method works as $n_1 - p$ is varied is a valuable direction. Second, the number of questions included in the regression model may be increased, and data exchanges between countries or servers become much more convenient with the fast development of data science. Thus, it would be worthwhile for researchers to further explore variable selection and the corresponding distributed-computation methods in similar situations. Third, this article focuses on the situation of a single response and that its values in the data are observed. When several regression analyses are to be conducted with different response variables, it would be helpful to consider this in the survey design stage so as to ensure the collection of the information for all respondents as much as possible. If, unfortunately, some values are missing in the SQD data for the response variable being considered, we may remove those cases with missing observations or consider a suitable imputation method if available. Examination of effects of the relevant factors (e.g., nature of the missing mechanism) in the process and deriving more efficient techniques require significant future work. Fourth, the BWLS is only based on continuous or dichotomous independent variables at present, while in many other cases, multi-categorical variables are of interest as well. For example, when the impact of different levels of

educations and occupations on earnings is concerned, levels of education and occupation are usually treated as multi-categorical variables. The extension of our proposed framework to these new situations is also worthwhile to look into. This article is focused on a simple random sampling for our methodology and we only consider the influence of the design on the sample size requirements of the survey items. Extensions to complex samplings possibly with key regression parameters (in addition to the means of the survey items) to be considered in the survey design are interesting future research projects.

6. Appendix

6.1. Results for Case 2

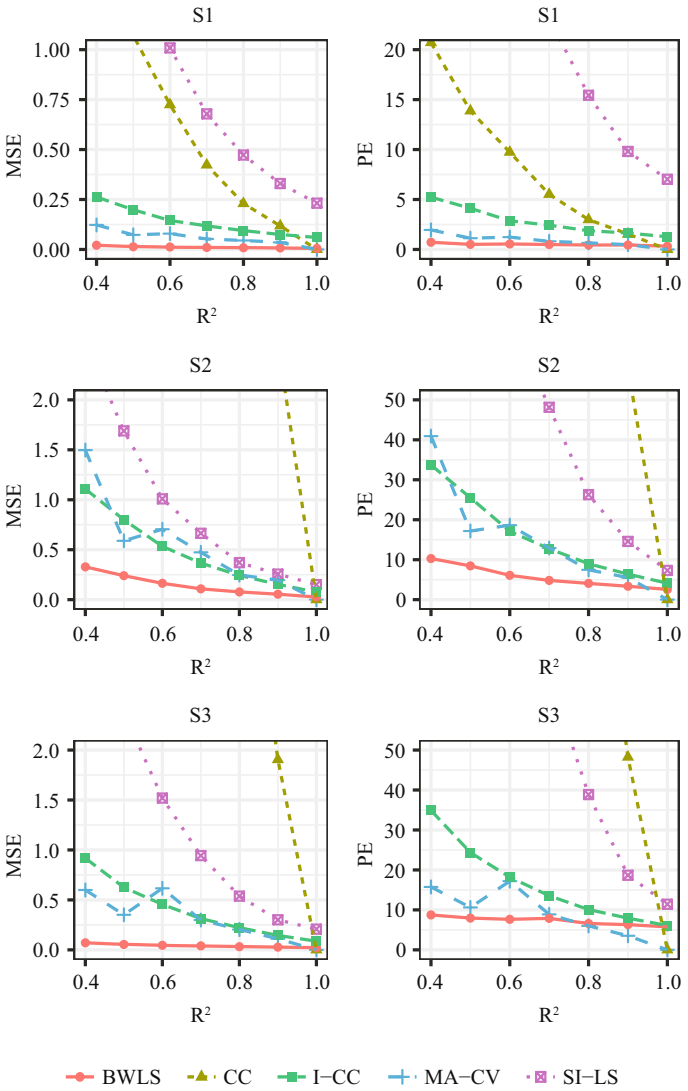


Fig. 8. The comparison of MSE and PE for different estimators in S1–S3 under case 2.

