SSC Annual Meeting, May 2021

Proceedings of the Survey Methods Section

ITEM IMPUTATION WITH A NONPROBABILITY SAMPLE

Abel Dasylva[1]

ABSTRACT

This study considers the problem of estimating a population total from two samples, including a probability sample and a larger nonprobability sample, where the variable of interest is present but missing at random in both samples. To this end, it evaluates various strategies for imputing the missing responses in the probability sample from the observed responses and covariates in both samples. These strategies include semiparametric solutions using a known conditional mean response function and Lasso regression. They also include nonparametric methods such as nearest neighbor and regression trees. These different solutions are evaluated in simulations and in an empirical study. They are also compared to the imputation of the missing responses within the probability sample.

KEY WORDS: data integration, missing data

RÉSUMÉ

Cette étude considère le problème de l'estimation d'un total de population à partir de deux échantillons, incluant un échantillon probabiliste et un plus grand échantillon nonprobabiliste, où la variable d'intérêt est présente mais manquante au hasard dans les deux échantillons. À cette fin, elle évalue diverses stratégies pour imputer les réponses manquantes dans l'échantillon probabiliste à partir des réponses et des variables explicatives des deux échantillons. Ces stratégies incluent des solutions paramétriques basées sur une fonction connue donnant la moyenne conditionnelle de la réponse et la régression Lasso. Elles incluent aussi des méthodes nonparamétriques telles que le plus proche voisin et les arbres de régression. Ces différentes solutions sont évaluées par des simulations et une étude empirique. Elles sont aussi comparées à l'imputation des réponses à l'intérieur de l'échantillon probabiliste.

MOTS CLÉS : intégration de données, données manquantes

1 INTRODUCTION

Statistics Canada has undertaken many activities to collect data on the impact of the COVID 19 pandemic, including the third iteration of the Canadian Perspective Survey Series (CPSS in short), a probability web panel, from June 15 to June 21, 2020. In the same period, the agency also collected data through a crowd-sourced sample, using some of the CPSS questions, resulting in a much larger sample of respondents and virtually no missing responses. Given the shared content and the sizes of the two samples, an interesting question is how to use the crowd-sourced data to mitigate the nonresponse in the web panel, ideally with minimal changes to the existing methodology.

This example is a specific instance of the general problem where a probability sample and a much larger nonprobability sample are available, both containing some auxiliary variables and responses, which are possibly missing in the probability sample. The general question is how to best combine the information from the two samples to estimate the finite population total or the related mean. Potential options include

[1] Abel Dasylva, R.-H. Coats building, 100 Tunney's pasture driveway Ottawa ON, K1A0T6, abel.dasylva@statcan.gc.ca

imputing each missing item in the probability sample with the help of the nonprobability sample and re-weighting the nonprobability sample. In what follows, our focus is on the first option, i.e. imputing the missing items. A special case occurs when all the responses are missing in the probability sample. For this situation, Chen, Li & Wu (2020) have described a re-weighting solution, while Rivers (2007), Kim et al. (2020), Chen, Yang & Kim (2020) and Yang et al. (2021) have developed mass imputation methods, including semiparametric and nonparametric mean imputation, fractional imputation and nearest-neighbor imputation. This paper explores possible extensions of these imputation strategies, when the probability sample contains some responses, to shed some light on the following basic questions. When is it beneficial to use the nonprobability sample to impute the missing values in the probability sample? In such cases, what is the best way to combine the information from the two samples? To gain some insight, this work considers categorical variables, a single response, and the imputation of each missing item by a single draw from a Bernoulli distribution according to the conditional mean response, where the latter is estimated semi-parametrically or non-parametrically. The proposed estimators are also shown to be consistent and asymptotically normal under general conditions.

The remaining sections are organized as follows. Section 2 describes the notations and assumptions. It is followed by the theory (Section 3), the simulations (Section 4), the empirical study (Section 5) and the conclusion (Section 6).

2 NOTATIONS AND ASSUMPTIONS

Let U_N denote the finite population that is comprised of N units, where unit k is associated with the binary response y_k and the fixed categorical covariates $\boldsymbol{x}_k$ in some finite set $\mathcal{X}$ with cardinal $|\mathcal{X}|$. Let $Y = \sum_{k \in U_N} y_k$ and $\overline{Y} = Y/N$ denote the finite population total and the related mean, respectively. Our goal is estimating the finite population mean $\overline{Y} = Y/N$ that is equal to the finite population proportion because the response is binary. The two samples include the size-n_P probability sample s_P and the size-n_{NP} nonprobability sample s_{NP}. For simplicity, suppose that all the responses are observed in the nonprobability sample, without any measurement error. For sample s_P and unit k, let μ_k, π_k, r_k denote the mean response, the first order inclusion probability and response indicator, respectively. Also define $a_k = I(k \in s_P)$ and $\delta_k = I(k \in s_{NP})$.

