A cautionary tale on instrumental calibration for the treatment of nonignorable unit nonresponse in surveys

Éric Lesage[*], David Haziza[†] and Xavier D'Haultfœuille [‡]

March 12, 2018

Abstract

Response rates have been steadily declining over the last decades, making survey estimates vulnerable to nonresponse bias. To reduce the potential bias, two weighting approaches are commonly used in National Statistical Offices: the one-step and the two-step approaches. In this paper, we focus on the one-step approach, whereby the design weights are modified in a single step with two simultaneous goals in mind: reduce the nonresponse bias and ensure the consistency between survey estimates and known population totals. In particular, we examine the properties of instrumental calibration, a special case of the one-step approach that has received a lot of attention in the literature in recent years. Despite the rich literature on the topic, there remain some important gaps that this paper aims to fill. First, we give a set of sufficient conditions required for establishing the consistency of instrumental calibration estimators. Also, we show that the latter may suffer from a large bias when some of these conditions are violated. Results from a simulation study support our findings.

Key words: Bias amplification, calibration variables, instrumental calibration, nonresponse bias, unit nonresponse, variance amplification.

[*]INSEE, Paris, France
[†]Department of mathematics and statistics, Université de Montréal, Montreal, Canada
[‡]CREST-ENSAE, Paris, France

1 Introduction

Response rates have been steadily declining over the last decades, making survey estimates vulnerable to nonresponse bias. To reduce the potential bias, two weighting approaches may be used: the one-step and the two-step approaches. The latter is commonly used in National Statistical Offices. It can be described as follows: in the first step, the design (or basic) weights are multiplied by the inverse of the estimated response probabilities. In the second step, the weights obtained in the first step are further modified so that survey weighted estimates agree with known population totals. This step is often referred to as calibration. In the first step, survey statisticians aim at reducing the nonresponse bias. Key to achieving an efficient bias reduction is the availability of fully observed variables related to both the probability of response and the survey variables. The estimated response probabilities are obtained by fitting a parametric or a nonparametric model. A common procedure consists of first dividing the respondents and nonrespondents into weighting classes and adjusting the design weights of respondents in a given class by the inverse of the response rate within the same class; see, for example, Eltinge and Yansaneh (1997) and Little (1986). Calibration procedures require variables that are observed on the respondents and whose population total is available from external sources such as the census. Commonly used calibration procedures include post-stratification and generalized raking; see Deville and Särndal (1992) and Deville et al. (1993).

An alternative weighting approach that has gained in popularity in recent years, is the so-called one-step approach, whereby the design weights are modified in a single step with two simultaneous goals in mind: reduce the nonresponse bias and ensure the consistency between survey estimates and known population totals; e.g., see Särndal and Lundström (2005). Unlike the two-step approach, explicit estimation of the response probabilities is not required. We focus on instrumental calibration, also called generalized calibration, a special version of the one-step approach, that has recently received a lot of attention in the literature; see Deville (2002), Sautory (2003), Kott (2006, 2009), Chang and Kott (2008), Kott and Chang (2010) and Kott and Liao (2012), among others. Instrumental calibration permits the use

of variables that are observed only on the respondents. Although it is not possible to test whether or not the data are missing at random (Molenberghs et al., 2008), instrumental calibration may prove useful if it is suspected that we are in the presence of nonignorable nonresponse (Deville, 1998, 2002; Kott and Chang, 2010).

Despite the rich literature discussing instrumental calibration, there remain some important gaps that we aim to fill in this paper. We start by giving a set of sufficient conditions for the consistency of the instrumental calibration estimator. We also show that the latter can nevertheless display a large asymptotic variance. Finally, we show that the instrumental calibration estimator may suffer from bias amplification when some of the conditions for consistency are violated. The terminology bias amplification was coined by Pearl (2010) in the context of causal inference; see also Bhattacharya and Vogt (2007), Myers et al. (2011) and Wooldridge (2016) for a related literature.

This paper is not the first to criticize the one-step approach. Kott and Liao (2015) and Haziza and Lesage (2016) argue that, unlike the one-step approach, the two-step approach makes it possible to assess separately the effect of weighting for nonresponse and that of weighting for calibration purposes. Also, in the case of nonignorable nonresponse, Kott and Liao (2012) advise against this one-step approach. We add to these papers by showing that the behaviour of the instrumental calibration estimator is highly sensitive to the validity of some key conditions. The latter were first examined in D'Haultfœuille (2010); see also Wang et al. (2014).

This paper is organized as follows: in Section 2, we introduce the notation and present the theoretical set-up. In Section 3, we lay out the conditions required for establishing the consistency of instrumental calibration estimators. In Section 4, we examine their properties when the conditions are violated. Variance estimation based on the reverse framework (Shao and Steel, 1999) is discussed in Section 5. The results of an empirical investigation, assessing the performance of several estimators in terms of bias and efficiency, are presented in Section 6. We make some final remarks in Section 7. The technical details are relegated to the

Appendix.

2 Theoretical set-up

Consider a finite population P of size N. We are interested in estimating the population total, $t_y = \sum_{k \in P} Y_k$, of a survey variable Y. A sample S, of size n, is selected from P according to a given sampling design $p(S)$ with first-order inclusion probabilities π_k and second-order inclusion probabilities π_{kl}, $k \neq l$. In the absence of nonresponse, a design-unbiased estimator of t_y is the expansion estimator

$$\widehat{t}_\pi = \sum_{k \in S} d_k Y_k,$$

where $d_k = 1/\pi_k$ denotes the design weight attached to unit k. In the presence of unit nonresponse, only a subset S_r of S is observed, which makes $\widehat{t}_\pi$ impossible to compute. In this case, a naive estimator of t_y is the unadjusted estimator

$$\widehat{t}_{un} = \widehat{N}_\pi \frac{\sum_{k \in S} d_k R_k Y_k}{\sum_{k \in S} d_k R_k}, \tag{1}$$

where $\widehat{N}_\pi = \sum_{k \in S} d_k$ denotes the estimated population size and R_k is a response indicator attached to unit k such that $R_k = 1$ if unit k is a respondent and $R_k = 0$, otherwise. The unadjusted estimator (1) is not consistent unless the data are Missing Completely At Random (Rubin, 1976).

To define a nonresponse adjusted estimator of t_y, we assume that a vector of calibration variables $\mathbf{X}$ is observed for $k \in S_r$ and that the corresponding vector of population totals, $\mathbf{t_x} = \sum_{k \in P} \mathbf{X}_k$, is available from an external source. The X-variables are called instrumental variables. Although we have assumed that the instrumental variables are observed for the respondents and that their population total is known, note that any variable observed for all the sample units (respondents and nonrespondents) may also play the role of instrumental variable. In addition, we assume that a vector of variables $\mathbf{Z}$, with $\dim(\mathbf{Z}) \leq \dim(\mathbf{X})$, is available for $k \in S_r$. Neither the vector of population totals, $\mathbf{t_z} = \sum_{k \in P} \mathbf{Z}_k$, nor the complete data estimator, $\widehat{\mathbf{t}}_{\mathbf{z},\pi} = \sum_{k \in S} d_k \mathbf{Z}_k$, is assumed to be available. The Z-variables are called response model variables (Kott and Liao, 2017) and are assumed to be related to the

probability of response. We assume that the first component of both the $\mathbf{X}$-vector and the $\mathbf{Z}$-vector is equal to one for all the population units.

Our terminology is consistent to what has been used in the econometric literature as the X-variables satisfy exclusion restrictions (see Equations (9) and (10) below); see, e.g., D'Haultfœuille (2010). However, it is different from the terminology used in the survey literature, where the Z-variables rather than the X-variables are called instrumental variables; see, e.g., Deville (1998; 2002) and Kott (2006), among others. Although we distinguish the response model variables $\mathbf{Z}$ from the survey variables, they are similar in nature as both are observed on the respondents only. In fact, some of the survey variables may be used in the calibration process as discussed below. This was advocated by Deville (2002) in the context of nonignorable nonresponse.

We consider an adjusted estimator of t_y of the form

$$\widehat{t}_C = \sum_{k \in S} w_k R_k Y_k, \tag{2}$$

where

$$w_k = d_k F(\widehat{\boldsymbol{\lambda}}^\top \mathbf{Z}_k) \tag{3}$$

is the calibrated weight attached to unit k and $F(\cdot)$ is a calibration function. We refer to (2) as the instrumental calibration estimator of t_y. The weights w_k in (3) are constructed so that the calibration constraints

$$\sum_{k \in S} w_k R_k \mathbf{X}_k = \sum_{k \in P} \mathbf{X}_k \tag{4}$$

are exactly satisfied when $\dim(\mathbf{X}) = \dim(\mathbf{Z})$, or hold approximately when $\dim(\mathbf{X}) > \dim(\mathbf{Z})$:

$$\widehat{\boldsymbol{\lambda}} \in \arg\min_{\boldsymbol{\lambda} \in \boldsymbol{\Lambda}} \left\| \sum_{k \in S} d_k R_k F(\boldsymbol{\lambda}^\top \mathbf{Z}_k) \mathbf{X}_k - \sum_{k \in P} \mathbf{X}_k \right\|, \tag{5}$$

where $\boldsymbol{\Lambda} \subset \mathbb{R}^{\dim(\mathbf{Z})}$ and $\|\cdot\|$ denotes the Euclidean norm. The calibration weight w_k in (3) is expressed as the product of the design weight d_k and a calibration adjustment factor $F(\widehat{\boldsymbol{\lambda}}^\top \mathbf{Z}_k)$. When $\mathbf{Z}_k = \mathbf{X}_k$, the instrumental calibration estimator (2) reduces to the conventional one-step calibration estimator; e.g., Särndal and Lundström (2005) and Haziza and

Lesage (2016).