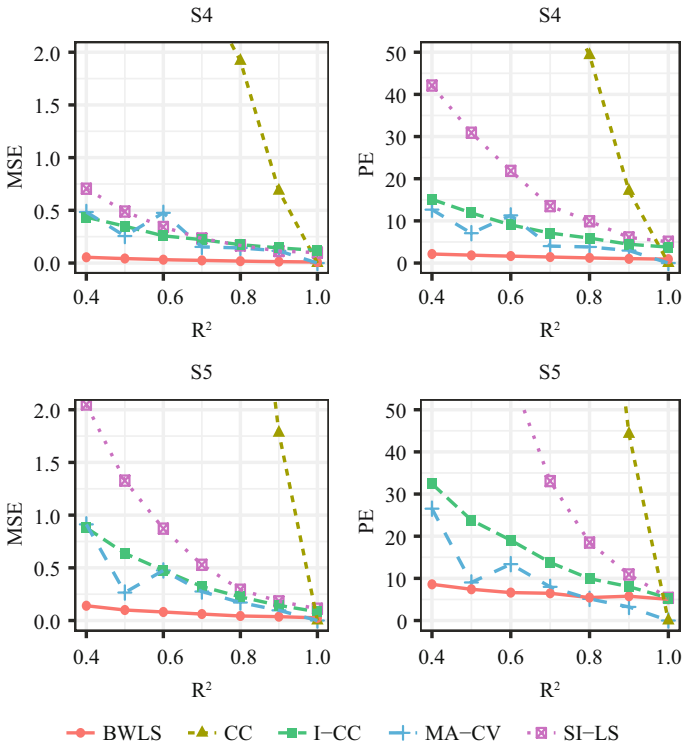


Fig. 9. The comparison of MSE and PE for different estimators in S4−S5 under case 2.

