A semiparametric propensity score weighting method to adjust for nonresponse using multivariate auxiliary information

Da Silva, Damião N.
University of Southampton (and Universidade Federal do Rio Grande do Norte, Brasil), Southampton Statistical Sciences Research Institute (S3RI) (and Departamento de Estatística)
Southampton, Hampshire, United Kingdom, SO17 1BJ
E-mail: d.dasilva@soton.ac.uk

Skinner, Chris
University of Southampton, Southampton Statistical Sciences Research Institute (S3RI)
Southampton, Hampshire, United Kingdom, SO17 1BJ
E-mail: C.J.Skinner@soton.ac.uk

1. Introduction

Weighting by propensity scores is a widely used adjustment procedure to deal with unit nonresponse in surveys. In this procedure, weights are defined in terms of the probability that the respondent would respond to the survey if sampled. This procedure can be motivated in a two–phase sampling framework given by the sampling design and the mechanism generating the responses (Oh and Scheuren 1983). Under this framework, if the response probabilities could be estimated without error, then the resulting Horvitz–Thompson estimator would be unbiased estimator for the corresponding finite population parameter. Hence, if the estimated response probabilities were reasonable approximations to the respective true probabilities, the propensity score adjusted estimator is expected to present "small" nonresponse bias.

The estimation of these response probabilities is usually implemented by fitting parametric regression models for binary variables (see, e.g., Laaksonen 2006). Although parametric response modeling has many advantages, the resulting adjusted estimators can be seriously biased under misspecification of the underlying models. An alternative is to estimate those probabilities by nonparametric regression. Giommi (1984) introduced this idea with the use of kernel smoothing. A detailed account of the statistical properties for the corresponding adjusted estimator by this approach is given in Da Silva and Opsomer (2006). Da Silva and Opsomer (2009) investigated the same properties by considering an extension to Giommi's estimator, based on the estimation of the probabilities of response with local polynomial regression. In both cases, the adjusted estimator by these two methods is unbiased (for large samples) and consistent under the two-phase framework of Oh and Scheuren (1983) without having to assume the form of the regression curve is known.

One possibly restrictive limitation for many sampling surveys to directly implement these nonparametric procedures for nonresponse is their requirement to depend on only one continuous auxiliary variable. In this article, we discuss a semiparametric kernel–based method to estimate response propensities that still require at least one auxiliary variable to be continuous. However, the method allows categorical variables that affect nonresponse to be taken into account.

2. The proposed method

Consider a population of elements denoted by $U = \{1, 2, ..., N\}$ with N finite. Let y_i denote the value of a study variable y and $\boldsymbol{x}_i = (\boldsymbol{z}_i^T, \boldsymbol{t}_i^T)^T$ a $k \times 1$ vector of values of k auxiliary variables for the i–th element in U, where $\boldsymbol{z}_i$ is $k_1 \times 1$ and $\boldsymbol{t}_i$ is $(k - k_1) \times 1$. Let also $\boldsymbol{y}_U = (y_1, y_2, ..., y_N)^T$ and the $\boldsymbol{X}_U = [\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N]^T$. Suppose we are interested in the estimation of the population mean of $\bar{y}_U = N^{-1} \sum_{i=1}^{N} y_i$ using the information in a sample s of size n selected from U. Let

$\boldsymbol{I}_U = (I_1, I_2, ..., I_N)^T$, where I_i is an indicator variable that element i is included in the sample s. Suppose s is chosen by a probability sampling design $p(\cdot)$ with the property

$$\Pr(\boldsymbol{I}_U = \boldsymbol{i}_U \,|\, \boldsymbol{X}_U, \boldsymbol{y}_U) = \Pr(\boldsymbol{I}_U = \boldsymbol{i}_U \,|\, \boldsymbol{X}_U),$$

for all possible realizations $\boldsymbol{i}_U$ of $\boldsymbol{I}_U$ and all possible values of $\boldsymbol{X}_U$ and $\boldsymbol{y}_U$. The first order inclusion probabilities are denoted $\pi_i = \Pr(I_i = 1 \,|\, \boldsymbol{X}_U)$ and in the event the information to be collected in the elements of s can be fully observed, we could estimate $\bar{y}_U$ with, for instance, the Hájek estimator

$$(1) \quad \bar{y}_H = \frac{\sum_{i\in s} \pi_i^{-1} y_i}{\sum_{i\in s} \pi_i^{-1}}.$$