The assumptions include independent samples, inclusion at random in each sample, missing at random responses in s_P and no measurement error in the nonprobability sample, where the first assumption means that $[a_k]_{k \in U_N}$ and $[\delta_k]_{k \in U_N}$ are independent, while the second assumption means that $[y_k]_{k \in U_N}$ and $[(a_k, \delta_k)]_{k \in U_N}$ are also independent. It is also important to note that the two samples are assumed to come from the same finite population; a reasonable assumption when considering the CPSS and crowd-sourced samples in a given period. When no response is missing, the population mean may be estimated from the probability sample with the design weights, i.e. by $\widehat{\overline{Y}} = \left(\sum_{k \in s_P} \pi_k^{-1}\right)^{-1} \sum_{k \in s_P} \pi_k^{-1} y_k$. With missing responses, one may instead use the estimator $\widehat{\overline{Y}}^{(I)} = \left(\sum_{k \in s_P} \pi_k^{-1}\right)^{-1} \sum_{k \in s_P} \pi_k^{-1} \left(r_k y_k + (1 - r_k)\widehat{y}_k\right)$, where $\widehat{y}_k$ is a function of the observed responses in the probability sample or both samples.

3 THEORY

In official statistics, single item imputation is the preferred option for imputing items, including deterministic procedures and random procedures (Chen & Haziza 2019). In this work, such a procedure is considered, where $\widehat{y}_k \sim Bernoulli(\widehat{\mu}_k)$ and $\widehat{\mu}_k$ is estimated from the observed responses in the probability sample or both samples.

Semiparametric setup: In this setting, the estimation of μ_k is based on the logistic model $\mu_k = \mu\left(\boldsymbol{\beta}; \boldsymbol{x}_k\right) = e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}} \left(1 + e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}\right)^{-1}$. Then $\widehat{\mu}_k = \mu\left(\widehat{\boldsymbol{\beta}}; \boldsymbol{x}_k\right)$, where $\widehat{\boldsymbol{\beta}}$ may be based on the solution of the following estimating equation,

$$\sum_{k \in U_N} (a_k r_k + z\delta_k) \boldsymbol{x}_k \left(y_k - \mu \left(\widehat{\boldsymbol{\beta}}; \boldsymbol{x}_k \right) \right) = \mathbf{0},$$

where $z = 0$ if only using the responses from the probability sample and $z = 1$ if using the responses from both samples. When estimating $\boldsymbol{\beta}$, it is possible to ignore the sample inclusion probabilities and to pool the responses from both samples, because the units are assumed to be included in the samples at random and the responses are assumed to be missing at random in the probability sample. In the appendix, it is shown that the resulting estimator of the finite population mean is consistent with an asymptotic normal distribution, under general conditions.

Nonparametric setup: For categorical covariates, one may use a random hot-deck solution by partitioning the respondents into classes where the mean response is as homogeneous as possible. Then $\widehat{y}_k$ is obtained by uniformly drawing a donor from the same class. This is equivalent to drawing $\widehat{y}_k$ according to $\textit{Bernoulli}\left(\widehat{\mu}_k\right)$, where $\widehat{\mu}_k$ is the proportion of donors having $y = 1$, within the corresponding class. As in the semiparametric case, $\widehat{\mu}_k$ may be based on the responses from the probability sample alone, or from both samples. When imputing the missing responses one must balance the additional variance with the potential bias. Indeed, the imputation of a missing response is a source of variance, which decreases with the size of the related imputation class. It also generates some bias when the missing response is imputed with a donor having a differing mean response. When building the imputation classes, one may only use the observed covariates among the respondents or also use their responses.