Special cases of (3) include linear and exponential weighting. For linear weighting and $\dim(\mathbf{X}_k) = \dim(\mathbf{Z}_k)$, the weights (3) reduce to

$$w_k = d_k(1 + \widehat{\boldsymbol{\lambda}}^\top \mathbf{Z}_k), \tag{6}$$

where

$$\widehat{\boldsymbol{\lambda}} = \left(\sum_{k \in S} d_k R_k \mathbf{Z}_k \mathbf{X}_k^\top \right)^{-1} \left(\sum_{k \in P} \mathbf{X}_k - \sum_{k \in S} d_k R_k \mathbf{X}_k \right); \tag{7}$$

see Särndal and Lundström (2005). For exponential weighting, the weights (3) reduce to

$$w_k = d_k \exp\left(\widehat{\boldsymbol{\lambda}}^\top \mathbf{Z}_k \right), \tag{8}$$

but unlike in the linear case, there is no closed form expression of $\widehat{\boldsymbol{\lambda}}$.

3 Consistency of the instrumental calibration estimator

In this section, we give a set of sufficient conditions for establishing the consistency of the instrumental calibration estimator $\widehat{t}_C$, as the population and sample sizes tend to infinity. The data consists of the random vectors $(R_k, \boldsymbol{X}_k^\top, Y_k, \boldsymbol{Z}_k^\top)$, for $k \in P$. These vectors are supposed to be mutually independent and identically distributed. In practice, the indicators R_k are not observed for the nonsampled units. However, at least conceptually, nothing precludes defining these indicators for the units outside the sample. We assume that the indicators R_k satisfy the following conditions.

Assumption 1. *(Exclusion restrictions)*

$$\mathbb{C}ov\left(\boldsymbol{X}_k,\ R_k \mid \boldsymbol{Z}_k\right) = 0, \tag{9}$$

$$\mathbb{C}ov\left(Y_k,\ R_k \mid \boldsymbol{Z}_k\right) = 0. \tag{10}$$

Assumption 2. *(Response probability model)*
We have $\mathbb{E}\left\{R_k \mid \boldsymbol{Z}_k\right\} = 1/F\left(\boldsymbol{\lambda}_0^\top \boldsymbol{Z}_k\right)$*, where* $\boldsymbol{\lambda}_0$ *is a vector of unknown coefficients, which belongs to the interior of the compact set* $\boldsymbol{\Lambda}$*.*

The conditions (9) and (10) are often referred to as exclusion restrictions (D'Haultfœuille, 2010) and are key to establishing the consistency of $\widehat{t}_C$. We also assume that $F(\cdot)$ provides an adequate description of the relationship between the inverse of the probability of response and the Z-variables. This is a strong assumption as selecting the appropriate functional through model diagnostics seems to be challenging, the values corresponding to the Z-variables being only recorded for the respondents. To the best of our knowledge, a statistical procedure for selecting the function $F(\cdot)$ does not seem to be currently available in the literature, in which case the choice of $F(\cdot)$ is essentially an "act of faith". Our goal is to show that, even if Assumption 2 holds, the validity of the instrumental calibration estimator (2) still relies on (9) and (10). That is, we argue in Section 4 that the instrumental calibration estimator (2) may be highly biased and/or inefficient if (9) and (10) do not hold, even if $F(\cdot)$ is correctly specified.

In addition to Assumptions 1-2, we impose the three additional assumptions below. We consider here an asymptotic framework where N tends to infinity. The population and sampling design depends on N but to ease notation, we leave this dependence implicit hereafter.

Assumption 3. *(regularity of the sampling design)*
Let $\mathcal{U}$ be the $\sigma-$algebra generated by $(R_k, \boldsymbol{X}_k^\top, Y_k, \boldsymbol{Z}_k^\top)_{k\in P}$ and Φ denote the standard normal cumulative distribution. For every random variable T_k such that $\mathbb{E}(T_k^2) < +\infty$,

(i) $\mathbb{E}\left[\mathbb{V}\left(\frac{1}{N}\sum_{k\in S} d_k T_k \middle| \mathcal{U}\right)\right] \to 0$;

(ii) There exists $L \geq 0$ such that $\mathbb{V}\left(\sum_{k\in P} T_k\right) / \mathbb{V}\left(\sum_{k\in S} d_k T_k \middle| \mathcal{U}\right) \xrightarrow{P} L$;

(iii) $\forall t \in \mathbb{R}$, $\left|\Pr\left(\mathbb{V}\left(\sum_{k\in S} d_k T_k \middle| \mathcal{U}\right)^{-1/2}\left(\sum_{k\in S} d_k T_k - \sum_{k\in P} T_k\right) \leq t \middle| \mathcal{U}\right) - \Phi(t)\right| = o_P(1)$.

Assumption 4. *(additional technical conditions for consistency)*

(i) $F(\cdot)$ is strictly increasing;

(ii) $\mathbb{E}(\boldsymbol{X}_k | R_k = 1, \boldsymbol{Z}_k) = \boldsymbol{\Gamma}\boldsymbol{Z}_k$ *with $\boldsymbol{\Gamma}$ a matrix of rank* $\dim(\boldsymbol{Z}_k)$;

(iii) $\boldsymbol{Z}_k$ *has a compact support and* $\mathbb{E}(\boldsymbol{Z}_k \boldsymbol{Z}_k^\top)$ *is nonsingular;*

(iv) $\mathbb{E}(|Y_k|) < \infty$ *and* $\mathbb{E}(|X_{j,k}|) < \infty$ *for* $j = 1, \ldots, \dim(\boldsymbol{X}_k)$*, where* $X_{j,k}$ *denotes the* j*-th element of* $\boldsymbol{X}_k$*.*

Assumption 5. *(additional technical conditions for asymptotic normality)*

(i) $F(\cdot)$ *is continuously differentiable;*

(ii) $\mathbb{E}(Y_k^2) < \infty$ *and* $\mathbb{E}(X_{j,k}^2) < \infty$*, for* $j = 1, \ldots, \dim(\boldsymbol{X}_k)$*.*

The first condition of Assumption 3 ensures that expansion-type estimators are consistent, which requires that $n \to \infty$ as $N \to \infty$. The second condition holds, for instance, for simple random sampling without replacement if the sampling rate $n/N \to r \in [0, 1)$ as $N \to \infty$. The third condition states that, conditionally on $\mathcal{U}$, expansion-type estimators are asymptotically normal. There does not exist a Central Limit Theorem applicable to every sampling design. However, results on asymptotic normality for specific sampling designs can be found in Hájek (1960, 1964), Rosen (1972), Krewski and Rao (1981), Bickel and Freedman (1984), Chen and Rao (2007) and Breidt et al. (2015), among others.

The following theorem establishes the consistency and asymptotic normality of $\widehat{t}_C$ under the above assumptions.

Theorem 1. *(Consistency of* $\widehat{t}_C$*)*
Suppose that Assumptions 1-3 hold. Then,

1. *Under Assumption 4,*

$$\widehat{\boldsymbol{\lambda}} \xrightarrow{P} \boldsymbol{\lambda}_0$$

and the normalized error of $\widehat{t}_C$ *converges to zero; i.e.,*

$$(\widehat{t}_C - t_y)/N \xrightarrow{P} 0.$$

2. *Let* $\rho_k = F'(\boldsymbol{\lambda}_0^\top \boldsymbol{Z}_k)/F(\boldsymbol{\lambda}_0^\top \boldsymbol{Z}_k)$ *with* $F'(.)$ *the derivative of* $F(.)$ *and* $\mathbf{G} = \mathbb{E}\left(\rho_k \, \boldsymbol{X}_k \boldsymbol{Z}_k^\top\right)$*. Under Assumptions 4-5,* $\widehat{t}_C$ *is asymptotically normal:*

$$V_a^{-1/2} \left(\widehat{t}_C - t_y\right) \xrightarrow{d} \mathcal{N}(0, 1),$$

where

$$V_a = \mathbb{V}\left\{\sum_{k\in S} d_k R_k F(\boldsymbol{\lambda}_0^\top \boldsymbol{Z}_k)(Y_k - \boldsymbol{\gamma}^\top \boldsymbol{X}_k)\Big|\mathcal{U}\right\} + N^2\,\mathbb{V}(W_k) \tag{11}$$

with

$$\boldsymbol{\gamma} = \mathbf{G}\left(\mathbf{G}^\top\mathbf{G}\right)^{-1}\mathbb{E}\left(\rho_k\, Y_k \boldsymbol{Z}_k\right)$$

and

$$W_k = \left\{R_k F(\boldsymbol{\lambda}_0^\top \boldsymbol{Z}_k) - 1\right\}(Y_k - \boldsymbol{\gamma}^\top \boldsymbol{X}_k).$$

The proof of Theorem 1 is presented in Appendix A.1. When the relationship between the X-variables and the Z-variables is weak, the residuals $Y_k - \boldsymbol{\gamma}^\top \boldsymbol{X}_k$ may become large in absolute value, as we illustrate in Example 1 below. It follows from (11) that the approximate variance V_a of $\widehat{t}_C$ is amplified in this case. Note that both terms on the right-hand side of (11) are affected by large values of $Y_k - \boldsymbol{\gamma}^\top \boldsymbol{X}_k$. The potential inefficiency of $\widehat{t}_C$ is due to the fact that the calibration equations provide very little information on $\boldsymbol{\lambda}_0$, which, in turn, is poorly estimated. This is similar to what is encountered in the context of linear regression with endogenous regressors and instrumental variables. The two-stage least square estimator of the slope coefficient becomes very unstable when the instrumental variables are weakly correlated with the endogenous regressors; e.g., Wooldridge (2002, section 5.2.6) and Bound et al. (1995).