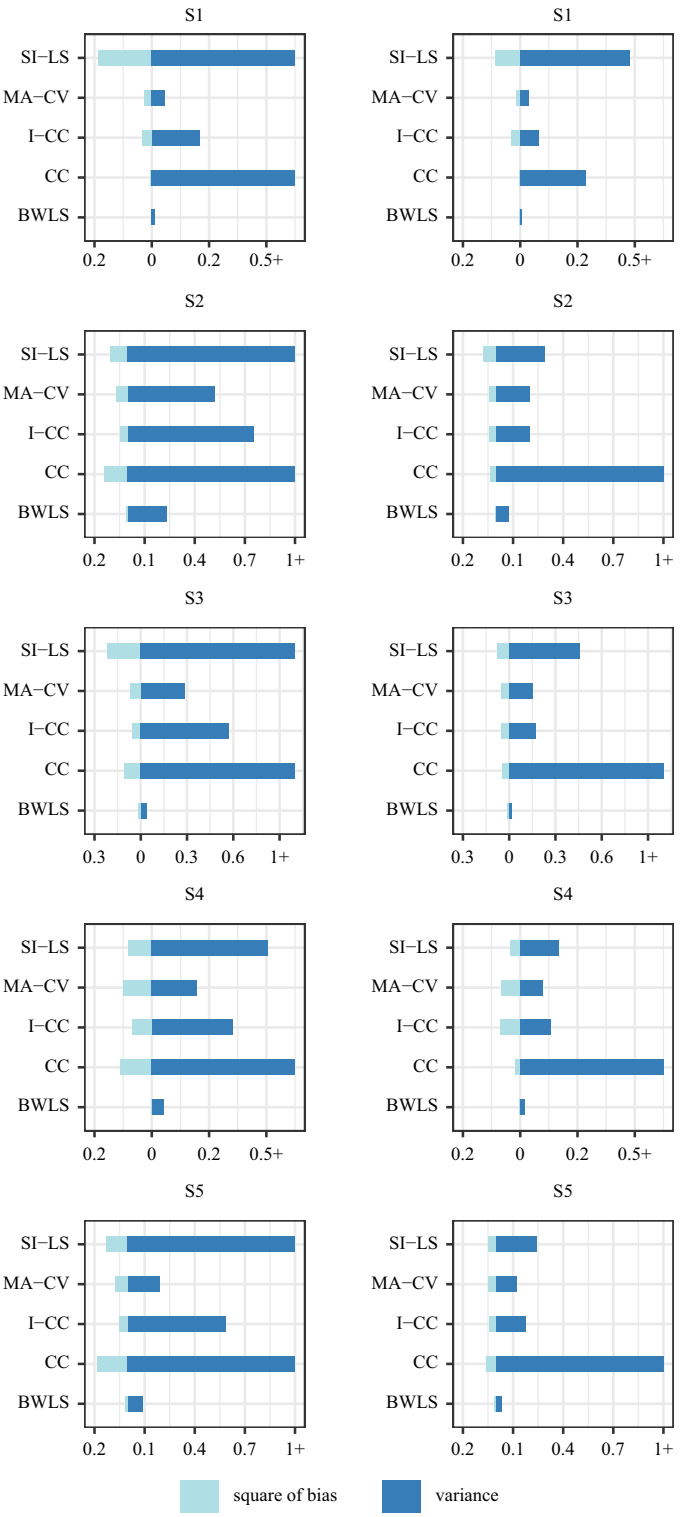
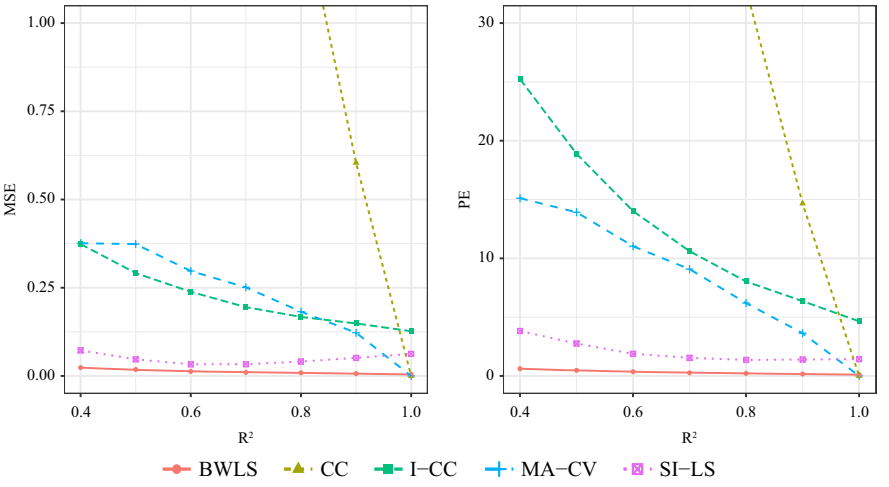
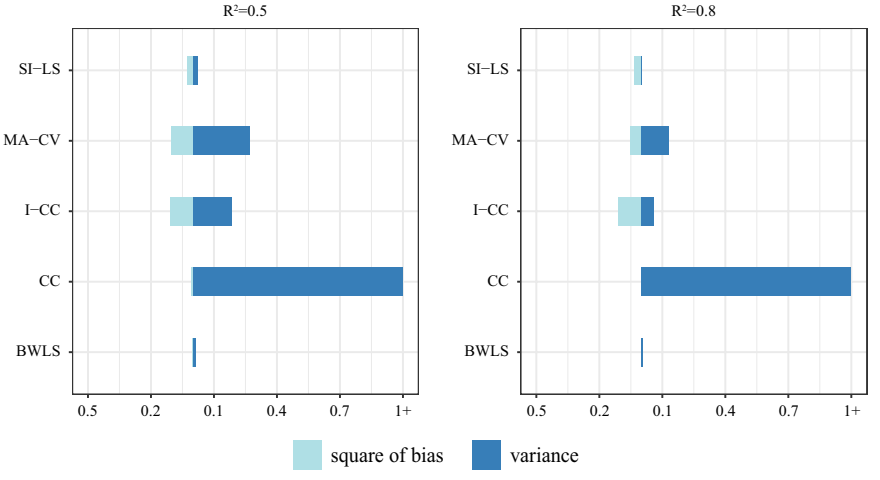


Fig. 10. The decomposition of MSE in case 2.

