Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

A note on multiply robust predictive mean matching imputation with complex survey data

by Sixia Chen, David Haziza and Alexander Stubblefield

Release date: June 24, 2021

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

Email at STATCAN.infostats-infostats.STATCAN@canada.ca

Telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

• Statistical Information Service	1-800-263-1136
• National telecommunications device for the hearing impaired	1-800-363-7629
• Fax line	1-514-283-9350

Depository Services Program

• Inquiries line	1-800-635-7943
• Fax line	1-800-565-7757

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

An HTML version is also available.

Cette publication est aussi disponible en français.

Survey Methodology, June 2021
Vol. 47, No. 1, pp. 215-222
Statistics Canada, Catalogue No. 12-001-X

A note on multiply robust predictive mean matching imputation with complex survey data

Sixia Chen, David Haziza and Alexander Stubblefield[1]

Abstract

Predictive mean matching is a commonly used imputation procedure for addressing the problem of item nonresponse in surveys. The customary approach relies upon the specification of a single outcome regression model. In this note, we propose a novel predictive mean matching procedure that allows the user to specify multiple outcome regression models. The resulting estimator is multiply robust in the sense that it remains consistent if one of the specified outcome regression models is correctly specified. The results from a simulation study suggest that the proposed method performs well in terms of bias and efficiency.

Key Words: Multiple robustness; Nearest-neigbour imputation; Survey data; Variance estimation.

1. Introduction

Predictive mean matching (PMM), a procedure closely related to nearest-neighbour imputation (NNI, Chen and Shao, 2000; Beaumont and Bocci, 2009; Yang and Kim, 2019), is a popular imputation procedure in practice (Little, 1988; Yang and Kim, 2020). In NNI, a missing value to a survey variable y is replaced by the y-value of the closest respondent with respect to a vector of fully observed variables $\mathbf{x}$. However, with NNI, the resulting imputed estimator may suffer from a non-negligible bias when the dimension of $\mathbf{x}$ is large (Yang and Kim, 2019), a problem often referred to as the curse of dimensionality. In contrast, PMM starts with fitting a parametric model (e.g., a linear regression model) based on the responding units with y as the response variable and $\mathbf{x}$ as the set of explanatory variables. This leads to a set of predicted values or scores, $\hat{m}$, for all the sample units (respondents and nonrespondents). A missing value to the survey variable y is then replaced by the y-value of the closest respondent with respect to $\hat{m}$. The latter may be viewed as a scalar summary of the information contained in the vector $\mathbf{x}$. Therefore, unlike NNI, PMM is not sensitive to the dimension of $\mathbf{x}$, which is a desirable feature.

Both NNI and PMM belong to the class of nonparametric procedures. Therefore, both procedures are less vulnerable to model misspecification unlike parametric methods (e.g., linear regression imputation). Also, both NNI and PMM belong to the class of donor imputation procedures; that is, they produce eligible imputed values as they use actual observed values "borrowed" from the respondents.

In the first step of PMM, the information contained in the vector $\mathbf{x}$ is compressed into a single score $\hat{m}$ through the use of a parametric model (e.g., a linear regression model). If the specified model provides an accurate description of the relationship linking y and $\mathbf{x}$, we expect PMM to perform well in terms of bias. On the other hand, if the specified model is grossly misspecified, PMM may yield biased estimators.

1. Sixia Chen, Department of Biostatistics and Epidemiology, University of Oklahoma Health Sciences Center, Oklahoma City, OK 73104, U.S.A.; David Haziza, Department of mathematics and statistics, University of Ottawa, Ottawa, Canada. E-mail: dhaziza@uottawa.ca; Alexander Stubblefield, Department of Economics, Michigan State University, East Lansing, MI 48823, U.S.A.