We make the following additional remarks.

Remark 1. *When $dim(\boldsymbol{Z}_k) = dim(\boldsymbol{X}_k)$, the expression of $\boldsymbol{\gamma}$ reduces to*

$$\begin{aligned}\boldsymbol{\gamma} &= \left\{\mathbb{E}\left(\rho_k\, \boldsymbol{Z}_k \boldsymbol{X}_k^\top\right)\right\}^{-1}\mathbb{E}\left(\rho_k\, \boldsymbol{Z}_k Y_k\right)\\ &= \left(\boldsymbol{\Gamma}^\top\right)^{-1}\left\{\mathbb{E}\left(\rho_k\, \boldsymbol{Z}_k \boldsymbol{Z}_k^\top\right)\right\}^{-1}\mathbb{E}\left(\rho_k\, \boldsymbol{Z}_k Y_k\right).\end{aligned}$$

In this case, $\boldsymbol{\gamma}$ is a weighted version of the customary two-stage least square estimator of the slope coefficient. Note that for linear weighting (6), we have $\rho_k = (1 + \boldsymbol{\lambda}_0^\top \boldsymbol{Z}_k)^{-1}$, whereas $\rho_k = 1$ in the case of exponential weighting (8).

Remark 2. *In the case of a census, Assumption 3 no longer holds but it can be shown that the estimator is still consistent and asymptotically normal. In such a case, the first term on*

the right-hand side of (11) vanishes and we get:

$$\frac{1}{\sqrt{N}}\left(\widehat{t}_C - t_y\right) \xrightarrow{d} \mathcal{N}\left(0, \mathbb{V}(W_k)\right).$$

To illustrate the risk of variance amplification of $\widehat{t}_C$, we consider the following example.

Example 1. *Consider the census case with* $\boldsymbol{X}_k = (1, X_k)^\top$, $\boldsymbol{Z}_k = (1, Z_k)^\top$, *where* $(X_k, Z_k) \in \mathbb{R}^2$. *We assume that the relationship between* Y *and* Z *is described by*

$$Y_k = \beta_0 + \beta_1 Z_k + \varepsilon_k,$$

where β_0 *and* β_1 *are unknown coefficients and* $\mathbb{E}(\varepsilon_k | R_k, \boldsymbol{Z}_k) = 0$. *Also, we assume that the variables* X *and* Z *are related through*

$$X_k = \Gamma_0 + \Gamma_1 Z_k + \nu_k$$

with $\mathbb{E}(\nu_k | R_k, \boldsymbol{Z}_k) = 0$ *and* $\mathbb{C}ov(\varepsilon_k, \nu_k \mid \boldsymbol{Z}_k, R_k) = 0$. *Then Assumption 1 is satisfied, and the asymptotic variance of* $\widehat{t}_C$ *is equal to*

$$\begin{aligned}
\mathbb{V}(W_k) &= \mathbb{V}\left[\left\{R_k F(\boldsymbol{\lambda}_0^\top \boldsymbol{Z}_k) - 1\right\}\left(\varepsilon_k - \frac{\beta_1}{\Gamma_1}\nu_k\right)\right] \\
&= \mathbb{V}\left[\left\{R_k F(\boldsymbol{\lambda}_0^\top \boldsymbol{Z}_k) - 1\right\}\varepsilon_k\right] + \left(\frac{\beta_1}{\Gamma_1}\right)^2 \mathbb{V}\left[\left\{R_k F(\boldsymbol{\lambda}_0^\top \boldsymbol{Z}_k) - 1\right\}\nu_k\right]. \qquad (12)
\end{aligned}$$

The second term on the right-hand side of (12) vanishes in the following situations: (i) $\beta_1 = 0$, *which implies that* Y *is unrelated to* Z; *(ii)* $Z_k = X_k$ *for all* k, *implying that* $\nu_k = 0$, *which corresponds to the conventional one-step calibration procedure. In both (i)and (ii), the asymptotic variance of* $\widehat{t}_C$ *depends solely on the strength of the relationship between* Y *and* X. *Otherwise, the second term on the right-hand side of (12) contributes to the amplification of the asymptotic variance. The amplification increases as* Γ_1 *decreases, that is to say as the association between* X *and* Z *becomes weaker.*

4 Bias amplification of the instrumental calibration estimator

In Section 3, we established the consistency of the instrumental calibration estimator $\widehat{t}_C$ under the exclusion restrictions (9) and (10). We now examine the situation where a wrong

choice of the vector of the X-variables entails a violation of (9). For simplicity, we focus hereafter on the case $\dim(\boldsymbol{X}_k) = \dim(\boldsymbol{Z}_k)$.

A violation of (9) may occur when there exists an unobserved variable U, independent of $\mathbf{Z}$ and Y, which is related to both the response indicator variable R and the X-variables (see Figure 1). For example, suppose that a household survey is conducted and that the domain of interest is a metropolis. Let $Y = Z$ be the household income, which is only observed for $k \in S_r$. Let X be the household square footage for which the population total is known. In large urban areas, it is reasonable to assume that there is a relationship between household income and household square footage. Let U be an unobserved variable representing the distance between the workplace and the home of the selected unit. It is not unrealistic to assume that U is related to R because the greater the distance between the workplace and the home, the less chance for a unit to be contacted in the case of face-to-face or telephone interviews as the sample unit would typically spend a significant amount of time commuting between both places. Also, as the distance between the workplace and the home increases, we can expect the household square footage to increase because cities are usually more expensive than suburbs in term of price by square footage. Here, the variable U is related to both R and X and hence, the exclusion restriction (9) is violated. As a result, the household square footage would not be a good candidate for playing the role of an instrumental variable.

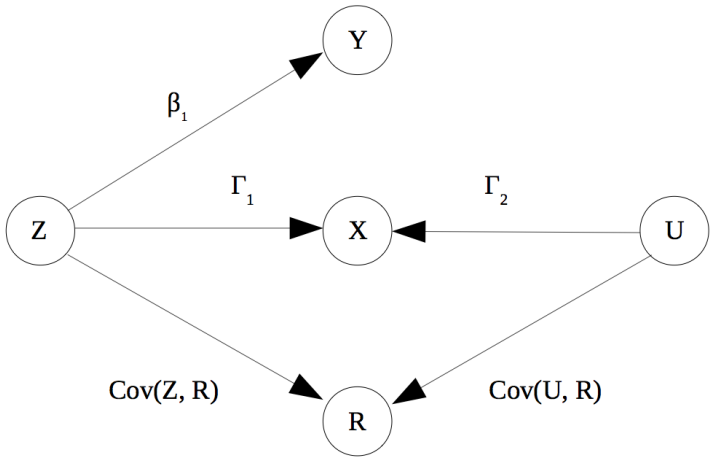

Figure 1: Relations between the variables $\boldsymbol{X}_k$, $\boldsymbol{Z}_k$, Y_k, U_k and R_k

.

The error, $\widehat{t}_C - t_y$, can be expressed as the sum of four terms:

$$\begin{aligned}\widehat{t}_C - t_y &= \left(\widehat{t}_\pi - t_y\right) + \sum_{k\in S} d_k \left\{R_k\, F(\boldsymbol{\lambda}_0^\top \boldsymbol{Z}_k) - 1\right\} Y_k \\ &+ \sum_{k\in S} d_k R_k \left\{F(\widehat{\boldsymbol{\lambda}}^\top \boldsymbol{Z}_k) - F(\boldsymbol{\lambda}_\infty^\top \boldsymbol{Z}_k)\right\} Y_k \\ &+ \sum_{k\in S} d_k R_k \left\{F(\boldsymbol{\lambda}_\infty^\top \boldsymbol{Z}_k) - F(\boldsymbol{\lambda}_0^\top \boldsymbol{Z}_k)\right\} Y_k, \end{aligned} \tag{13}$$

where $\boldsymbol{\lambda}_0$ is the vector of parameters in the nonresponse model and $\boldsymbol{\lambda}_\infty$ is the probability limit of $\widehat{\boldsymbol{\lambda}}$. The first term on the right-hand side of (13) is the sampling error, the second term is the nonresponse error assuming that the response probabilities are known, the third term is the error arising from estimating $\boldsymbol{\lambda}_\infty$, whereas the last term is due to the fact that the exclusion restriction (9) is violated, implying $\boldsymbol{\lambda}_\infty \neq \boldsymbol{\lambda}_0$, in general.