Now suppose that there is nonresponse and y_i and only observed for $r = \{i \in s : R_i = 1\}$, where R_i is an indicator variable that element $i \in U$ responds to the survey. In this case, *naive* adjustments that apply the base sampling weights to the respondents result in biased estimators for $\bar{y}_U$ when y is correlated with nonresponse. To attempt to reduce this bias, define the propensity scores as

$$\phi_i = \Pr\{R_i = 1 \,|\, \boldsymbol{X}_U\}, \quad i = 1, 2, ..., N.$$

The adjustment by the propensity scores requires the estimation of the response probabilities ϕ_i for all respondents. The estimated probabilities, denoted by $\{\widehat{\phi}_i : i \in r\}$, yield the adjusted estimator

$$(2) \quad \bar{y}_{H,\widehat{\phi}} = \frac{\sum_{i\in r} \pi_i^{-1} \widehat{\phi}_i^{-1} y_i}{\sum_{i\in r} \pi_i^{-1} \widehat{\phi}_i^{-1}},$$

which is nearly unbiased when the response propensities are estimated without error ($\widehat{\phi}_i = \phi_i$).

A successful application of the propensity score procedure depends on how well the response propensities can be estimated. This task is intrinsically related to one's ability in postulating a sensible working model for the true propensity scores. The model of interest in this article assumes that

(R.1) For all possible realizations $\boldsymbol{r}_U = (r_1, r_2, ..., r_N)^T$ of $\boldsymbol{R}_U = (R_1, R_2, ..., R_N)^T$ and all $\boldsymbol{y}_U$ and $\boldsymbol{X}_U$,

$$\Pr\{\boldsymbol{R}_U = \boldsymbol{r}_U \,|\, \boldsymbol{y}_U, \boldsymbol{X}_U\} = \prod_{i=1}^{N} \Pr\{R_i = r_i \,|\, \boldsymbol{X}_U\};$$

(R.2) For a given known and monotone link function $g(\cdot)$,

$$\phi_i \equiv \Pr\{R_i = 1 \,|\, \boldsymbol{X}_U\} = g^{-1}\big(\boldsymbol{z}_i^T \boldsymbol{\beta} + \gamma(\boldsymbol{t}_i)\big), \quad \text{for all } i \in U \text{ and all } \boldsymbol{X}_U,$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector of parameters and $\gamma : \mathbb{R}^{k-k_1} \to \mathbb{R}$ is an unknown smooth function.

Condition (R.1) assumes the nonresponse process follows a *missing–at–random* response mechanism, as the ϕ_i do not depend on the study variable given the values of the auxiliary variables are fixed. This condition assumes also the response indicators are conditionally independent given the auxiliary variables. Condition (R.2) postulates a model for the response propensities that is a special case of the semiparametric regression model proposed by Severini and Staniswalis (1994). Categorical variables are modeled through the term $\boldsymbol{\beta}^T \boldsymbol{z}$ and continuous auxiliary variables with possibly nonlinear effects on the ϕ_i are modeled nonparametrically through the function $\gamma(\cdot)$, which allows also taking into account curvature in the link scale even if the link function is incorrectly specified.

The estimation of the response propensities defined in (R.2) requires estimating the vector of parameters $\boldsymbol{\beta}$ and the function $\gamma(\cdot)$. If a census could have been undertaken over the population U, estimates $\widehat{\boldsymbol{\beta}}$ and $\widehat{\gamma}(\cdot)$ could be obtained by working with the population–based log–likelihood function

$$\ell \equiv \ln \Pr\{R_1 = r_1, ..., R_N = r_N \,|\, \boldsymbol{X}_U\} = \sum_{i=1}^{N} \big[r_i \ln \phi_i + (1 - r_i) \ln (1 - \phi_i)\big],$$