Multiply robust approaches with multiple outcome regression and nonresponse models have been shown to improve the robustness against model misspecification, see Han and Wang (2013), Han (2014), and Chen and Haziza (2019a) among others. In this note, we propose a novel PMM procedure that allows for multiple models, each which may be based on a different functional and/or a different set of explanatory variables. Postulating multiple models may prove useful in a number of situations; e.g., see Chen and Haziza (2017) and Chen and Haziza (2019b) for a discussion. The specified models may be parametric or nonparametric. The rationale behind the proposed method is to fit each of these specified models based on the responding units, which leads to multiple set of predicted values (scores) for all the sample units. After describing the theoretical setup in Section 2, we show how to combine these scores to construct the imputed values in Section 3. The proposed PMM procedure is multiply robust in the sense that the resulting estimator is consistent if all but one model are misspecified. Because the true model linking y and $\mathbf{x}$ is unknown, the proposed approach is attractive because it provides some protection against model misspecification. Also, unlike the multiply robust imputation procedure considered in Chen and Haziza (2017), the proposed method belongs to the class of donor imputation procedures. In Section 4, we conduct a simulation study to assess the performance of the proposed method in terms of bias and efficiency.

2. Basic setup

Consider a finite population $\mathscr{F}_N = \{(\mathbf{x}_i, y_i), i = 1, 2, \ldots, N\}$, assumed to have been generated from the following superpopulation model:

$$y_i = m(\mathbf{x}_i) + \varepsilon_i, \tag{2.1}$$

where $m(\cdot)$ is an unknown functional, $\mathbf{x}_i$ is a vector of fully observed variables attached to unit i, and the ε_i's are mutually independent random variables such that $\mathbb{E}(\varepsilon_i \mid \mathbf{x}_i) = 0$ and $\mathbb{V}(\varepsilon_i \mid \mathbf{x}_i) = \sigma^2$. For simplicity, we assume that the variance structure is homoscedastic but our method can be easily extended to the case of unequal variances.

The interest lies in estimating the population mean, $\theta = \mathbb{E}(y)$. Given the finite population, a probability sample S, of size n, is selected according to a sampling design with first-order inclusion probabilities π_i and second-order inclusion probabilities π_{ij}. The sampling weight attached to unit i is denoted by $w_i = \pi_i^{-1}$.

Let r_i be response indicator attached to unit i such that $r_i = 1$ if y_i is observed, and $r_i = 0$ if y_i is missing. Let $S_r = \{i \in S: r_i = 1\}$ denote the set of respondents to the survey variable y. We assume that the data are Missing At Random (MAR):

$$\Pr(r_i = 1 \mid \mathbf{x}_i, y_i) = \Pr(r_i = 1 \mid \mathbf{x}_i). \tag{2.2}$$

The customary PMM procedure can be described as follows. We first postulate a parametric outcome regression model $\mathscr{M} = \{m(\mathbf{x}_i; \boldsymbol{\beta})\}$, where $\boldsymbol{\beta}$ is a vector of unknown parameters (Yang and Kim, 2020). For $i \in S$, we compute the score $\hat{m}_i = m(\mathbf{x}_i; \hat{\boldsymbol{\beta}})$, where $\hat{\boldsymbol{\beta}}$ is a suitable estimator of $\boldsymbol{\beta}$ based on the responding units. Then, the imputed value for the missing y_i is $y_i^* = y_j$, where j is the index of the nearest-neighbour of unit i, which satisfies $\mathcal{D}(\hat{m}_j, \hat{m}_i) \le \mathcal{D}(\hat{m}_{j'}, \hat{m}_i)$ for any $j' \in S_r$, where $\mathcal{D}(\cdot,\cdot)$ denotes a distance function; e.g., the Euclidean distance. In order for PMM to be robust against misspecification, the specified parametric model must satisfy the Lipschitz continuity condition (Yang and Kim, 2020). This condition may not be satisfied for some commonly used models and functional forms, including quadratic models; see Yang and Kim (2020) for a discussion.

3. Proposed method

The proposed method allows the user to specify multiple outcome regression models for the survey variable y. This grants a greater probability of selecting a model that performs well at replicating the relationship between the response variable and the explanatory variables, making the approach multiply robust without requiring the Lipschitz continuity condition to hold. As long as one of the specified models is correctly specified, the resulting estimator will be consistent.