6.2. Results for Real Data Simulation



(a) The comparison of MSE and PE for different estimators.



(b) The decomposition of MSE.

Fig. 11. Results for real data simulation.

6.3. Results for BWLS and Mean Imputaion

We conducted a simulation study to compare the results of BWLS and mean imputation in terms of MSE and PE for Case 2. The results are presented in Figures 12 and 13

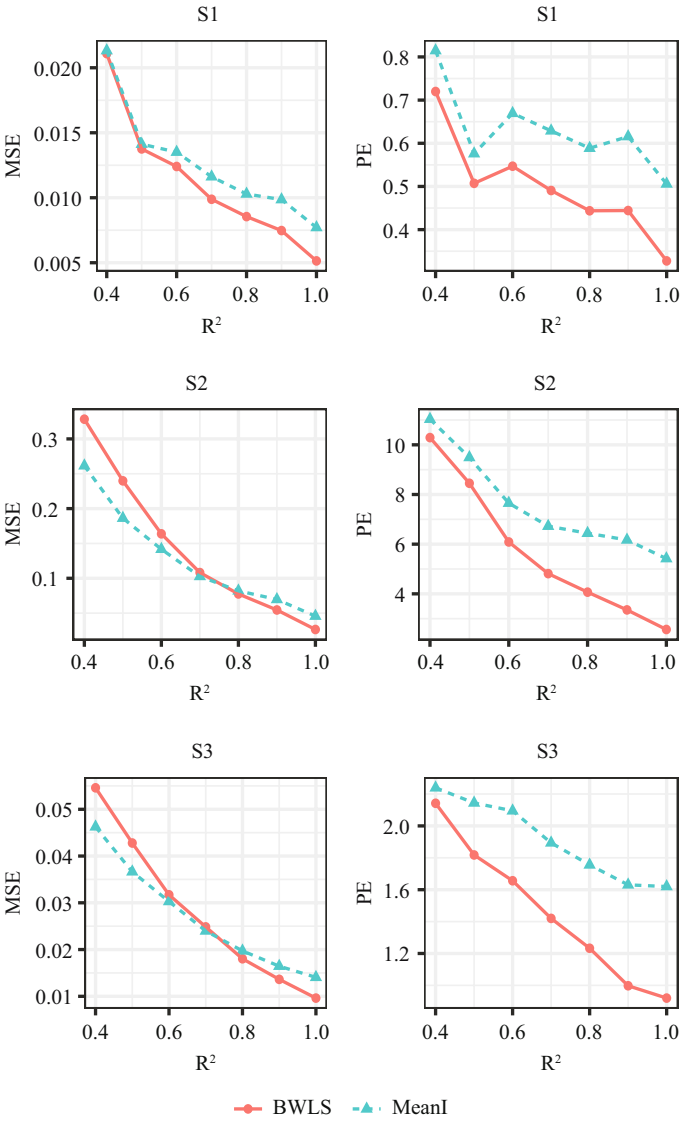


Fig. 12. The comparison of MSE and PE between BWLS and MeanI in S1–S3 under case 2.