An example of the former strategy is to place two respondents in the same class if they have the same covariates, which is essentially equivalent to the nearest-neighbor imputation when the samples are sufficiently large compared to $|\mathcal{X}|$. An example of the latter strategy (i.e. also using the observed responses) is to build the imputation classes with a regression tree based on recursive partitioning. In general, using the observed responses provides a more flexible way of optimizing the imputation bias-variance trade-off, i.e. one may hope for fewer imputation classes (hence a smaller imputation variance) while keeping the bias small. Recursive partitioning is an attractive solution for building the imputation classes. In the appendix, it is argued that recursive partitioning produces homogeneous classes with a high probability, when the two samples become arbitrarily large while $|\mathcal{X}|$ is fixed. Then the resulting estimator is essentially equivalent to the semiparametric estimator under the saturated model, which is consistent and asymptotically normal under the general conditions given in the appendix. Recursive partitioning presents an advantage (over nearest-neighbor when N is large) if it builds no more than $|\mathcal{X}|-1$ homogeneous imputation classes, and ideally the smallest number of such classes. Yet this may not be the case in some pathological situations, such as the following simple example. Consider two binary covariates, where $\mathcal{X} = \{(0,0),(0,1),(1,0),(1,1)\}$, $E\left[y|\boldsymbol{x} = (0,0)\right] = E\left[y|\boldsymbol{x} = (1,1)\right] = a$ and $E\left[y|\boldsymbol{x} = (0,1)\right] = E\left[y|\boldsymbol{x} = (1,0)\right] = b \neq a$. In this case, it is easily seen that the only trees producing homogeneous classes are the two full trees, which produce four classes, i.e. the maximum number of imputation classes.

Variance estimation: To estimate the variance, a re-sampling procedure is proposed, which draws its inspiration from previous work by Shao & Sitter (1996) and Kim et al. (2020). It is a three-step procedure, where the two samples are drawn with replacement independently in the first step. In the second step, these samples are used to impute the missing responses. Finally, the mean is estimated with the imputed responses. Although the performance of this procedure is studied in the simulations, it must be noted that, in the semiparametric case, the consistency of the resulting variance estimator does not follow from the consistency proof by Kim et al. (2020), because therein $\widehat{\boldsymbol{\beta}}$ is only based on the nonprobability sample instead of using both samples as proposed here. However the extension of this proof is deferred to future work.

4 SIMULATIONS

The simulations are design-based with a single finite population with $N = 10,000$ units and two covariates. Three scenarios are considered where four estimators are compared based on the logistic model, the lasso approach, recursive partitioning and nearest neighbor. The first scenario is ideal, where x_1 and x_2 are independent $Bernoulli(1/2)$ variables and the response is such that $logit\left(E\left[y\left|\boldsymbol{x}\right.\right]\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ with $\beta_0 = 0$, $\beta_1 = 1$ and $\beta_2 = -1$. Both samples are independent simple random samples without replacement with sizes $n_P = 100$ and $n_{NP} = 1,000$, and the responses are missing completely at random with probability 1/2 in the probability sample. The other scenarios are mild deviations from this ideal situation. In the second scenario, the nonprobability sample is drawn in two phases, where the first phase is based on a simple random sample of 1,000 units. In the second phase, a unit with the response y is kept in the sample with the probability $e^{-\tau y}$ with $\tau = 1/10$. Thus the inclusion mechanism is not ignorable. In the third scenario, the response probability is $e^{\tau y}/\left(1 + e^{\tau y}\right)$, where $\tau = 1/10$, i.e. the responses are not missing at random in the probability sample.

To evaluate the performance of the proposed variance estimation procedure, the bootstrap variance is computed and compared to the actual variance in the first scenario. It is based on 100 bootstrap samples and it is only computed for the estimators based on the logistic model, recursive partitioning and nearest neighbor. The bootstrap variance is not computed for the lasso because the procedure is computer intensive. All the estimates are computed in R. For the logistic model, the estimates are obtained by calling the function `glm()` (Marschner 2011). For the lasso and recursive partitioning, the estimates are computed with the packages `glmnnet` (Friedman et al. 2010) and `rpart` (Therneau & Atkinson 2019), respectively. For the nearest neighbor estimator, the imputed response is drawn from a Bernoulli distribution where the probability corresponds to the proportion of respondents (in the probability sample or both samples) having $y = 1$ and the same values for the covariates.