To get a better understanding of the last point, note that $\boldsymbol{\lambda}_\infty$ is the solution of

$$\mathbb{E}\left\{R_k F(\boldsymbol{\lambda}^\top \boldsymbol{Z}_k)\boldsymbol{X}_k\right\} = \mathbb{E}\left(\boldsymbol{X}_k\right), \tag{14}$$

which are the moment equations corresponding to the calibration equations (4). When the exclusion restriction (9) is violated, these equalities are not satisfied by $\boldsymbol{\lambda} = \boldsymbol{\lambda}_0$. More specifically, we show in Appendix A.2 that

$$\boldsymbol{\lambda}_\infty - \boldsymbol{\lambda}_0 = -\left[\mathbb{E}\left\{f_k F(\boldsymbol{\lambda}_0^\top \boldsymbol{Z}_k)^{-1}\boldsymbol{X}_k \boldsymbol{Z}_k^\top\right\}\right]^{-1} \mathbb{E}\left\{F(\boldsymbol{\lambda}_\infty^\top \boldsymbol{Z}_k)\, \mathbb{C}ov\left(\boldsymbol{X}_k,\, R_k \mid \boldsymbol{Z}_k\right)\right\}, \tag{15}$$

where

$$\begin{cases} f_k = \frac{F(\boldsymbol{\lambda}_\infty^\top \boldsymbol{Z}_k) - F\left(\boldsymbol{\lambda}_0^\top \boldsymbol{Z}_k\right)}{(\boldsymbol{\lambda}_\infty - \boldsymbol{\lambda}_0)^\top \boldsymbol{Z}_k} & \text{if } \boldsymbol{\lambda}_\infty^\top \boldsymbol{Z}_k \neq \boldsymbol{\lambda}_0^\top \boldsymbol{Z}_k, \\ f_k = 1 & \text{otherwise.} \end{cases} \tag{16}$$

For the special case of linear weighting (6), we have $f_k = 1$ for all k.

A consequence of $\boldsymbol{\lambda}_\infty \neq \boldsymbol{\lambda}_0$ is that the last term on the right-hand side of (13) does not vanish, which implies that $\widehat{t}_C$ is inconsistent. The following theorem gives the expression of its asymptotic bias.

Theorem 2. *(Inconsistency of $\widehat{t}_C$)*
Suppose that (10) *and Assumptions 2-4 hold,* $\dim(\boldsymbol{X}_k) = \dim(\boldsymbol{Z}_k)$ *and* (14) *admits at least one solution. Then,*

1. *The solution of* (14) *is unique. Moreover, denoting this solution by* $\boldsymbol{\lambda}_\infty$*,*

$$\widehat{\boldsymbol{\lambda}} \xrightarrow{P} \boldsymbol{\lambda}_\infty.$$

2. *The normalized calibration estimator* $N^{-1}(\widehat{t}_C - t_y)$ *is inconsistent in general, and the asymptotic bias is given by*

$$(\widehat{t}_C - t_y)/N \xrightarrow{P} -\boldsymbol{\gamma}_{fR}^\top \, \mathbb{E}\left\{F(\boldsymbol{\lambda}_\infty^\top \boldsymbol{Z}_k) \, \mathbb{C}ov\left(\boldsymbol{X}_k,\ R_k \mid \boldsymbol{Z}_k\right)\right\}, \tag{17}$$

where

$$\boldsymbol{\gamma}_{fR} = \left\{\mathbb{E}(f_k R_k \boldsymbol{Z}_k \boldsymbol{X}_k^\top)\right\}^{-1} \ \mathbb{E}(f_k R_k \boldsymbol{Z}_k Y_k) \tag{18}$$

and f_k *is defined in* (16)*.*

The proof of Theorem 2 is presented in Appendix A.2. As we shall see in Example 2 below, the term $\boldsymbol{\gamma}_{fR}$ in (18) can be viewed as a bias amplifier term.

Example 2. *Once again, we consider the census case,* $n = N$*. We assume that the relationship between* Y *and* Z *is described by the following linear regression model:*

$$Y_k = \beta_0 + \beta_1 Z_k + \varepsilon_k,$$

where $\mathbb{E}(\varepsilon_k | R_k, U_k, Z_k) = 0$*. Assume that the variables* X*,* Z *and* U *are related through*

$$X_k = \Gamma_0 + \Gamma_1 Z_k + \Gamma_2 U_k + \nu_k,$$

where $\mathbb{E}(\nu_k | R_k, U_k, Z_k) = 0$*. In this case, (17) reduces to*

$$(\widehat{t}_C - t_y)/N \xrightarrow{P} -\beta_1 \, \frac{\Gamma_2}{\Gamma_1} \, \mathbb{E}\left\{F(\boldsymbol{\lambda}_\infty^\top \boldsymbol{Z}_k) \, \mathbb{C}ov\left(U_k,\ R_k \mid Z_k\right)\right\}. \tag{19}$$

If $\beta_1 = 0$*, i.e., there is no relationship between* Y *and* Z*, the asymptotic bias* (19) *vanishes, as expected. If either* $\Gamma_2 = 0$ *or* $\mathbb{C}ov\left(U_k,\ R_k \mid Z_k\right) = 0$*, the asymptotic bias is equal to zero as the exclusion restriction* (9) *is satisfied. Otherwise, the asymptotic bias is large when* Γ_2 *is large (i.e., there is a strong association between* X *and* U*) and/or if* $\mathbb{C}ov\left(U_k,\ R_k \mid Z_k\right)$ *is large (i.e., there is a strong association between* R *and* U*). For a given value of* Γ_2 $\mathbb{C}ov\left(U_k,\ R_k \mid Z_k\right) \neq 0$*, the bias increases as* Γ_1 *decreases; i.e., as the relationship between* X *and* Z *becomes weaker.*

5 Variance estimation

Drawing valid inferences relies on the availability of a consistent point and variance estimators. Here, we derive an estimator of $\mathbb{V}(\widehat{t}_C)$ through the reverse approach of Shao and Steel (1999). The proposed variance estimator is consistent for the true variance of the instrumental calibration estimator, provided that the sampling fraction n/N is small. For simplicity, we focus on the case $\dim(\boldsymbol{Z}_k) = \dim(\boldsymbol{X}_k)$. The total variance of $\widehat{t}_C$ in (2), denoted by V_{tot}, can be expressed as the sum of two terms:

$$\mathrm{V}_{tot} = \mathrm{V}_1 + \mathrm{V}_2, \tag{20}$$

where

$$\mathrm{V}_1 = \mathbb{E}\left[\mathbb{V}\left(\widehat{t}_C \mid \mathcal{U}\right)\right]$$

and

$$\mathrm{V}_2 = \mathbb{V}\left[\mathbb{E}\left(\widehat{t}_C - t_y \mid \mathcal{U}\right)\right].$$

The decomposition of the variance (20) is often referred to as the reverse decomposition; see, e.g., Shao and Steel (1999) and Kim and Rao (2009). Under mild regularity conditions, the first term on the right-hand side of (20) is $O(N^2/n)$, whereas the second term is $O(N)$. Therefore, the contribution of V_2 to the total variance V_{tot} is negligible provided that the sampling fraction n/N is negligible (Shao and Steel, 1999).

Assuming that the sampling fraction n/N is negligible, an estimator of V_{tot} is obtained by estimating V_1. To that end, it suffices to obtain a consistent estimator of $\mathbb{V}(\widehat{t}_C \mid \mathcal{U})$, which represents the sampling variance of $\widehat{t}_C$ conditional on all the other quantities. Since $\widehat{t}_C$ in (2) can be expressed as a smooth function of estimated totals, any complete data variance estimation procedure designed for estimating the design variance of smooth functions of totals can be used. An estimator of V_1 based on a first-order Taylor expansion procedure leads to

$$\widehat{\mathrm{V}}_1 = \sum_{k\in S}\sum_{l\in S} \frac{\pi_{kl} - \pi_k\pi_l}{\pi_k\pi_l\pi_{kl}}\widehat{\eta}_k\widehat{\eta}_l, \tag{21}$$

where

$$\widehat{\eta}_k = R_k\, F\left(\widehat{\boldsymbol{\lambda}}^\top \mathbf{Z}_k\right)\, \left(Y_k - \widehat{\boldsymbol{\gamma}}^\top \boldsymbol{X}_k\right)$$

and

$$\widehat{\boldsymbol{\gamma}} = \left\{ \sum_{k \in S} d_k R_k F'(\widehat{\boldsymbol{\lambda}}^\top \mathbf{Z}_k)\, \mathbf{Z}_k \mathbf{X}_k^\top \right\}^{-1} \sum_{k \in S} d_k R_k F'(\widehat{\boldsymbol{\lambda}}^\top \mathbf{Z}_k)\, \mathbf{Z}_k Y_k;$$

e.g., Demnati and Rao (2004) and D'Arrigo and Skinner (2010). It is worth noting that the estimator (21) provides a consistent estimator of V_1 even if $\widehat{t}_C$ is biased. A 95% confidence interval for t_y is given by

$$\widehat{t}_C \pm 1.96\sqrt{\widehat{\mathrm{V}}_1}.$$

We expect the coverage rate of this interval to be close to 95% provided that $\widehat{t}_C$ exhibits a small bias, which would occur if the exclusion restrictions (9) and (10) are satisfied.