We consider a class of outcome regression models: $\mathcal{M} = \{m^{(k)}(\mathbf{x}_i; \boldsymbol{\beta}^{(k)}), k = 1, 2, \ldots, K\}$. To impute the missing values, we proceed as follows:

(Step1). Obtain the estimators $\hat{\boldsymbol{\beta}}^{(k)}$ of $\boldsymbol{\beta}^{(k)}, k = 1, 2, \ldots, K,$ by solving the following survey weighted estimating equations:

$$\hat{U}_m^{(k)}(\hat{\boldsymbol{\beta}}^{(k)}) = \sum_{i \in S} w_i r_i \{y_i - m^{(k)}(\mathbf{x}_i; \boldsymbol{\beta}^{(k)})\} \frac{\partial m^{(k)}(\mathbf{x}_i; \boldsymbol{\beta}^{(k)})}{\partial \boldsymbol{\beta}^{(k)}} = 0. \tag{3.1}$$

(Step2). For $i \in S$, obtain the K-vector of predicted values

$$\mathbf{V}_i = \left(m^{(1)}(\mathbf{x}_i; \hat{\boldsymbol{\beta}}^{(1)}), m^{(2)}(\mathbf{x}_i; \hat{\boldsymbol{\beta}}^{(2)}), \ldots, m^{(K)}(\mathbf{x}_i; \hat{\boldsymbol{\beta}}^{(K)})\right)^{\top}.$$

(Step3). Fit a weighted linear regression model without intercept with y as the response variable and $\mathbf{V}$ as the vector of explanatory variables. Let $\hat{M}(\mathbf{x}_i; \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\eta}})$ be the resulting predicted value attached to unit i:

$$\hat{M}(\mathbf{x}_i; \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\eta}}) = \mathbf{V}_i^{\top} \hat{\boldsymbol{\eta}},$$

where $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}^{(1)}, \hat{\boldsymbol{\beta}}^{(2)}, \ldots, \hat{\boldsymbol{\beta}}^{(K)})$ and

$$\hat{\boldsymbol{\eta}} = \left\{\sum_{i \in S} w_i r_i \mathbf{V}_i \mathbf{V}_i^{\top}\right\}^{-1} \sum_{i \in S} w_i r_i \mathbf{V}_i y_i. \tag{3.2}$$

(Step4). The imputed value for the missing y_i is $y_i^* = y_j$, where j is the index of the nearest-neighbour of unit i, which satisfies $\mathcal{D}\{\hat{M}(\mathbf{x}_j; \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\eta}}), \hat{M}(\mathbf{x}_i; \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\eta}})\} \le \mathcal{D}\{\hat{M}(\mathbf{x}_{j'}; \hat{\boldsymbol{\eta}}), \hat{M}(\mathbf{x}_i; \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\eta}})\}$ for any $j' \in S_r$.

After applying (Step1)-(Step4), we construct the imputed estimator of θ:

$$\hat{\theta}_{\mathrm{MR}} = \frac{1}{\hat{N}} \sum_{i \in S} w_i \{r_i y_i + (1 - r_i)\ y_i^*\}, \tag{3.3}$$

where $\hat{N} = \sum_{i \in S} w_i$. Using an approach similar to the one used by Yang and Kim (2020), it can be shown that the estimator $\hat{\theta}_{\mathrm{MR}}$ is multiply robust in the sense that it is consistent if all but one model are misspecified.