The results appear in Tables 1-4. In the first scenario, the results show that using both samples reduces the bias but slightly increases the variance. This is especially true with the nearest neighbor estimator where the bias is large when only using the probability sample. The MSE increases when going from the logit model, to the lasso, to recursive partitioning and then to the nearest neighbor method, regardless of whether the nonprobability sample is used. Table 2 shows the relative bias of the bootstrap variance, where it can be seen that the bias is non-negligible even if the variance is estimated with the correct order of magnitude (the relative bias is no greater than 30%). Overall the first scenario results demonstrate the advantage of using the nonprobability sample when all the assumptions are met. For the second scenario, Table 3 shows that the bias is increased, when using the nonprobability sample, without surprise since the units are not included at random in this sample. However, the variance is decreased when using the nonprobability, except for the logit model. Overall, the MSE is increased except for the nearest neighbor method. For the third scenario, Table 4 shows that using the nonprobability sample tends to reduce the bias and the variance, except for the slightly increased lasso bias and the slightly increased recursive partitioning variance. Overall these results confirm the advantage of using the nonprobability sample when all the assumptions are satisfied.

5 EMPIRICAL STUDY

In the empirical study, the probability sample is based on the third iteration of the Canadian Perspectives Surveys Series (Statistics Canada 2020); a probability web-panel that ran from June 15 to June 21, 2020. The nonprobability sample is based on a voluntary web survey that ran in the same period. Both surveys contain questions about the willingness to get vaccinated as well as the age, education level, immigration status and whether one has children. The willingness to get vaccinated is based on the question "When a COVID-19 vaccine becomes available, how likely is it that you will choose to get it?", with the answer being one of "Very likely", "Somewhat likely", "Somewhat unlikely" and "Very unlikely", when non-missing. For

Table 1: Performance of the different estimators in the first scenario.

Estimator	Both samples	Relative bias (%)	Variance ($\times 10^{-5}$)	MSE ($\times 10^{-5}$)
Logit	no	0.41	0.90	1.30
	yes	0.37	1.06	1.40
Lasso	no	0.39	1.16	1.53
	yes	0.30	1.19	1.41
Recursive partitioning	no	0.46	1.08	1.60
	yes	0.39	1.40	1.78
Nearest neighbor	no	-10.09	0.85	247.15
	yes	-1.15	0.89	4.07

Table 2: Performance of the bootstrap variance in the first scenario.

Estimator	Both samples	Relative bias of the estimated variance (%)
Logit	no	21.20
	yes	7.56
Recursive partitioning	no	-0.74
	yes	-20.78
Nearest neighbor	no	19.11
	yes	26.19

Table 3: Performance of the different estimators in the second scenario.

Estimator	Both samples	Relative bias (%)	Variance ($\times 10^{-5}$)	MSE ($\times 10^{-5}$)
Logit	no	-0.29	1.32	1.53
	yes	-0.73	1.60	2.91
Lasso	no	-0.44	1.18	1.67
	yes	-0.65	1.00	2.06
Recursive partitioning	no	-0.23	1.23	1.36
	yes	-0.70	1.16	2.36
Nearest neighbor	no	-0.23	118.66	118.79
	yes	-2.28	13.35	26.18

Table 4: Performance of the different estimators in the third scenario.

Estimator	Both samples	Relative bias (%)	Variance ($\times 10^{-5}$)	MSE ($\times 10^{-5}$)
Logit	no	1.93	1.31	10.59
	yes	1.73	1.04	8.57
Lasso	no	1.87	1.32	10.07
	yes	1.88	1.25	10.11
Recursive partitioning	no	1.97	1.10	10.76
	yes	1.77	1.13	8.95
Nearest neighbor	no	2.34	125.82	139.57
	yes	0.95	9.61	11.86

the study, the willingness to get vaccinated is treated as the response. It is coded as a 0-1 variable, which is set to 1 if the answer is "Very likely". The explanatory variables include the age, education level, immigration status and whether one has children, which are all coded as 0-1 variables in this study. The age is set to 1 when the person's age is greater than or equal to 65. The education level is set to 1 if the individual has at least a bachelor's degree. The immigration status is set to 1 if the respondent was born in Canada. Finally whether one has children is coded as a 1 if that is indeed the case.

In the probability sample, the response is set to missing according to an independent *Bernoulli*(1/2). As for the nonprobability sample, it is limited to the respondents where all the variables are non-missing. Both samples are treated as simple random samples. This means that the design weights of the probability sample are ignored, which implies a non-informative sampling design. The relative error is measured with respect to the mean of the probability sample with all the responses. The variance is estimated with the previously described bootstrap procedure.

In Table 5, it can be seen that both the relative error and the variance tend to increase when using the nonprobability sample, for each estimator, unlike what one would expect if all the assumptions were met. This suggests a possible departure from some of these assumptions, which must be checked, such as the inclusion at random of a unit in each sample.

Table 5: Results of the empirical study.