where $\phi_i = g^{-1}\big(\boldsymbol{z}_i^T\boldsymbol{\beta} + \gamma\big(\boldsymbol{t}_i\big)\big)$. However, in our present context, the only components of $\boldsymbol{R}_U$ that are observed are $\{R_i : i \in s\}$. Hence, we suggest estimating $\boldsymbol{\beta}$ and $\gamma(\cdot)$ as the $\widehat{\boldsymbol{\beta}}$ and $\widehat{\gamma}(\cdot)$ that maximize the design–weighted log-likelihood

$$(3) \quad \widehat{\ell} \equiv \sum_{i\in s} \pi_i^{-1}\big\{r_i \ln\big[g^{-1}\big(\boldsymbol{z}_i^T\boldsymbol{\beta} + \gamma\big(\boldsymbol{t}_i\big)\big)\big] + (1-r_i)\ln\big[1 - g^{-1}\big(\boldsymbol{z}_i^T\boldsymbol{\beta} + \gamma\big(\boldsymbol{t}_i\big)\big)\big]\big\},$$

which is a design–unbiased estimator of ℓ. The maximizers of (3) can be obtained by applying firstly local likelihood estimation to estimate the γ term, for fixed $\boldsymbol{\beta}$, and then estimating $\boldsymbol{\beta}$ using the gamma estimated previously. This requires a two–step computing algorithm, such as the one proposed by Severini and Staniswalis (1994), that needs to be iterated until some convergence criteria is met. An adaptation of this algorithm to our setting is as follows:

Step 1: For each $\boldsymbol{t}$ and $\boldsymbol{\beta}$ fixed, obtain the estimate $\widehat{\gamma}_\beta \equiv \widehat{\gamma}_\beta(\boldsymbol{t})$ as the value γ_β that solves the equation

$$\frac{\partial}{\partial\gamma_\beta}\Bigg\{\sum_{i\in s}\pi_i^{-1}W\bigg(\frac{\boldsymbol{t}_i - \boldsymbol{t}}{b}\bigg)\big\{r_i\ln\big[g^{-1}\big(\boldsymbol{z}_i^T\boldsymbol{\beta}+\gamma_\beta\big)\big] + (1-r_i)\ln\big[1-g^{-1}\big(\boldsymbol{z}_i^T\boldsymbol{\beta}+\gamma_\beta\big)\big]\big\}\Bigg\} =$$
$$\sum_{i\in s}\pi_i^{-1}W\bigg(\frac{\boldsymbol{t}_i - \boldsymbol{t}}{b}\bigg)\big[r_i - g^{-1}\big(\boldsymbol{z}_i^T\boldsymbol{\beta}+\gamma_\beta\big)\big] = 0,$$

where $W(\cdot)$ is a kernel function on $\mathbb{R}^{k-k_1}$ and b is its bandwidth parameter. Also, obtain the estimate $\widehat{\boldsymbol{\gamma}}' \equiv (\partial/\partial\boldsymbol{\beta})\widehat{\gamma}(\boldsymbol{t})$ as the vector $\boldsymbol{\gamma}'_\beta$ which is the solution of the system

$$\frac{\partial}{\partial\gamma_\beta}\Bigg\{\sum_{i\in s}\pi_i^{-1}W\bigg(\frac{\boldsymbol{t}_i - \boldsymbol{t}}{b}\bigg)\big[r_i - g^{-1}\big(\boldsymbol{z}_i^T\boldsymbol{\beta}+\widehat{\gamma}_\beta\big)\big]\big\}\Bigg\} =$$
$$\sum_{i\in s}\pi_i^{-1}W\bigg(\frac{\boldsymbol{t}_i - \boldsymbol{t}}{b}\bigg)g^{-1}\big(\boldsymbol{z}_i^T\boldsymbol{\beta}+\widehat{\gamma}_\beta\big)\big[1-g^{-1}\big(\boldsymbol{z}_i^T\boldsymbol{\beta}+\widehat{\gamma}_\beta\big)\big]\big(\boldsymbol{z}_i + \widehat{\boldsymbol{\gamma}}'\big)\big\} = \boldsymbol{0}.$$