Estimating the variance of $\hat{\theta}_{\mathrm{MR}}$ can be done through replication variance estimation procedures; see e.g., Rust and Rao (1996) and Wolter (2007). In the context of PMM for survey data, Yang and Kim (2020) also considered replication procedures. Let L denote the number of replicates and $w_i^{(g)}$ be a replication weight attached to unit i in the g^{th} replicate. A replication variance estimator of $\hat{\theta}_{\mathrm{MR}}$ is given by

$$\hat{V}_{\mathrm{rep}}\left(\hat{\theta}_{\mathrm{MR}}\right) = \sum_{g=1}^{L} c_g \left(\hat{\theta}_{\mathrm{MR}}^{(g)} - \hat{\theta}_{\mathrm{MR}}\right)^2, \tag{3.4}$$

where

$$\hat{\theta}_{\mathrm{MR}}^{(g)} = \sum_{i \in S} w_i^{(g)} \{r_i y_i + (1 - r_i)\ y_i^{*(g)}\}$$

denote the estimator $\hat{\theta}_{\mathrm{MR}}$ in the g^{th} replicate with $y_i^{*(g)}$ denoting the imputed value attached to unit i in the g^{th} replicate, obtained from (Step1)-(Step4) above, based on the replication weight $w_i^{(g)}$ instead of the original weights w_i. The factor c_g in (3.4) is determined by the replication method. For instance, with the delete-one jackknife, we have $L = n,\ c_g = n/(n-1)$ and $w_i^{(g)} = n/(n-1)\, w_i$ if $i \neq g$ and $w_i^{(g)} = 0$ if $i = g$.

4. Simulation study

To assess the performance of the proposed method in terms of bias and efficiency, we conducted a limited simulation study. We generated $B = 2{,}000$ finite populations, each of size $N = 20{,}000$. First, the explanatory variables x_1 - x_4 were generated from a multivariate standard normal distribution. Then, given x_1 - x_4, we generated the survey variable y according to the following outcome regression models:

(M1). $y = 1 + x_1 + x_2 + x_3 + x_4 + \varepsilon,$ where $\varepsilon \sim N(0, 1)$.

(M2). $y = 1 + x_1^2 + x_2^2 + x_3 + x_4 + x_3 x_4 + \varepsilon,$ where $\varepsilon \sim N(0, 1)$.

Note that both (M1) and (M2) are linear models based on the explanatory variables x_1 - x_4, except that (M2) includes quadratic terms and an interaction term.

From each finite population, a probability sample S was selected according to probability proportional-to-size (PPS) systematic sampling based on the size variable $z_i = \log(0.1\,|\,y_i + v_i\,| + 4)$, where $v_i \sim N(0, 1)$. The first-order inclusion probabilities are given by $\pi_i = nz_i \big/ \sum_{i=1}^{N} z_i$ with $n =$ 200, 500 and 1,000.

In each sample, the response indicators r_i were generated from a Bernoulli distribution with probability p_i, where

$$p_i = 0.1 + 0.9 \times \frac{\exp(\alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \alpha_3 x_{3i} + \alpha_4 x_{4i})}{1 + \exp(\alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \alpha_3 x_{3i} + \alpha_4 x_{4i})}. \quad (4.1)$$

We used two sets of values for $(\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4)$: $(0, 1, 1, 1, 1)$ and $(1.38, 1, 1, 1, 1)$. These led to response rates approximately equal to 70%, and 50%, respectively.

We computed the following estimators of θ

(Naive). The weighted mean of the respondents, $\hat{\theta}_{\text{naive}} = \sum_{i \in S_r} w_i y_i \big/ \sum_{i \in S_r} w_i$.

(Reg). The imputed estimator based on deterministic linear regression imputation, assuming the model (M1).

(PMM1). The imputed estimator based on PMM, where the score $\hat{m}_i,\, i \in S,$ was obtained by fitting the model (M1).

(New1). The imputed estimator based on the proposed multiply robust PMM procedure using both models (M1) and (M2).

(New2). The imputed estimator based on the proposed multiply robust PMM procedure using models (M1), (M2), and two additional models (M3) and (M4), where (M3) uses x_1 only as the predictor and (M4) uses x_1^2 only as the predictor.