Estimator	Both samples	Relative error (%)	Variance ($\times 10^{-5}$)
Logit	no	-1.11	1.45
	yes	-6.00	2.05
Lasso	no	-0.89	1.46
	yes	-5.95	2.30
Recursive partitioning	no	-1.44	1.84
	yes	-6.39	1.74
Nearest neighbors	no	-1.68	1.19
	yes	-5.95	2.14

6 CONCLUSION

In conclusion, it is beneficial to use the nonprobability sample under the four stated assumptions, including non-informative samples and missing at random responses in the probability sample. However, these assumptions are critical and must be checked. The issue of variance estimation must also be revisited as the proposed procedure may estimate the variance with a non-negligible bias, even when all the assumptions are met.

7 DISCLAIMER

The content of this paper represents the authors' opinions and not necessarily those of Statistics Canada. It describes theoretical methods that may not reflect those currently implemented by the Agency.

REFERENCES

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

Chen, S. & Haziza, D. (2019), ‘Recent developments in dealing with item non-response in surveys: A critical review. international statistical review’, *International Statistical Review* **87**, S192–S218.

Chen, S., Yang, S. & Kim, J.-K. (2020), ‘Nonparametric mass imputation for data integration’, *Journal of Survey Statistics and Methodology* **0**, 1–24.

Chen, Y., Li, P. & Wu, C. (2020), ‘Doubly robust inference with nonprobability survey samples’, *Journal of the American Statistical Association* **115**, 2011–2021.

Friedman, J., Hastie, T. & Tibshirani, R. (2010), ‘Regularization paths for generalized linear models via coordinate descent’, *Journal of Statistical Software* **33**, 1–22.
URL: *https://www.jstatsoft.org/v33/i01/*

Fuller, W. (2009), *Sampling statistics*, Wiley.

Kim, J.-K., Park, S., Chen, Y. & Wu, C. (2020), ‘Combining non-probability and probability survey samples through mass imputation’. arXiv:1812.10694.

Marschner, I. (2011), ‘glm2: Fitting generalized linear models with convergence problems’, *The R Journal* **3**, 12–15.

Rivers, D. (2007), Sampling for web surveys, *in* American Statistical Association, ed., ‘In Proceedings of the Section on Survey Research Methods’, pp. 1–26.

Shao, J. & Sitter, R. R. (1996), ‘Bootstrap for imputed survey data’, *Journal of the American Statistical Association* **91**, 1278–1288.

Statistics Canada (2020), ‘Canadian perspectives survey series 3: Resuming economic and social activities during covid-19’, `https://www150.statcan.gc.ca/n1/daily-quotidien/200708/dq200708a-eng.htm`. posted 08-July-2020.

Therneau, T. & Atkinson, B. (2019), *rpart: Recursive Partitioning and Regression Trees.* R package version 4.1-15.
URL: *https://CRAN.R-project.org/package=rpart*

Yang, S., Kim, J.-K. & Hwang, Y. (2021), ‘Integration of data from probability surveys and big found data for finite population inference using mass imputation’, *Survey Methodology* **47**, 29–58.

A PROOFS

A.1 Consistency of recursive partitioning

The consistency of recursive partitioning has been discussed extensively in previous work including Breiman et al. (1984) and references therein, when dealing with continuous covariates. However in this work, the interest lies with categorical covariates when $|\mathcal{X}|$ is fixed regardless of N. In such cases, it seems intuitive that recursive partitioning will produce homogeneous classes with a high probability when each sample becomes arbitrarily large. Indeed, with this procedure, the tree is built in two phases including a splitting phase followed by a pruning phase. In the splitting phase, the data points are partitioned recursively into disjoint regions, with a greedy procedure to minimize a cost function based on a sum of squares. At each step, a region is further partitioned into subregions according to the variable yielding the largest decrease in the sum of squares, until a stopping criterion is met, e.g. a minimum number of observations in a region or a maximum number of regions. The pruning phase selects a subtree with a cost not exceeding that of the tree. To avoid over-fitting different data points are used for the two phases, through cross-validation. However, in the asymptotic regime where $|\mathcal{X}|$ is bounded and the two samples become arbitrarily large, this cross-validation is likely to have a negligible impact. Thus it is ignored in the following heuristic discussion.