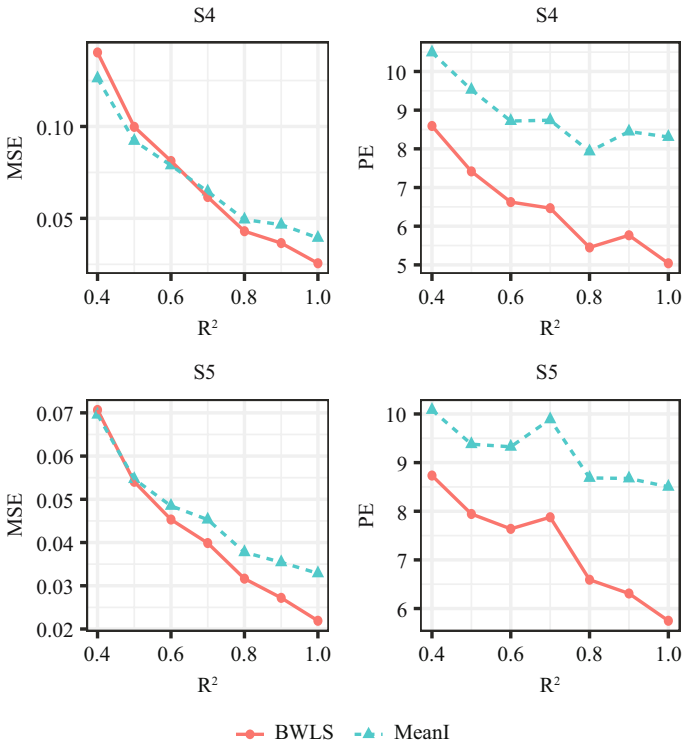


Fig. 13. The comparison of MSE and PE between BWLS and MeanI in S4–S5 under case 1.

6.4. Results for BWLS and Multiple Imputation

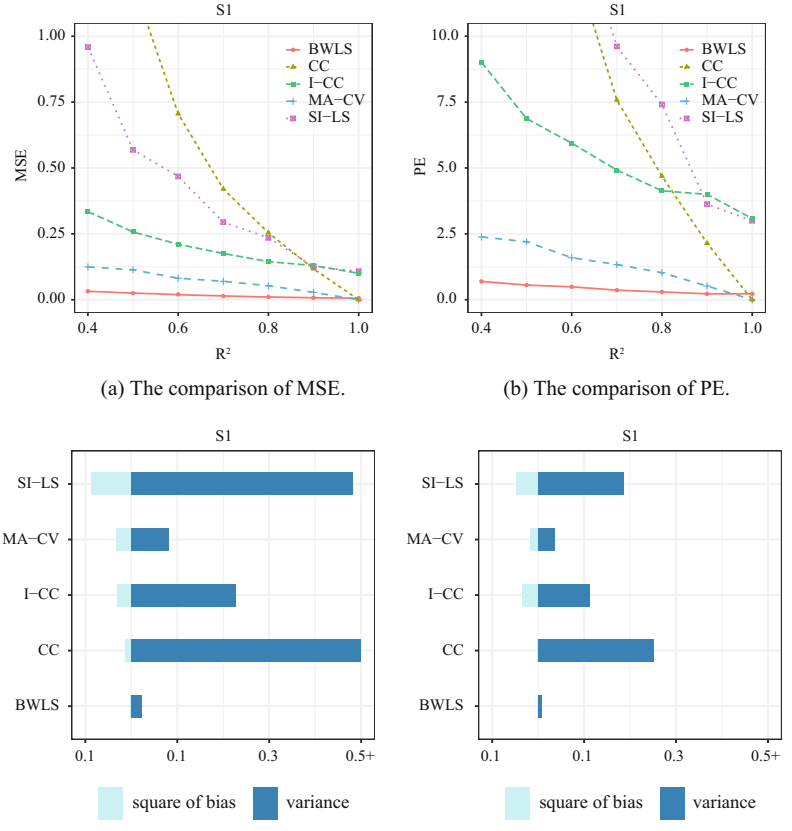
We conducted a simulation study to compare the results of BWLS and multiple imputation (MI). Here, we present the results in Table 5 of Scenarios S1-S4 with correlation structure in Case 1 ($\lambda_1 = 0.3$ and $\lambda_2 = 0.1$) considered in Simulation 1.

Table 5. Simulation results for BWLS and MI.