We computed the Monte Carlo relative bias (MCRB), the Monte Carlo relative standard error (MCRSE) and the Monte Carlo relative root mean squared error (MCRMSE), defined respectively as

$$\text{MCRB} = \frac{2{,}000^{-1} \sum_{b=1}^{2{,}000} (\hat{\theta}_b - \theta_b)}{\theta_{\text{MC}}},$$

$$\text{MCRSE} = \frac{\sqrt{(B-1)^{-1} \sum_{b=1}^{B} (\hat{\theta}_b - \hat{\theta}_{\text{MC}})^2}}{\theta_{\text{MC}}}$$

and

$$\text{MCRMSE} = \frac{\sqrt{(B-1)^{-1} \sum_{b=1}^{B} (\hat{\theta}_b - \theta_{\text{MC}})^2}}{\theta_{\text{MC}}},$$

where θ_b denotes the population mean in the b^{th} population, $\hat{\theta}_b$ denotes the estimator $\hat{\theta}$ in the b^{th} sample, $b = 1, \ldots, 2{,}000$, and

$$\theta_{\text{MC}} = \frac{1}{2{,}000} \sum_{b=1}^{2{,}000} \theta_b, \quad \hat{\theta}_{\text{MC}} = \frac{1}{2{,}000} \sum_{b=1}^{2{,}000} \hat{\theta}_b.$$

The results are presented in Tables 4.1 and 4.2. The naive estimator exhibited a significant bias in all the scenarios, as expected. When the true model was given by (M1), we note from Table 4.1 that linear regression imputation performed very well in terms of bias, as expected. Both PMM and the proposed method showed negligible bias for $n = 1{,}000$ and a slight bias for $n = 500$ and $n = 200$. For instance, for $n = 200$ and a response rate of 70%, the value of RB was equal to 2.4% for PMM, New1 and New2. In terms of efficiency, linear regression imputation slightly outperformed both PMM and the proposed methods, as expected. For instance for $n = 1{,}000$ and a response rate of 70%, the value of RMSE was equal to 7.5% for linear regression imputation and equal to 8.0% for both PMM, New1 and New2. It is worth pointing out that both PMM and the proposed methods exhibited almost identical performances in all the scenarios presented in Table 4.1. Therefore, incorporating two additional models did not seem to affect the efficiency of the resulting estimator (New2).

When the true model was given by (M2), we note from Table 4.2 that both linear regression imputation and PMM led to significant biases in all the scenarios, as expected. Being a parametric imputation procedure, linear regression imputation is vulnerable to model misspecification. On the other hand, PMM showed smaller biases than linear regression imputation, suggesting some robustness against model misspecification. For instance, for $n = 1{,}000$ and a response rate of 70%, the value of RB was equal to -9.2% for linear regression imputation and -3.7% for PMM. The proposed methods outperformed both linear regression imputation and PMM in terms of bias, standard error and mean square error in all the scenarios. Finally, both New1 and New2 exhibited almost identical performances.

Table 4.1
Monte Carlo relative bias (MCRB), relative standard error (MCRSE), and relative root mean squared error (MCRMSE) when the true model is (M1)

			Method				
Response rate	**Sample Size**	**Measure $(\times 10^2)$**	**Naive**	**Reg**	**PMM1**	**New1**	**New2**
70%	1,000	MCRB	64.7	-0.1	0.4	0.4	0.4
		MCRSE	7.5	7.5	8.0	8.0	8.0
		MCRMSE	65.1	7.5	8.0	8.0	8.0
70%	500	MCRB	65.3	0.5	1.4	1.4	1.4
		MCRSE	10.7	10.4	11.2	11.2	11.2
		MCRMSE	66.1	10.4	11.3	11.3	11.3
70%	200	MCRB	64.6	0.3	2.4	2.4	2.4
		MCRSE	16.5	16.7	17.5	17.5	17.6
		MCRMSE	66.7	16.7	17.7	17.7	17.7
50%	1,000	MCRB	99.3	0.0	0.7	0.7	0.6
		MCRSE	8.8	8.1	9.0	9.0	9.0
		MCRMSE	99.7	8.1	9.1	9.1	9.1
50%	500	MCRB	98.9	-0.1	1.3	1.3	1.3
		MCRSE	12.1	11.2	12.5	12.5	12.5
		MCRMSE	99.6	11.2	12.6	12.6	12.6
50%	200	MCRB	99.8	0.8	4.3	4.3	4.4
		MCRSE	19.3	17.7	19.6	19.6	19.6
		MCRMSE	101.6	17.7	20.1	20.1	20.0