Denote by Π the set of all partitions of $\mathcal{X}$ and by $\mathcal{P} = \left\{\mathcal{R}_{\mathcal{P}1}, \ldots, \mathcal{R}_{\mathcal{P}|\mathcal{P}|}\right\}$ such a partition, where $\mathcal{R}_{\mathcal{P}1}$,...,$\mathcal{R}_{\mathcal{P}|\mathcal{P}|}$ are nonempty, disjoints and such that $\mathcal{X} = \bigcup_{t=1}^{|\mathcal{P}|} \mathcal{R}_{\mathcal{P}t}$. Let Π^* denote the subset of Π, which comprises of all the partitions with constituent subsets that are level sets of the mean response. In other words, $\mathcal{P} \in \Pi^*$ if and only if the mean response is a constant over $\mathcal{R}_{\mathcal{P}t}$ for each $t = 1, \ldots, |\mathcal{P}|$. For $t = 1, \ldots, |\mathcal{P}|$, let $n_{\mathcal{P}t} = \sum_{k\in U_N} I\left(\boldsymbol{x}_k \in \mathcal{R}_{\mathcal{P}t}\right)\left(a_k r_k + z\delta_k\right)$ and define the cost function

$$C(\mathcal{P}) = \sum_{t=1}^{|\mathcal{P}|} \sum_{k\in U_N} I\left(\boldsymbol{x}_k \in \mathcal{R}_{\mathcal{P}t}\right)\left(a_k r_k + z\delta_k\right)\left(y_k - \frac{1}{n_{\mathcal{P}t}} \sum_{k'\in U_N} I\left(\boldsymbol{x}_{k'} \in \mathcal{R}_{\mathcal{P}t}\right)\left(a_{k'} r_{k'} + z\delta_{k'}\right) y_{k'}\right)^2 .$$

Suppose that the following limits hold when $N \to \infty$

$$\frac{1}{\sum_{t=1}^{|\mathcal{P}|} n_{\mathcal{P}t}} \sum_{k\in U_N} \left(a_k r_k + z\delta_k\right) \mu_k \left(1 - \mu_k\right) \xrightarrow{p} \tau^2$$

and

$$\frac{1}{\sum_{t=1}^{|\mathcal{P}|} n_{\mathcal{P}t}} \sum_{t=1}^{|\mathcal{P}|} \sum_{k\in U_N} I\left(\boldsymbol{x}_k \in \mathcal{R}_{\mathcal{P}t}\right)\left(a_k r_k + z\delta_k\right)\left(\mu_k - \frac{1}{n_{\mathcal{P}t}} \sum_{k'\in U_N} I\left(\boldsymbol{x}_{k'} \in \mathcal{R}_{\mathcal{P}t}\right)\left(a_{k'} r_{k'} + z\delta_{k'}\right) \mu_{k'}\right)^2 \xrightarrow{p} v_{\mathcal{P}}^2,$$

for all $\mathcal{P} \in \Pi$, where $v_{\mathcal{P}}^2 > 0$ if and only if $\mathcal{P} \notin \Pi^*$. Then

$$C(\mathcal{P}) = \left(\sum_{t=1}^{|\mathcal{P}|} n_{\mathcal{P}t}\right)\left(\tau^2 + v_{\mathcal{P}}^2 + o_p(1)\right).$$

When $\mathcal{P} \in \Pi^*$, $v_{\mathcal{P}}^2 = 0$ so that $C(\mathcal{P}) = \left(\sum_{t=1}^{|\mathcal{P}|} n_{\mathcal{P}t}\right)\left(\tau^2 + o_p(1)\right)$. When $\mathcal{P} \notin \Pi^*$, $v_{\mathcal{P}}^2 \geq \min_{\mathcal{P}'\notin\Pi^*} v_{\mathcal{P}'}^2 > 0$ so that $C(\mathcal{P}) > \max_{\mathcal{P}'\in\Pi^*} C(\mathcal{P}')$ and $\min_{\mathcal{P}\in\Pi-\Pi^*} C(\mathcal{P}) > \max_{\mathcal{P}\in\Pi^*} C(\mathcal{P})$ with a high probability.

Now suppose that the tree is built through recursive partitioning and that the stopping criterion is based on having no more than $|\mathcal{X}|$ terminal nodes in the growth phase. Then with a high probability as $N \to \infty$, the growth phase is to produce a maximal tree, where the leaves are the singleton subsets of $\mathcal{X}$, i.e. the partition $\mathcal{P}_0 = [\{\boldsymbol{x}\}]_{\boldsymbol{x}\in\mathcal{X}} \in \Pi^*$. The pruning phase is to produce a subtree $\mathcal{P}'$ having a cost no greater than $C(\mathcal{P}_0)$. Then with a high probability,

$$\min_{\mathcal{P}\in\Pi-\Pi^*} C(\mathcal{P}) > \max_{\mathcal{P}\in\Pi^*} C(\mathcal{P}) \geq C(\mathcal{P}_0) \geq C(\mathcal{P}'),$$

which implies that $\mathcal{P}' \in \Pi^*$, i.e. $\mathcal{P}'$ produces homogeneous imputation classes with a high probability.