6 Simulation study

We conducted a simulation study to assess the properties of the instrumental calibration estimator in terms of bias and efficiency. We generated finite populations of size $N = 1\,000$, each consisting of a variable of interest Y, an instrumental variable X, a response model variable Z and an unobserved variable U. Let $\boldsymbol{X}_k = (1, X_k)^\top$ and $\boldsymbol{Z}_k = (1, Z_k)^\top$. The variables Z and U were first generated independently from a uniform distribution $\left(-\sqrt{3}, \sqrt{3}\right)$, so that $\mathbb{E}(Z) = \mathbb{E}(U) = 0$ and $\mathbb{V}(Z) = \mathbb{V}(U) = 1$. Then, given Z, the variable Y was generated according to two models:

(i) a linear regression model:

$$Y_{1,k} = 10 + 5Z_k + \varepsilon_{1,k}, \tag{22}$$

where the errors $\varepsilon_{1,k}$ were generated from a normal distribution with mean equal to 0 and variance equal to 4. The resulting coefficient of determination was equal to 85%;

(ii) an exponential model:

$$Y_{2,k} = \exp(2.5\ Z_k) + \varepsilon_{2,k}, \tag{23}$$

where the errors $\varepsilon_{2,k}$ were generated from a normal distribution with mean equal to 0 and variance equal to 4. Figure 2 shows the relationship between Y and Z for both types of populations.

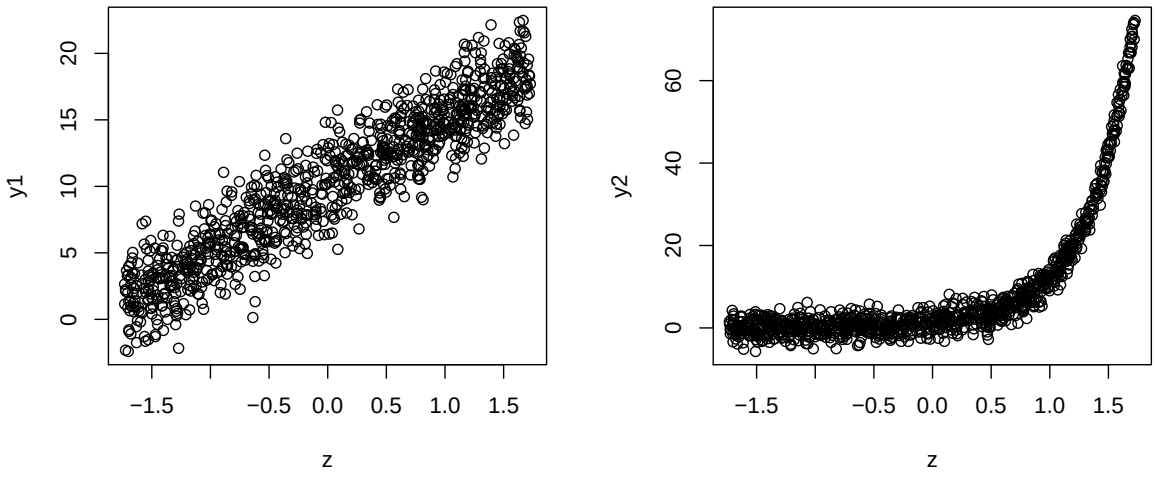

Figure 2: Relationship between Y and Z

Finally, given the values of Z and U, the X-values were generated according to the linear regression model

$$X_k \;\; = \;\; \Gamma_1 Z_k + \Gamma_2 U_k + \nu_k,$$

where $\nu_k \sim \mathcal{N}(0, 1 - \Gamma_1^2 - \Gamma_2^2)$. Then $\mathbb{V}(X_k) = 1$ and Γ_1 (respectively Γ_2) corresponds to the correlation coefficient between X_k and Z_k (respectively U_k). We used the following values for Γ_1 and Γ_2: $\Gamma_1 \in \{0.2, 0.4, 0.6\}$ and $\Gamma_2 \in \{0, 0.1, 0.3, 0.5\}$.

The response indicators R_k were generated independently from a Bernoulli distribution with parameter p_k equal to

$$p_k = \frac{1}{2 + 0.35\ Z_k} + 0.1\ U_k. \tag{24}$$

This led to an overall response rate of around 50%. Figure 3 shows the relationship between Z_k and p_k. Note that (24) implies that $\mathbb{C}ov\,(U_k,\ R_k \mid Z_k) \neq 0$. As a result, $\mathbb{C}ov\,(X_k,\ R_k \mid Z_k) \neq 0$ unless $\Gamma_2 = 0$.

Finally, we considered the census case where $n = N = 1\ 000$. In each population, the whole process (i.e., generating the finite population and generating nonresponse), was repeated

$K = 10,000$ times, leading to $K = 10,000$ sets of respondents.

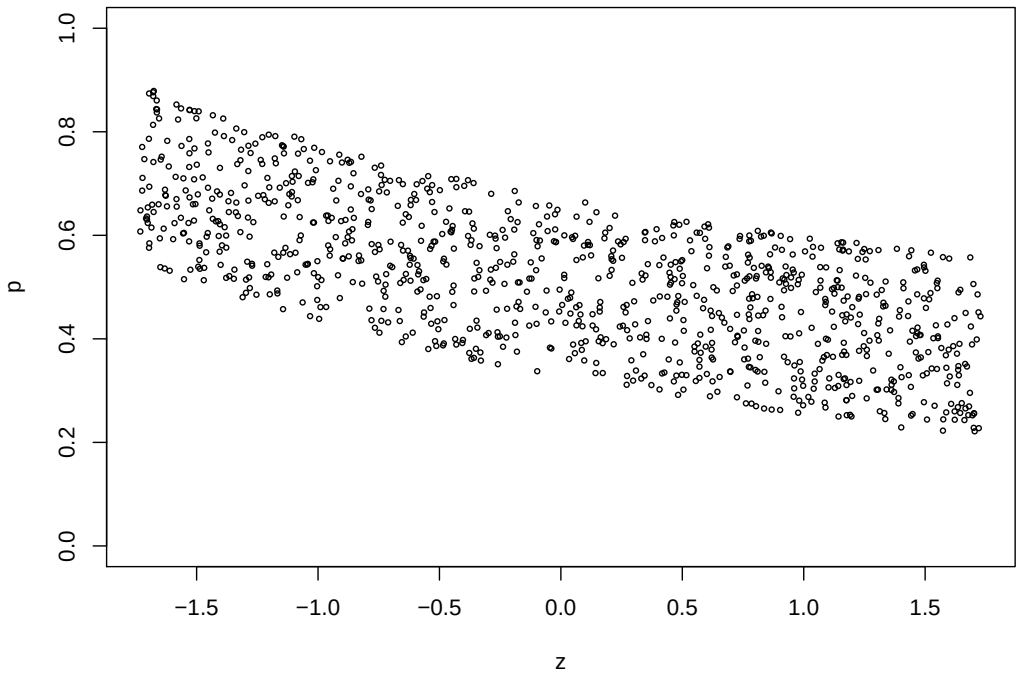

Figure 3: Relationship between Z and the probability of response p given by (24)

We computed the following estimators: (i) the unadjusted estimator $\widehat{t}_{un}$ given by (1); (ii) the instrumental calibration estimator $\widehat{t}_C$, based on linear weighting (6); (iii) the conventional one-step calibration estimator $\widehat{t}_{Conv}$ given by (2), but where $\mathbf{X}_k$ is used in place of $\mathbf{Z}_k$ in the computation of the weights w_k. That is, the latter does not make use of the variable Z_k. Nevertheless, we show below that it can outperform $\widehat{t}_C$, because of the variance and bias amplification phenomena described above.

Equation (24) implies that $\mathbb{E}\{R_k \mid \boldsymbol{Z}_k\}^{-1}$ is linear. To ensure that Assumption 2 holds, we therefore used linear weighting ($F(u) = u$) to construct $\widehat{t}_C$. The estimator $\widehat{t}_{Conv}$ was also based on a linear weighting procedure.

For an estimator $\widehat{t}$, we computed the (Monte Carlo percent) relative bias given by

$$RB_{MC}\left(\widehat{t}\right) = 100 \times \frac{1}{K}\sum_{j=1}^{K}\frac{\left(\widehat{t}_{(j)} - t_{y(j)}\right)}{t_{y(j)}},$$

where $\widehat{t}_{(j)}$ and $t_{y(j)}$ denote, respectively, the estimator $\widehat{t}$ and the true population total t_y at the j-th iteration, $j = 1, \cdots, K$.

As a measure of variability of $\widehat{t}$, we also computed a measure of percent relative standard error, given by

$$RSD_{MC}(\widehat{t}) = 100 \times \frac{\left[\frac{1}{K}\sum_{j=1}^{K}\left(\widehat{t}_{(j)} - t_{y(j)}\right)^2 - \left\{E_{MC}(\widehat{t} - t_y)\right\}^2\right]^{0.5}}{E_{MC}(t_y)},$$

where $E_{MC}(\widehat{t}) = \sum_{j=1}^{K}\widehat{t}_{(j)}/K$ and $E_{MC}(t_y) = \sum_{j=1}^{K} t_{y(j)}/K$.

Tables 1 and 2 show the percent relative bias and standard error (in parentheses) of $\widehat{t}_C$ for populations generated according to (22) and (23), respectively. For populations generated according to (22), the unadjusted estimator exhibited a relative bias of -9.0% and a relative standard error equal to 2.4%. For populations generated according to (23), the relative bias of the unadjusted estimator was equal to -21.2% and its relative standard error equal to 7.3%.