	R_2	*S*1		*S*2		*S*3		*S*4	
		BWLS	MI	BWLS	MI	BWLS	MI	BWLS	MI
MSE	0.4	0.0263	0.0490	0.1445	0.3860	0.0401	0.0853	0.0703	0.1132
	0.5	0.0187	0.0346	0.1014	0.2765	0.0300	0.0625	0.0534	0.0848
	0.6	0.0150	0.0274	0.0748	0.2088	0.0228	0.0494	0.0427	0.0626
	0.7	0.0112	0.0187	0.0568	0.1399	0.0172	0.0357	0.0365	0.0477
	0.8	0.0091	0.0139	0.0434	0.0952	0.0144	0.0274	0.0307	0.0337
	0.9	0.0074	0.0086	0.0327	0.0543	0.0111	0.0181	0.0261	0.0216
PE	0.4	0.5529	0.8364	5.8703	12.3396	1.7240	2.7974	4.7488	3.8173
	0.5	0.4219	0.5868	4.5167	8.7446	1.3387	2.0661	4.1224	2.7836
	0.6	0.3679	0.4685	3.7295	6.5603	1.1338	1.6330	3.7433	2.0942
	0.7	0.2926	0.3161	3.1695	4.3525	0.9484	1.1789	3.5797	1.5803
	0.8	0.2605	0.2357	2.7958	3.0022	0.8511	0.9038	3.4674	1.1129
	0.9	0.2291	0.1466	2.3951	1.6991	0.7417	0.6005	3.2825	0.7123

6.5. Results Under MAR

Based on the settings in Scenario S1 with Case 1 in Simulation 1, we generate one more variable, namely, $p = 21$. Let X_{i1} be observed for all respondents, and the remaining 20 variables are evenly divided into five modules. For generating MAR data, let $\delta_{ij} = 1$ if the i-th respondent answers the j-th sub-questionnaire and otherwise $\delta_{ij} = 0, i = 1, 2, ..., n$; $j = 1, 2, ..., q$. Here, we set $\delta_{ij} = I(\tau_{j-1} < X_{i1} <_{\tau_j}), i = 1, 2, ..., n$, where τ_j is (j/q)-th quantile of X_{i1}, which means that the missing probability depends on the value of Xi_1. The results are shown in Figure 14. It is observed that the performance of BWLS is superior to the competing methods in terms of MSE and PE, which is consistent with the results under missing completely at random. The conclusion is similar under other settings in the article and we omit the details here due to space limitation.



(a) The comparison of MSE.

(b) The comparison of PE.

(c) The decomposition of MSE under $R^2 = 0.5$.

(d) The decomposition of MSE under $R^2 = 0.8$.

Fig. 14. Simulation results for S1 with case 1 under MAR.

7. References

Adigüzel, F., and M. Wedel. 2008. "Split questionnaire design for massive surveys." *Journal of Marketing Research* 45(5): 608–617. DOI: https://doi.org/10.1509/jmkr.45.5.608.

Andreadis, I., and E. Kartsounidou. 2020. "The impact of splitting a long onlin equestionnaire on data quality." *Survey Research Methods* 14(1): 31–42. DOI: https://doi.org/10. 18148/srm/2020.v14i1.7294.

Beichelt, T., I. Hahn, F. Schimmelfennig, and S. Worschech. 2014. *Civil society and democracy promotion*. Springer. DOI: https://doi.org/10.1057/9781137291097.

Best, H. and, C. Wolf. 2013. *The SAGE handbook of regression analysis and causal inference*. Sage. DOI: https://doi.org/10.4135/9781446288146.

Cai, J., E.J. Candès, and Z. Shen. 2010. "A singular value thresholding algorithm for matrix completion." *SIAM Journal on Optimization* 20(4): 1956–1982. DOI: https://doi.org/10.1137/080738970.

Chipperfield, J.O., M.L. Barr, and D.G. Steel. 2018. "Split questionnaire designs: collecting only the data that you need through MCAR and MAR designs." *Journal of Applied Statistics* 45(8): 1465–1475. DOI: https://doi.org/10.1080/02664763.2017.1375085.