A.2 Large sample properties

This section proves the consistency and asymptotic normality of $\widehat{\overline{Y}}^{(I)}$, in the semiparametric case, under general conditions that are inspired by Fuller (2009).

Suppose that $E[y|\boldsymbol{x}] = \mu(\boldsymbol{\beta}_0; \boldsymbol{x})$, where $\mu(.; \boldsymbol{x})$ is sufficiently differentiable for each $\boldsymbol{x}$-value. Let $\mu_k = \mu\left(\boldsymbol{\beta}_0; \boldsymbol{x}_k\right)$, $\overline{\mu}_N = N^{-1}\sum_{k\in U_N} \mu_k$ and suppose that $\widehat{\boldsymbol{\beta}}$ is the solution of the following estimating equation for some sufficiently differentiable (with respect to its first argument) score function $\boldsymbol{g}(.; ., .)$.

$$\sum_{k\in U_N} \left(a_k r_k + \delta_k\right) \boldsymbol{g}\left(\widehat{\boldsymbol{\beta}}; \boldsymbol{x}_k, y_k\right) = \mathbf{0}.$$

Let $\widehat{\mu}_k = \mu\left(\widehat{\boldsymbol{\beta}}; \boldsymbol{x}_k\right)$ and $n = |s_P|$ for notational convenience. Also define $\widehat{T} = \sum_{k\in U_N} \pi_k^{-1} a_k y_k$, $\widehat{T}^{(I)} = \sum_{k\in U_N} \pi_k^{-1} a_k \left(r_k y_k + (1-r_k)\widehat{y}_k\right)$ and $\widehat{N} = \sum_{k\in U_N} \pi_k^{-1} a_k$.

Suppose that $\max_{1\le k\le N} \pi_k^{-1}(n/N) = O(1)$ and

$$\begin{bmatrix} \sqrt{n}\left(N^{-1}\sum_{k\in U_N} \pi_k^{-1} a_k \mu_k - \overline{\mu}\right) \\ \\ \sqrt{n}\left(\widehat{N}/N - 1\right) \end{bmatrix} \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}).$$

Further suppose that $n = o_p(s_{NP})$ such that $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = o_p(n^{-1/2})$ and $\max_{1\le k\le N} |\widehat{\mu}_k - \mu_k| = o_p(n^{-1/2})$. Also suppose that

$$\frac{1}{n}\sum_{k\in U_N}\left(\frac{n\pi_k^{-1}}{N}\right)^2 a_k(1-r_k)\mu_k\left(1-\mu_k\right) \xrightarrow{p} \sigma^2.$$

Then, it can be shown that $\widehat{\overline{Y}}^{(I)}$ is consistent with an asymptotic normal distribution. Indeed for $\boldsymbol{\omega} = [\omega_1\ \omega_2]^\top \in I\!R^2$, let

$$\begin{aligned} \Delta_N &= \omega_1\sqrt{n}\left(\frac{\widehat{T}^{(I)}}{N} - \overline{\mu}_N\right) + \omega_2\sqrt{n}\left(\frac{\widehat{N}}{N} - 1\right) \\ &= \underbrace{\omega_1\sqrt{n}\left(\frac{\widehat{T}_y^I}{N} - \frac{1}{N}\sum_{k\in U_N}\pi_k^{-1}a_k\left(r_k\mu_k + (1-r_k)\,\widehat{\mu}_k\right)\right)}_{=\Delta_{N1}} + \underbrace{\frac{\omega_1\sqrt{n}}{N}\sum_{k\in U_N}\pi_k^{-1}a_k\left(1-r_k\right)\left(\widehat{\mu}_k - \mu_k\right) +}_{=o_p(1)} \\ &\quad \underbrace{\omega_1\sqrt{n}\left(\frac{1}{N}\sum_{k\in U_N}\pi_k^{-1}a_k\mu_k - \overline{\mu}\right) + \omega_2\sqrt{n}\left(\frac{\widehat{N}}{N} - 1\right)}_{=\Delta_{N2}}. \end{aligned}$$