	Γ_2			
Γ_1	0.0	0.1	0.3	0.5
0.6	0.0	-1.6	-4.9	-8.1
	(2.2)	(2.2)	(2.1)	(2.1)
0.4	0.1	-2.4	-7.3	-12.0
	(3.8)	(3.7)	(3.6)	(3.5)
0.2	0.4	-4.8	-14.3	-23.5
	(8.6)	(8.1)	(7.9)	(8.1)

Table 1: Relative bias and standard error (in parentheses) of $\widehat{t}_C$ for a population generated according to (22)

	Γ_2			
Γ_1	0.0	0.1	0.3	0.5
0.6	-0.2	-4.0	-11.7	-19.2
	(6.9)	(6.8)	(6.5)	(6.2)
0.4	-0.0	-6.0	-17.4	-28.5
	(9.9)	(9.6)	(9.2)	(9.0)
0.2	0.8	-11.6	-34.0	-55.6
	(20.8)	(19.5)	(19.0)	(19.4)

Table 2: Relative bias and standard error (in parentheses) of $\widehat{t}_C$ for a population generated according to (23)

From Tables 1 and 2, the calibration estimator $\widehat{t}_C$ showed negligible bias when $\Gamma_2 = 0$. These results are not surprising since the restriction exclusion condition (9) was satisfied in this case. However, the variance of $\widehat{t}_C$ increased rapidly as Γ_1 decreased. This is consistent with (12). For example, in Table 1, for $\Gamma_1 = 0.6$, the relative standard error of $\widehat{t}_C$ was equal to 2.2%, whereas it was equal to 8.6% for $\Gamma_1 = 0.2$. Similar results were obtained in Table 2 for $\Gamma_2 = 0$. Our results are consistent with those obtained by Osier (2012).

We now turn to the case $\Gamma_2 \neq 0$. The restriction exclusion condition (9) is no longer satisfied as both Γ_2 and $\mathbb{C}ov\,(U_k,\ R_k \mid Z_k)$ are different from zero. In this case, the instrumental calibration estimator $\widehat{t}_C$ exhibited some bias. The bias increased as Γ_2 increased. For a given value of Γ_2, the bias of $\widehat{t}_C$ rapidly increased as Γ_1 decreased. This is consistent with (19). The same remark can be made about the variance of $\widehat{t}_C$ that increased rapidly as Γ_1 decreased. For example, in Table 1, for $\Gamma_1 = 0.6$ and $\Gamma_2 = 0.3$ the relative bias and standard error were equal to -4.9% and 2.1%, respectively, whereas they were equal to -14.3% and 7.9% for $\Gamma_1 = 0.2$ and $\Gamma_2 = 0.3$. These results suggest that the instrumental calibration estimator may suffer simultaneously from bias and variance amplification in the case of a weak correlation between the instrumental variable X and the model response variable Z. Similar results were obtained in Table 2. It is worth noting that $\widehat{t}_C$ exhibited a larger bias and a larger relative standard deviation than that of $\widehat{t}_{un}$ for some pairs (Γ_1, Γ_2). For example, for $\Gamma_1 = 0.4$ and $\Gamma_2 = 0.5$ the relative bias and standard error were equal to -12.0% and 3.5%, respectively; see Table 1.

Figures 4 and 5 show side-by-side boxplots of the relative error of $\widehat{t}_C$, $100 \times (\widehat{t}_C - t_y)/t_y$ corresponding to the populations generated according to (22) and (23), respectively. Each boxplot shows the distribution of the relative error of $\widehat{t}_C$ for a given pair (Γ_1, Γ_2). These figures are in line with the results of Tables 1 and 2. While the distribution of relative errors is centered around zero when $\Gamma_2 = 0$, they shift to the left as Γ_2 increases. The dispersion of this distribution also increases as Γ_1 tends to zero.

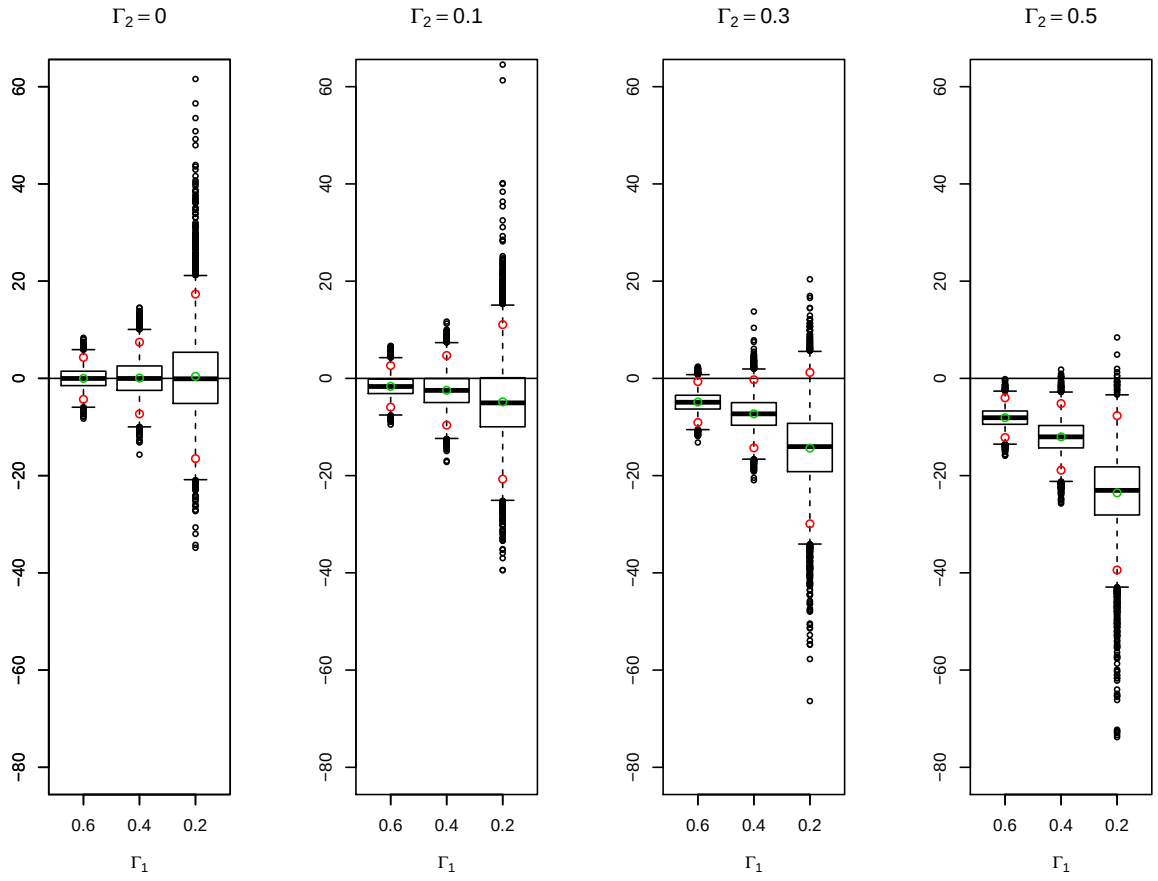

Figure 4: Relative error of $\widehat{t}_C$ for different pairs (Γ_1, Γ_2) with a linear model for Y.

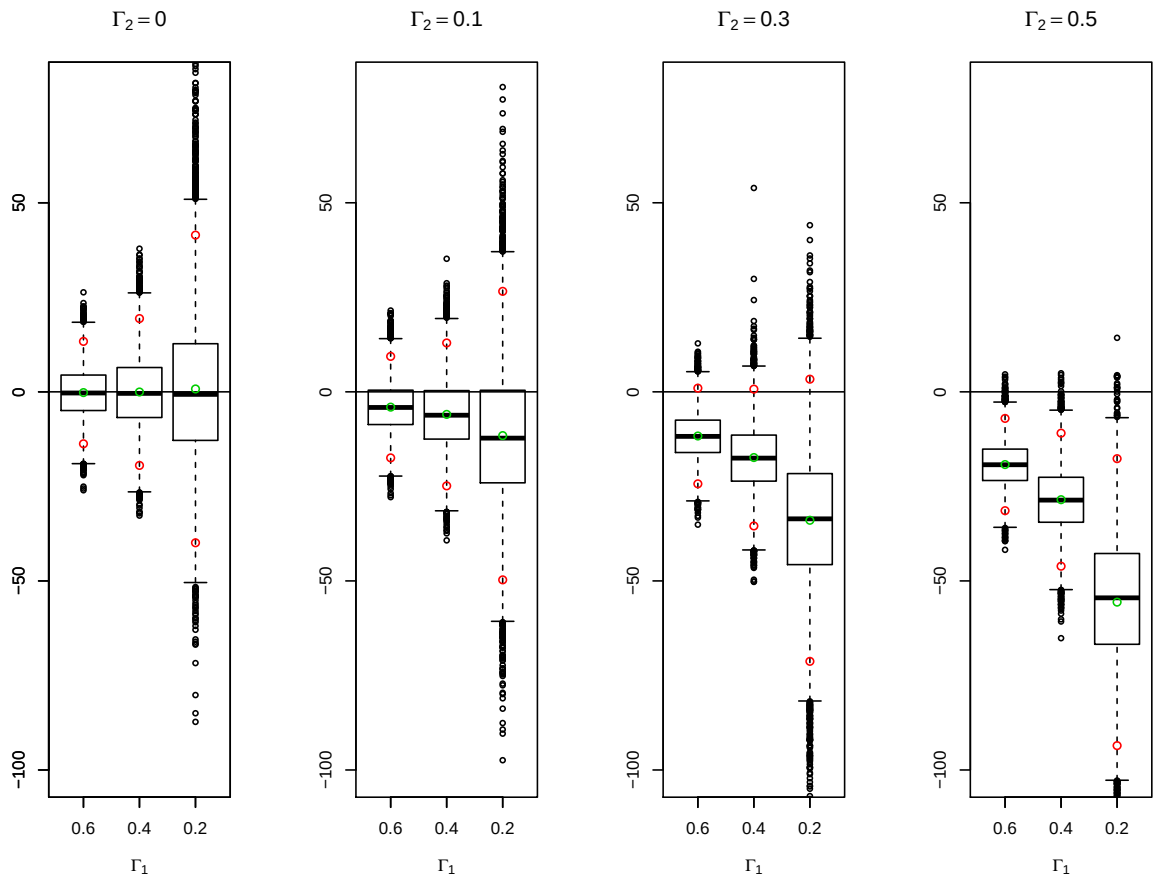

Figure 5: Relative error of $\widehat{t}_C$ for different pairs (Γ_1, Γ_2) with an exponential model for Y.