Step 2: Given the values $\widehat{\gamma}_\beta \equiv \widehat{\gamma}_\beta(\boldsymbol{t})$ and $\widehat{\boldsymbol{\gamma}}' \equiv (\partial/\partial\boldsymbol{\beta})\widehat{\gamma}(\boldsymbol{t})$ obtained in Step 1, compute the estimate $\widehat{\boldsymbol{\beta}}$ as the value for the vector $\boldsymbol{\beta}$ that solves the system

$$\frac{\partial}{\partial\boldsymbol{\beta}}\Bigg\{\sum_{i\in s}\pi_i^{-1}\big\{r_i\ln\big[g^{-1}\big(\boldsymbol{z}_i^T\boldsymbol{\beta}+\widehat{\gamma}_\beta\big)\big] + (1-r_i)\ln\big[1-g^{-1}\big(\boldsymbol{z}_i^T\boldsymbol{\beta}+\widehat{\gamma}_\beta\big)\big]\big\}\Bigg\} =$$
$$\sum_{i\in s}\pi_i^{-1}\big[r_i - g^{-1}\big(\boldsymbol{z}_i^T\boldsymbol{\beta}+\widehat{\gamma}_\beta\big)\big]\big(\boldsymbol{z}_i + \widehat{\boldsymbol{\gamma}}'\big) = \boldsymbol{0}.$$

Steps 1 and 2 are repeated until convergence and, after this, the final estimate of $\gamma(\cdot)$ is taken to be $\widehat{\gamma}_{\widehat{\beta}}(\cdot)$.

3. Simulation study

We now demonstrate statistical properties of the semiparametric adjustment described in Section 2 and compare this with the adjustment obtained by fitting a logistic regression model to the response propensities. It is of particular interest to address how well the the semiparametric adjustment is capable of correcting for misspecification of the linear predictor used in the estimation of the propensity scores. We considered a population of $N = 4{,}000$ elements and the following variables

$$X_1 \sim \text{Bernouilli}(1/2), \;\; X_2 = 1 + I\big(U \le 1/2\big) \;\; [\text{where } U \sim \text{U}(0,1)], \quad T \sim \text{U}(0,1),$$

$$Y_1 \sim N(100 + 6X_1 + 10X_2, 16), \quad Y_2 \sim N(200 + 100T, 25),$$

$$Y_3 \;\sim N(400 + 20X_1 - 25X_2 + 100T, 81) \;\; \text{and} \;\; Y_4 = 100I(Y_3 \le \text{Median}(Y_3))$$

were generated for each element by *i.i.d.* sampling. Keeping those fixed, $B = 5{,}000$ simple random samples (without replacement) were taken from the population. For each sample, one population response indicator vector was generated under three response mechanisms (I, II and III) by *i.i.d.* sampling from Bernouilli distribution. The response propensity functions of these mechanisms are

$$\begin{aligned}
\phi_I &\equiv \Pr(R=1\,|\,X_1=x_1, X_2=x_2, T=t) = \text{logit}^{-1}(\beta_1 x_1 + \beta_2 x_2 + \beta_0 + \beta_3 t) \\
\phi_{II} &\equiv \Pr(R=1\,|\,X_1=x_1, X_2=x_2, T=t) = \text{logit}^{-1}(\beta_1 x_1 + \beta_2 x_2 + \gamma(t)), \quad \text{where} \\
&\quad \gamma(t) = \beta_4 + \beta_5(t-0.2) + \beta_6\big[1 + (t-0.2)^2/0.05\big]^{-1} \\
\phi_{III} &\equiv \Pr(R=1\,|\,X_1=x_1, X_2=x_2, T=t) = 1 - \exp\big[-\exp\big(\beta_1 x_1 + \beta_2 x_2 + \gamma(t)\big)\big].
\end{aligned}$$

Response mechanisms I and II use the logit link function and have the same linear effect of X_1 and of X_2 in the link scale. However, mechanism II contains a nonlinear effect of T in contrast to mechanism I which has a linear effect of that variable. Mechanism III differs from II only by the change of the logit by the complementary log–log link. The constants β_0,..., β_6 are such that ϕ_I, ϕ_{II} and ϕ_{III} yield on average 20%, 45% and 35% nonresponse rates. For each set of respondents of a sample, the four methods were applied to adjust the Hájek estimator (2):

- True propensities: adjustment defined by applying the true response propensity scores ($\widehat{\phi} = \phi$).
- Semiparametric: $\widehat{\phi}$ is obtained using the semiparametric algorithm described in Section 2, with a fixed bandwidth parameter $b = 0.35$.
- Logistic: logistic regression with linear predictor that considers linear effects of x and t.
- Respondent mean: The naive adjustment by considering $\widehat{\phi} = c$, where c is a constant in $(0, 1]$.