Chipperfield, J.O., and D.G. Steel. 2009. Design and Estimation for Split Questionnaire Surveys. *Journal of Official Statistics* 25(2): 227–244. Available at: https://ro.uow.edu.au/infopapers/3334/.

Chipperfield, J.O., and D.G. Steel. 2011. "Efficiency of split questionnaire surveys." *Journal of Statistical Planning and Inference* 141(5): 1925–1932. DOI: https://doi.org/10.1016/j.jspi.2010.12.003.

Chipperfield, J.O., and D.G. Steel. 2012. "Multivariate random effect models with complete and incomplete data." *Journal of Multivariate Analysis* 109: 146–155. DOI: https://doi.org/10.1016/j.jmva.2012.02.014.

Davidov, E., J. Cieciuch, and P. Schmidt. 2018. "The cross-country measurement comparability in the immigration module of the European Social Survey 2014-15." 12(1): 15–27. DOI: https://doi.org/10.18148/srm/2018.v12i1.7212.

Dziura, J.D., L.A. Post, Q. Zhao, Z. Fu, and P. Peduzzi. 2013. "Strategies for dealing with missing data in clinical trials: from design to analysis." *The Yale journal of biology and medicine* 86(3): 343. Available at: https://pubmed.ncbi.nlm.nih.gov/24058309/.

Early, K. 2016. *Dynamic question ordering: obtaining useful information while reducing user burden proposal*. Ph. D. thesis, Carnegie Mellon University Pittsburgh, PA. DOI: https://doi.org/10.1184/R1/6716123.v1.

Fang, F., W. Lan, J. Tong, and J. Shao. 2019. "Model averaging for prediction with fragmentary data." *Journal of Business & Economic Statistics* 37(3): 517–527. DOI: https://doi.org/10.1080/07350015.2017.1383263.

Hoerl, A.E. and R.W. Kennard (1970). "Ridge regression: biased estimation for nonorthogonal problems." *Technometrics* 12(1): 55–67. DOI: *Technometrics* 12(1): 55–67. DOI: https://doi.org/10.1080/00401706.1970.10488634.

Ioannidis, E., T. Merkouris, L.-C. Zhang, M. Karlberg, M. Petrakos, F. Reis, and P. Stavropoulos. 2016. "On a Mmodular Approach to the Design of Integrated Social

Surveys." *Journal of Official Statistics* 32(2): 259–286. DOI: https://doi.org/10.1515/-jos-2016-0013.

Ledoit, O., and M. Wolf. 2004. "A well-conditioned estimator for large-dimensional covariance matrices." *Journal of Multivariate Analysis* 88(2): 365–411. DOI: https://doi.org/10.1016/S0047-259X(03)00096-4.

Lesperance, M.L., and J.D. Kalbfleisch. 1992. "An algorithm for computing the nonparametric MLE of a mixing distribution." *Journal of the American Statistical Association* 87(417): 120–126. DOI: https://doi.org/10.1080/01621459.1992.10475182.

Little, R.J. 1992. "Regression with missing X's: a review." *Journal of the American statistical association* 87(420): 1227–1237. DOI: https://doi.org/10.1080/01621459.1992.10476282.

Little, R.J., and M.D. Schluchter. 1985. "Maximum likelihood estimation for mixed continuous and categorical data with missing values." *Biometrika* 72(3): 497–512. DOI: https://doi.org/10.1093/biomet/72.3.497.

Liu, M., and L. Wronski. 2018. "Examining completion rates in web surveys via over 25,000 real-world surveys." *Social Science Computer Review* 36(1): 116–124. DOI: https://doi.org/10.1177/0894439317695581.

Mazumder, R., T. Hastie, and R. 2010. "Spectral regularization algorithms for learning large incomplete matrices." *Journal of Machine Learning Research* 11: 2287–2322. DOI: https://dl.acm.org/doi/10.5555/1756006.1859931.