We now turn to the case of the conventional one-step calibration approach. The results are shown in Tables 3 and 4. The conventional one-step estimator, $\widehat{t}_{Conv}$, was biased in all the scenarios. Unlike the instrumental calibration estimator, $\widehat{t}_{Conv}$ exhibited some bias when $\Gamma_2 = 0$. For a given value of Γ_1, the bias increased as Γ_2 increased. Also, for a given value of Γ_2, the bias increased as Γ_1 decreased. The properties of the conventional one-step calibration estimator depended solely on the strength of the relationship between the Y-variable and the X-variable. In the worst scenarios (e.g., small values of Γ_1 and large values of Γ_2), the bias of $\widehat{t}_{Conv}$ was similar to that of the unadjusted estimator, although the former was more efficient. In these situations, the relative bias and standard error of $\widehat{t}_{Conv}$ were considerably smaller than those of the instrumental calibration estimator. These results suggest that, when the exclusion restrictions are violated, the conventional one-step calibration estimator may outperform the instrumental calibration estimator.

	Γ_2			
Γ_1	0.0	0.1	0.3	0.5
0.6	-5.8	-6.3	-7.5	-8.6
	(1.4)	(1.4)	(1.4)	(1.4)
0.4	-7.5	-7.9	-8.7	-9.5
	(1.6)	(1.6)	(1.5)	(1.5)
0.2	-8.6	-8.8	-9.2	-9.6
	(1.6)	(1.6)	(1.6)	(1.6)

Table 3: Relative bias and standard error (in parentheses) of $\widehat{t}_{Conv}$ for different pairs (Γ_1, Γ_2) corresponding to population generated according to (22)

	Γ_2			
Γ_1	0.0	0.1	0.3	0.5
0.6	-13.7	-15.0	-17.7	-20.5
	(5.5)	(5.5)	(5.3)	(5.2)
0.4	-17.9	-18.8	-20.6	-22.5
	(5.6)	(5.6)	(5.5)	(5.4)
0.2	-20.4	-20.9	-21.8	-22.8
	(5.7)	(5.7)	(5.6)	(5.6)

Table 4: Relative bias and standard error (in parentheses) of $\widehat{t}_{Conv}$ for different pairs (Γ_1, Γ_2) corresponding to population generated according to (22)

7 Final remarks

In this paper, we showed that instrumental calibration may be successful in reducing the nonresponse bias even when the probability of response depends on the Y-variable subject to missingness. However, one needs to exercise some caution as the resulting estimator may be highly biased and/or unstable. We first argued that instrumental calibration leads to negligible bias provided that Assumption 2 and the restriction exclusions (9) and (10) are satisfied. However, a procedure for validating the choice of $F(\cdot)$ through model diagnostics when the probability of response depends on variables subject to missingness does not seem to be yet available. Also, a statistical procedure for testing whether or not (9) and (10) hold is currently not available in the literature.

In practice, the search for an instrumental variable X that satisfies (9) and (10) is key. Potential candidates include (i) variables that are observed among the respondents and for

which the corresponding population total is known and (ii) those that are observed on every sample unit (respondent and nonrespondent). In statistical agencies, commonly encountered variables of the type (i) include socio-demographic variables such as age and sex and geographical variables (e.g., region or province). Examples of variables of the type (ii) include paradata also called field process data (Couper, 1998). For example, in face-to-face surveys, paradata may include interviewer observations about the physical and social characteristics of the selected households (Durrant et al., 2011). When paradata are collected, it may be wise to attempt collecting some variables that are believed to be good candidates as instrumental variables.

If several potential candidates are available, one could improve the accuracy of instrumental calibration by deriving a scalar instrumental variable compressing the information contained in the multiple candidates. This could be achieved by regressing the variable Y on the vector $\mathbf{X}$ of potential candidates based on the responding units and use the predicted values $\widehat{\mu}$ as a scalar instrumental variable. This is a topic of future research.

We have used the generic notation Y to denote a survey variable. In practice, many surveys conducted by statistical agencies are multipurpose surveys in the sense that information is collected on a large number of survey variables. In such surveys, the probability of response may depend on multiple survey variables, which makes the application of instrumental calibration challenging.

When the search for an instrumental variable is unsuccessful, it may be more prudent to use the conventional one-step calibration procedure solely based on X-variables. Although one may not be able to reduce the bias to the same extent as with instrumental calibration, there is no risk of bias and variance amplification, which in turn offers some protection against an unduly large bias and/or variance. In this perspective, instrumental calibration approach could still be used as a sensitivity check, to assess the potential effect of nonignorable nonresponse.

Acknowledgement

The work of the second author was supported by grants from the Natural Sciences and Engineering Research Council of Canada and the Canadian Statistical Sciences Institute. The authors wish to thank the Editor, an Associate Editor and two referees for their constructive comments.

References

D'Arrigo, J. and Skinner, C. (2010). Linearization variance estimation for generalized raking estimators in the presence of nonresponse. *Survey Methodology* **36** 181–192.

Bhattacharya, J. and Vogt, W. B. (2007). Do Instrumental Variables Belong in Propensity Scores? NBER Technical Working Paper no. 343. Cambridge, MA: National Bureau of Economic Research.

Bickel, P. J., and Freedman, D. A. (1984). Asymptotic normality and the bootstrap in stratified sampling. *The Annals of Statistics* **12**, 470–482.

Bound, J., Jaeger, D. A., and Baker, R. M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association* **90**, 443–450.

Breidt, F. J., Opsomer, J. D. and Sanchez-Borrego, I. (2015). Nonparametric variance estimation under fine stratification: An alternative to collapsed strata. *Journal of the American Statistical Association* **514**, 822–833.

Chang, T. and Kott, P. S. (2008). Using calibration weighting to adjust for nonresponse under a plausible model. *Biometrika* **95**, 557–571.

Chen, J. and Rao, J. N. K. (2007). Asymptotic normality under two-phase sampling designs. *Statistica Sinica* **17**, 1047–64.

Couper, M.P. (1998). Measuring survey quality in a CASIC environment. *Proceedings of the Survey Research Methods Section, American Statistical Association* **48**, 743–772.

Demnati, A. and Rao, J. N. K. (2004). Linearization variance estimators for survey data (with discussion). *Survey Methodology* **30**, 17–34.

Deville, J-C. and Särndal, C. E. (1992). Calibration estimators in survey sampling. *Journal of the American statistical Association* **87**, 376–382.

Deville, J-C., Särndal, C. E. and Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American statistical Association* **88**, 1013–1020.

Deville, J-C. (1998). La correction de la non-réponse par calage ou par échantillonnage équilibré. *Paper presented at the Congrès de l'ACFAS, Sherbrooke, Quebec.*

Deville, J-C. (2002). La correction de la non-réponse par calage généralisé. *Actes des Journées de Méthodologie Statistique de l'Insee*, Paris, France.

D'Haultfœuille, X. (2010). A New Instrumental Method For Dealing with Endogenous Selection. *Journal of Econometrics* **154**, 1–10.

Durrant, G. B., D'Arrigo, J. and Steele, F. (2011). Using field process data to predict best times of contact conditioning on household and interviewer influences. *Journal of the Royal Statistical Society: series A* **174**, 1029–1049.

Eltinge, J. L. and Yansaneh, I. S. (1997). Diagnostics For Formation Of Nonresponse Adjustment Cells, With An Application To Income Nonresponse In The U.S. Consumer Expenditure Survey. *Survey Methodology* **23**, 33–40.

Hájek, J. (1960). Limiting distributions in simple random sampling from a finite population. *Publications of the Mathematics Institute of the Hungarian Academy of Science* **5**, 361–74.

Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics* **35**, 1491–1523.

Haziza, D. and Lesage, E. (2016). A discussion of weighting procedures in the presence of unit nonresponse. *Journal of Official Statistics* **32**, 129–145.

Kim, J. K. and Rao, J. N. K. (2009). Unified approach to linearization variance estimation from survey data after imputation for item nonresponse. *Biometrika* **96**, 917–932.

Kott, P. S. (2006). Using calibration weighting to adjust for nonresponse and undercoverage. *Survey Methodology* **32**, 133–142.

Kott, P. S. (2009). Calibration weighting: Combining probability samples and linear prediction models: *In Handbook of Statistics 29B, Sample Surveys: Inference and Analysis*. Pfeffermann, D. & Rao, C.R. (Eds.), Oxford, UK: Elsevier.

Kott, P. S. and Chang, T. (2010). Using calibration weighting to adjust for nonignorable unit nonresponse. *Journal of the American Statistical Association* **105**, 1265–1275.