Hence

$$\Delta_N = \Delta_{N1} + \boldsymbol{\omega}^\top \underbrace{\begin{bmatrix} \sqrt{n}\left(N^{-1}\sum_{k\in U_N} \pi_k^{-1} a_k \mu_k - \overline{\mu}\right) \\ \\ \sqrt{n}\left(\widehat{N}/N - 1\right) \end{bmatrix}}_{=\boldsymbol{W}_N} + o_p(1),$$

Next observe that

$$
\begin{aligned}
E\left[e^{\jmath\Delta_{N1}}\middle|\left[\begin{pmatrix} a_k \\ \delta_k \\ y_k \\ r_k \end{pmatrix}\right]_{k\in U_N}\right] &= E\left[\exp\left(\frac{\jmath\omega_1}{\sqrt{n}}\sum_{k\in U_N}\left(\frac{n\pi_k^{-1}}{N}\right)a_k(1-r_k)\left(\widehat{y}_k-\widehat{\mu}_k\right)\right)\middle|\left[\begin{pmatrix} a_k \\ \delta_k \\ y_k \\ r_k \end{pmatrix}\right]_{k\in U_N}\right] \\
&= \prod_{k\in U_N}\Big[\left(1-\widehat{\mu}_k\right)e^{-\jmath\omega_1(n\pi_k^{-1}/N)(1-r_k)\widehat{\mu}_k/\sqrt{n}} + \\
&\quad \widehat{\mu}_k e^{\jmath\omega_1(n\pi_k^{-1}/N)(1-r_k)(1-\widehat{\mu}_k)/\sqrt{n}}\Big]^{a_k} \\
&= \prod_{k\in U_N}\left[1-\frac{\omega_1^2}{2n}\left(\frac{n\pi_k^{-1}}{N}\right)^2(1-r_k)\widehat{\mu}_k\left(1-\widehat{\mu}_k\right)+O_p\left(n^{-3/2}\right)\right]^{a_k} \\
&= \exp\left(-\frac{\omega_1^2}{2}\left(\frac{1}{n}\sum_{k\in U_N}\left(\frac{n\pi_k^{-1}}{N}\right)^2 a_k(1-r_k)\widehat{\mu}_k\left(1-\widehat{\mu}_k\right)\right)+O_p\left(n^{-1/2}\right)\right) \\
&= \exp\Bigg(-\frac{\omega_1^2}{2}\left(\frac{1}{n}\sum_{k\in U_N}\left(\frac{n\pi_k^{-1}}{N}\right)^2 a_k(1-r_k)\mu_k\left(1-\mu_k\right)\right)+o_p(n^{-1/2})+ \\
&\quad O_p\left(n^{-1/2}\right)\Bigg) \\
&= e^{-\sigma^2\omega_1^2/2+o_p(1)}.
\end{aligned}
$$

Therefore

$$
\begin{aligned}
E\left[e^{\jmath\Delta_N}\right] &= E\left[E\left[e^{\jmath\Delta_{N1}}\middle|\left[\begin{pmatrix} a_k \\ \delta_k \\ y_k \\ r_k \end{pmatrix}\right]_{k\in U_N}\right]e^{\jmath\omega^\top\boldsymbol{W}_N+o_p(1)}\right] \\
&= E\left[e^{-\sigma^2\omega_1^2/2+\jmath\omega^\top\boldsymbol{W}_N+o_p(1)}\right] \\
&= e^{-\sigma^2\omega_1^2/2}E\left[e^{\jmath\omega^\top\boldsymbol{W}_N}\right]+o(1) \\
&= e^{-\sigma^2\omega_1^2/2}\left(e^{-\boldsymbol{\omega}^\top\boldsymbol{\Sigma}\boldsymbol{\omega}/2}+o(1)\right)+o(1)
\end{aligned}
$$

and $E\left[e^{\jmath\Delta_N}\right]\to e^{-\left(\sigma^2\omega_1^2+\boldsymbol{\omega}^\top\boldsymbol{\Sigma}\boldsymbol{\omega}\right)/2}$. This means that $\sqrt{n}\left(\widehat{T}^{(I)}/N-\overline{\mu}_N\right)$ and $\sqrt{n}\left(\widehat{N}/N-1\right)$ have a limiting joint distribution that is normal. Then Slutsky's theorem implies that $\sqrt{n}\left(\widehat{\overline{Y}}^{(I)}-\overline{\mu}_N\right)$ has an asymptotic normal distribution with zero mean.