Table 1: Monte Carlo properties of adjusted estimators for the population mean of four variables based on 5,000 SRS of size 400 from the population of 4,000 elements and 5,000 response indicators under mechanism I

Variable	Method	Mean	Bias	Variance	MSE	RB (%)	RMSE
Y_1	True propensities	118.1	-0.0	0.84	0.84	2.6	1.00
	Semiparametric	118.1	-0.0	0.81	0.81	2.6	0.95
	Logistic	118.1	-0.0	0.82	0.82	2.8	0.97
	Respondent mean	118.8	0.6	0.80	1.16	67.6	1.37
Y_2	True propensities	249.9	-0.0	4.63	4.63	1.2	1.00
	Semiparametric	250.8	0.8	3.85	4.56	42.9	0.98
	Logistic	249.9	-0.0	3.86	3.86	2.3	0.83
	Respondent mean	254.1	4.1	4.10	21.16	203.9	4.57
Y_3	True propensities	423.1	-0.1	23.16	23.16	1.6	1.00
	Semiparametric	424.0	0.9	21.95	22.69	18.4	0.98
	Logistic	423.1	-0.1	22.30	22.31	2.1	0.96
	Respondent mean	427.2	4.1	21.43	38.02	88.0	1.64
Y_4	True propensities	50.0	0.0	7.64	7.64	1.4	1.00
	Semiparametric	49.6	-0.4	7.36	7.52	14.6	0.98
	Logistic	50.1	0.1	7.45	7.46	1.9	0.98
	Respondent mean	48.0	-2.0	7.18	11.12	74.2	1.46

The Monte Carlo results for the estimation of the population mean of Y_1, Y_2, Y_3 and Y_4 are given in Tables 1–3. In what follows, the relative bias (RB) is defined as the ratio between the absolute bias

and the standard deviation. The relative mean square error (RMSE) is the MSE of the estimator divided by the MSE of the adjusted estimator by the true propensity scores.

Table 2: Monte Carlo properties of adjusted estimators for the population mean of four variables based on 5,000 SRS of size 400 from the population of 4,000 elements and 5,000 response indicators under mechanism II

Variable	Method	Mean	Bias	Variance	MSE	RB (%)	RMSE
Y_1	True propensities	118.2	0.0	2.93	2.93	1.0	1.00
	Semiparametric	118.2	0.0	2.09	2.09	1.8	0.71
	Logistic	118.3	0.1	2.97	2.98	6.7	1.02
	Respondent mean	119.1	1.0	1.24	2.17	86.7	0.74
Y_2	True propensities	250.1	0.2	16.18	16.22	4.7	1.00
	Semiparametric	251.5	1.6	8.80	11.33	53.7	0.70
	Logistic	247.1	-2.8	13.93	21.92	75.7	1.35
	Respondent mean	264.4	14.5	5.24	214.48	631.7	13.23
Y_3	True propensities	423.3	0.1	71.93	71.95	1.5	1.00
	Semiparametric	424.1	1.0	52.97	53.90	13.2	0.75
	Logistic	419.2	-4.0	74.18	89.85	46.0	1.25
	Respondent mean	437.6	14.4	30.47	238.89	261.5	3.32
Y_4	True propensities	49.9	-0.1	25.90	25.90	1.6	1.00
	Semiparametric	49.3	-0.7	16.83	17.30	16.8	0.67
	Logistic	51.6	1.6	22.23	24.87	34.4	0.96
	Respondent mean	42.8	-7.2	10.19	61.42	224.2	2.37

Table 3: Monte Carlo properties of adjusted estimators for the population mean of four variables based on 5,000 SRS of size 400 from the population of 4,000 elements and 5,000 response indicators under mechanism III