Merkouris, T. 2015. "An efficient estimation method for matrix survey sampling." *Survey Methodology* 41(1): 237–262. DOI: https://www150.statcan.gc.ca/n1/en/catalogue/ 12-001-X201500114174.

Neidorf, T., and M. Sheehan. 2014. "National Assessment of Educational Progress (NAEP)." In *Encyclopedia of Science Education*., edited by R. Gunstone. Dordrecht: Springer. DOI: https://doi.org/10.1007/978-94-007-6165-0_67-2.

Peytchev, A., and E. Peytcheva. 2017. "Reduction of measurement error due to survey length: evaluation of the split questionnaire design approach." *Survey Research Methods* 11(4): 361–368. DOI: https://doi.org/10.18148/srm/2017.v11i4.7145.

Raghunathan, T.E., and J.E. Grizzle. 1995. "A split questionnaire survey design." *Journal of the American Statistical Association* 90(429): 54–63. DOI: https://doi.org/10.1080/01621459.1995.10476488.

Rhemtulla, M., and T.D. Little. 2012. "Planned missing data designs for research in cognitive development." *Journal of Cognition and Development* 13(4): 425–438. DOI: https://doi.org/10.1080/15248372.2012.717340.

Rust, K.F., and E.G. Johnson. 1992. "Chapter 2: Sampling and weighting in the national assessment." *Journal of Educational Statistics* 17(2): 111–129. DOI: https://doi.org/10.3102/10769986017002111.

Schnaudt, C., M. Weinhardt, R. Fitzgerald, and S. Liebig. 2014. "The European Social Survey: contents, design, and research potential." *Journal of Contextual Economics: Schmollers Jahrbuch* 134(4): 487–506. DOI: https://doi.org/10.3790/schm.134.4.487.

Skinner, C.J., and O. Coker. 1996. "Regression analysis for complex survey data with missing values of a covariate." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 159(2): 265–274. DOI: https://doi.org/10.2307/2983173.

Van Ham, C., J.J. Thomassen, K. Aarts, and R.B. Andeweg. 2017. *Myth and reality of the legitimacy crisis: explaining trends and cross-national differences in established democracies*. Oxford University Press. DOI: https://doi.org/10.1093/oso/9780198793717.001.0001.

Vonneilich, N., D. Lüdecke, and O. von dem Knesebeck. 2019. "Educational inequalities in self-rated health and social relationships–analyses based on the European Social Survey 2002–2016." *Social Science and Medicine*. DOI: https://doi. org/10.1016/j.-socscimed.2019.112379.

Wu, H., and S.-O. Leung. 2017. "Can likert scales be treated as interval scales? – a simulation study." *Journal of Social Service Research* 43(4): 527–532. DOI: https://doi.org/10.1080/01488376.2017.1329775.

Xiang, S., L. Yuan, W. Fan, Y. Wang, P.M. Thompson, and J. Ye. 2014. "Bi-level multi-source learning for heterogeneous block-wise missing data." *NeuroImage* 102: 192–206. DOI: https://doi.org/10.1016/j.neuroimage.2013.08.015.

Yang, S., and J.K. Kim 2017. "A semiparametric inference to regression analysis with missing covariates in survey data." *Statistica Sinica* 27: 261–285. DOI: https://doi.org/10.5705/ss.2014.174.

Yu, G., Q. Li, D. Shen, and Y. Liu. 2020. "Optimal sparse linear prediction for blockmissing multi-modality data without imputation." *Journal of the American Statistical Association* 115(531): 1406–1419. DOI: https://doi.org/10.1080/01621459.2019.1632079.

Yuan, K.-H., and P.M. Bentler. 2010. "Consistency of normal-distribution-based pseudo maximum likelihood estimates when data are missing at random." *American Statistician* 64(3): 263–267. DOI: https://doi.org/10.1198/tast.2010.09203.

Received September 2021
Revised August 2022
Accepted March 2023