Table 4.2
Monte Carlo relative bias (MCRB), relative standard error (MCRSE), and relative root mean squared error (MCRMSE) when the true model is (M2)

			Method				
Response rate	**Sample Size**	**Measure $(\times 10^2)$**	**Naive**	**Reg**	**PMM1**	**New1**	**New2**
70%	1,000	MCRB	7.5	-9.2	-3.7	0.1	0.1
		MCRSE	3.5	3.5	3.9	3.1	3.1
		MCRMSE	8.2	9.9	5.4	3.1	3.1
70%	500	MCRB	7.5	-9.4	-4.0	0.2	0.2
		MCRSE	5.0	5.1	5.6	4.5	4.5
		MCRMSE	9.0	10.7	6.9	4.5	4.5
70%	200	MCRB	7.6	-9.2	-4.0	0.1	0.1
		MCRSE	7.8	7.9	8.5	6.8	6.8
		MCRMSE	10.9	12.1	9.4	6.8	6.8
50%	1,000	MCRB	16.6	-11.3	-3.1	0.3	0.3
		MCRSE	4.0	4.5	5.0	3.3	3.3
		MCRMSE	17.1	12.2	5.9	3.3	3.3
50%	500	MCRB	16.5	-11.5	-3.5	0.3	0.3
		MCRSE	5.7	6.3	7.0	4.8	4.7
		MCRMSE	17.5	13.2	7.8	4.8	4.8
50%	200	MCRB	16.5	-12.0	-3.9	-0.1	-0.1
		MCRSE	9.1	9.9	11.0	7.4	7.4
		MCRMSE	18.8	15.6	11.7	7.4	7.4

Acknowledgements

S. Chen was supported by the National Institute on Minority Health and Health Disparities (NIMHD) at National Institutes of Health (NIH) (1R21MD014658-01A1) and the Oklahoma Shared Clinical and Translational Resources (U54GM104938) with an Institutional Development Award (IDeA) from National Institute of General Medical Sciences. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The work of D. Haziza was supported by grants from the Natural Sciences and Engineering Research Council of Canada.

References

Beaumont, J.-F., and Bocci, C. (2009). Variance estimation when donor imputation is used to fill in missing values. *Canadian Journal of Statistics*, 37, 400-416.

Chen, S., and Haziza, D. (2017). Multiply robust imputation procedures for the treatment of item nonresponse in surveys. *Biometrika*, 102, 439-453.

Chen, S., and Haziza, D. (2019a). Multiply robust nonparametric multiple imputation for the treatment of missing data. *Statistica Sinica*, 29, 2035-2053.

Chen, S., and Haziza, D. (2019b). Recent developments in dealing with item nonresponse in surveys: A critical review. *International Statistical Review*, 87, S192-S218.

Chen, J., and Shao, J. (2000). Nearest-neighbour imputation for survey data. *Journal of Official Statistics*, 16, 583-599.

Han, P. (2014). Multiply robust estimation in regression analysis with missing data. *Journal of the American Statistical Association*, 109, 1159-1173.

Han, P., and Wang, L. (2013). Estimation with missing data: Beyond double robustness. *Biometrika*, 100, 417-430.

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

Rust, K.F., and Rao, J.N.K. (1996). Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research*, 5, 283-310.

Wolter, K. (2007). *Introduction to Variance Estimation, 2nd Edition*. Springer, Berlin.

Yang, S., and Kim, J.K. (2019). Nearest neighbor imputation for general parameter estimation in survey sampling. *Advances in Econometrics - The Econometrics of Complex Survey Data: Theory and Applications*, 39, 209-234.

Yang, S., and Kim, J.K. (2020). Predictive mean matching imputation in survey sampling. To appear in the *Scandinavian Journal of Statistics*.