Variable	Method	Mean	Bias	Variance	MSE	RB (%)	RMSE
Y_1	True propensities	118.2	0.0	2.55	2.55	1.6	1.00
	Semiparametric	118.0	-0.1	2.14	2.16	9.4	0.85
	Logistic	118.0	-0.1	5.65	5.67	5.3	2.22
	Respondent mean	119.0	0.8	0.99	1.62	79.6	0.63
Y_2	True propensities	250.1	0.1	15.15	15.16	3.4	1.00
	Semiparametric	250.6	0.6	10.02	10.40	19.5	0.69
	Logistic	240.7	-9.3	26.63	112.34	179.4	7.41
	Respondent mean	264.0	14.0	4.47	201.38	663.9	13.28
Y_3	True propensities	423.4	0.2	66.01	66.06	2.7	1.00
	Semiparametric	423.2	0.0	55.00	55.00	0.2	0.83
	Logistic	412.2	-11.0	135.33	255.91	94.4	3.87
	Respondent mean	437.3	14.2	25.46	226.32	280.9	3.43
Y_4	True propensities	49.9	-0.1	22.60	22.62	2.9	1.00
	Semiparametric	49.7	-0.3	16.33	16.40	6.3	0.72
	Logistic	54.9	4.9	39.61	63.26	77.3	2.80
	Respondent mean	43.0	-7.0	8.43	56.97	239.9	2.52

4. Discussion

In this article, we considered an adjustment procedure for unit nonresponse by a propensity score weighting method that uses the semiparametric regression model proposed by Severini and Staniswalis (1994). This model allows the estimation of response propensities by taking into account categorical and continuous auxiliary variables. Properties of this method were investigated in a Monte Carlo study which considered also the respondent mean and the adjusted estimators with the true propensity scores and with a logistic regression fit.

- Under the correct model for the logistic adjustment (Response mechanism I), this method outperforms the semiparametric method in terms of bias reduction and efficiency as well. However, the performances of the semiparametric method are not much worse.
- For response mechanisms II and III, the model underlying the logistic method is incorrect. In these cases, the semiparametric method has better performances than the logistic method. The former method yield smaller biases and smaller mean square errors. The superior performance of the semiparametric method under mechanism III demonstrates its potential to adjust for nonresponse under misspecification of the link function.
- The adjustment by true propensities is unbiased for all scenarios, as expected. However, this method is not as efficient as the semiparametric and not as the logistic, when the underlying response mechanism for this was correctly specified. On all scenarios, the respondent mean is highly biased illustrating the danger of not adjusting for the nonresponse.

These results indicate the semiparametric method adopted in this article can be a valuable addition to the class of propensity weighting adjustment methods. However, further investigations are needed to make the method more useful in practice. One issue is how to select properly the bandwidth parameter. Moreover, the implementation and the performances of the method should be investigated in the presence of interactions between the categorical and the continuous auxiliary variables.

REFERENCES (RÉFERENCES)

Da Silva, D. N. and J. D. Opsomer (2006). A kernel smoothing method of adjusting for unit non-response in sample surveys. *The Canadian Journal of Statistics 34*(4), 563–579.

Da Silva, D. N. and J. D. Opsomer (2009). Nonparametric propensity weighting for survey nonresponse through local polynomial regression. *Survey Methodology 35*(2), 165–176.

Giommi, A. (1984). A simple method for estimating individual response probabilities in sampling from finite populations. *Metron 42*(4), 185–200.

Laaksonen, S. (2005–2006). Does the choice of link function matter in response propensity modelling? *Model Assisted and Applications 1*(2), 95–100.

Oh, H. L. and F. J. Scheuren (1983). Weighting adjustments for unit non-response. In W. G. Madow, I. Olkin, and D. B. Rubin (Eds.), *Incomplete data in sample surveys (Vol. 2): Theory and bibliographies*, pp. 143–184. Academic Press (New York; London).

Severini, T. A. and J. G. Staniswalis (1994). Quasi-likelihood estimation in semiparametric models. *Journal of the American Statistical Association 89*, 501–511.