Kott, P. S. and Liao, D. (2012). Providing double protection for unit nonresponse with a nonlinear calibration-weighting routine. *Survey Research Methods* **6**, 105–111.

Kott, P. S. and Liao, D. (2015). One step or two? Calibration weighting from a complete list frame with nonresponse. *Survey Methodology* **41**, 165–181.

Kott, P. and Liao, D. (2017). Calibration Weighting for Nonresponse that is Not Missing at Random: Allowing More Calibration than Response-Model Variables, *Journal of Survey Statistics and Methodology* **5**, 159—174.

Krewski, D., and Rao, J. N. K. (1981). Inference from stratified samples: properties of the linearization, Jackknife and balanced repeated replication methods. *The Annals of Statistics* **9**, 1010–1019.

Little, R. J. A. (1986). Survey Nonresponse Adjustments for Estimates of Means. *International Statistical Review* **54**, 139–157.

Molenberghs, G., Beunckens, C., and Sotto, C. (2008). Every missingness not at random model has a missingness at random counterpart with equal fit. *Journal of Royal Statistical Society B* **70**, 371–388.

Myers, J. A., Rassen, J. A., Gagne, J. J., Huybrechts, K. F., Schneeweiss, S., Rothman and K. J., Glynn, R. J. (2011). Effects of adjusting for instrumental variables on bias and precision of effect estimates. *American Journal of Epidemiology* **174**, 1213–1222.

Osier, G. (2012). Traitement de la non-réponse non-ignorable par calage généralisé : une simulation à partir de l'enquête Budget des Ménages au Luxembourg. *Actes des Journées de Méthodologie Statistique de l'Insee*, Paris, France.

Pearl, J. (2010). On a class of bias-amplifying variables that endanger effect estimates. In: Grünwald P, Spirtes P, editors. *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence Corvallis*, 425–432.

Rosen, I. (1972). Asymptotic theory for successive sampling with varying probabilities without replacement. *The Annals of Mathematical Statistics* **43**, 373–397.

Rubin, D.B. (1976). Inference and Missing Data. *Biometrika* **63**, 581–590.

Särndal, C. E. and Lundström, S. (2005). *Estimation in Surveys with Nonresponse.* New York: John Wiley and Sons.

Sautory, O. (2003). Calmar 2: a new version of the Calmar calibration adjustment program. Proceedings of the Statistics Canada Symposium, Ottawa, Canada.

Shao, J. and Steel, P. (1999). Variance estimation for survey data with composite imputation and nonnegligible sampling fractions. *Journal of the American Statistical Association* **94**, 254–265.

van der Vaart, A. (2000). *Asymptotic Statistics.* Cambridge University Press.

Wang, S., Shao, J. and Kim, J. K. (2014). An instrument variable approach for identification and estimation with nonignorable nonresponse. *Statistica Sinica* **24** 1097–1116.

Wooldridge, J. (2002). *Econometric Analysis of Cross Section and Panel Data.* MIT Press.

Wooldridge, J. (2016). Should instrumental variables be used as matching variables? *Research in Economics* **70**, 232–237.

A Proofs

A.1 Proof of Theorem 1

1. **Consistency of $\widehat{\boldsymbol{\lambda}}$.**

Let us define

$$\boldsymbol{\psi}_n(\boldsymbol{\lambda}) = \frac{1}{N}\left\{\sum_{k\in S} d_k R_k F(\boldsymbol{\lambda}^\top \boldsymbol{Z}_k)\boldsymbol{X}_k - \sum_{k\in P}\boldsymbol{X}_k\right\},$$
$$\boldsymbol{\psi}(\boldsymbol{\lambda}) = \mathbb{E}\left\{\left(R_k F(\boldsymbol{\lambda}^\top \boldsymbol{Z}_k) - 1\right)\boldsymbol{X}_k\right\}.$$

Let $\Psi_n(\boldsymbol{\lambda}) = -||\boldsymbol{\psi}_n(\boldsymbol{\lambda})||$ and $\Psi(\boldsymbol{\lambda}) = -||\boldsymbol{\psi}(\boldsymbol{\lambda})||$. Hence, $\widehat{\boldsymbol{\lambda}} \in \arg\max_{\boldsymbol{\lambda}\in\boldsymbol{\Lambda}} \Psi_n(\boldsymbol{\lambda})$. We check the conditions of Theorem 5.9 in van der Vaart (2000). First,

$$\mathbb{E}(\boldsymbol{\psi}_n(\boldsymbol{\lambda})|\mathcal{U}) = \frac{1}{N}\sum_{k\in P}\left[R_k F(\boldsymbol{\lambda}^\top \boldsymbol{Z}_k) - 1\right]\boldsymbol{X}_k,$$

which implies that $\mathbb{E}\{\boldsymbol{\psi}_n(\boldsymbol{\lambda}\} = \boldsymbol{\psi}(\boldsymbol{\lambda})$. Since $R_k F(\boldsymbol{\lambda}^\top \boldsymbol{Z}_k) - 1$ is bounded, it follows that $\left\{R_k F(\boldsymbol{\lambda}^\top \boldsymbol{Z}_k) - 1\right\}\boldsymbol{X}_k$ admits second-order moments. Then, $\mathbb{V}[\mathbb{E}(\boldsymbol{\psi}_n(\boldsymbol{\lambda})|\mathcal{U})]$ converges to 0. Also,

$$\mathbb{E}\left[\mathbb{V}\{\boldsymbol{\psi}_n(\boldsymbol{\lambda})|\mathcal{U}\}\right] = \mathbb{E}\left[\mathbb{V}\left\{\frac{1}{N}\sum_{k\in S} d_k R_k F(\boldsymbol{\lambda}^\top \boldsymbol{Z}_k)\boldsymbol{X}_k\middle|\mathcal{U}\right\}\right]. \tag{25}$$

By Assumption 3-(i), the right-hand side of (25) tends to 0. Then $\boldsymbol{\psi}_n(\boldsymbol{\lambda}) \to \boldsymbol{\psi}(\boldsymbol{\lambda})$ in L^2, and thus in probability.

Because $\boldsymbol{Z}_k$ has compact support and $\boldsymbol{\Lambda}$ is compact, there exists a compact interval I including with probability one the interval $[\min_{\boldsymbol{\lambda}\in\boldsymbol{\Lambda}} \boldsymbol{\lambda}^\top \boldsymbol{Z}_k, \max_{\boldsymbol{\lambda}\in\boldsymbol{\Lambda}} \boldsymbol{\lambda}^\top \boldsymbol{Z}_k]$. Moreover, $F(.)$ is uniformly continuous on I. Now, fix ε and let δ be such that for any $(a,b) \in I^2$, $|a-b| < \delta$ implies $|F(a) - F(b)| < \varepsilon$. Let C be such that $\|\boldsymbol{Z}_k\| \leq C$ with probability one. Consider balls of center $\boldsymbol{\lambda}_b$ and of radius δ/C for $\boldsymbol{\lambda}$. Then, for $\boldsymbol{\lambda}$ within such a ball, we get, by the triangular and Cauchy-Schwarz inequalities,

$$\begin{aligned}\|\boldsymbol{\psi}_n(\boldsymbol{\lambda}) - \boldsymbol{\psi}_n(\boldsymbol{\lambda}_b)\| &\leq \frac{1}{N}\sum_{k\in S} d_k R_k \left|F(\boldsymbol{\lambda}^\top \boldsymbol{Z}_k) - F(\boldsymbol{\lambda}_b^\top \boldsymbol{Z}_k)\right| \|\boldsymbol{X}_k\| \\ &\leq \frac{\varepsilon}{N}\sum_{k\in S} d_k R_k \|\boldsymbol{X}_k\|. \end{aligned} \tag{26}$$

Similarly, $\|\boldsymbol{\psi}(\boldsymbol{\lambda}) - \boldsymbol{\psi}(\boldsymbol{\lambda}_b)\| \leq \varepsilon \mathbb{E}(\|\boldsymbol{X}_k\|)$. Now, by assumption, $\boldsymbol{\Lambda}$ is compact. It can then be recovered by B balls of centers $\boldsymbol{\lambda}_b$ $(b = 1...B)$ and of radius δ/C. Then, using $\big|\,||a|| - ||b||\,\big| \leq ||a - b||$, we get

$$\begin{aligned}\sup_{\boldsymbol{\lambda}\in\boldsymbol{\Lambda}} |\Psi_n(\boldsymbol{\lambda}) - \Psi(\boldsymbol{\lambda})| &\leq \sup_{\boldsymbol{\lambda}\in\boldsymbol{\Lambda}} ||\boldsymbol{\psi}_n(\boldsymbol{\lambda}) - \boldsymbol{\psi}(\boldsymbol{\lambda})|| \\ &\leq \max_{b=1...B} \|\boldsymbol{\psi}_n(\boldsymbol{\lambda}_b) - \boldsymbol{\psi}(\boldsymbol{\lambda}_b)\| + \varepsilon \left\{ \mathbb{E}(\|\boldsymbol{X}_k\|) + \frac{1}{N} \sum_{k\in S} d_k R_k \|\boldsymbol{X}_k\| \right\}.\end{aligned}$$

The first term on the right-hand side converges in probability to zero by pointwise consistency of $\boldsymbol{\psi}_n(\boldsymbol{\lambda})$. By Assumption 3-(i) and reasoning as above, the term within brackets converges in probability to $\mathbb{E}\{(1 + R_k)\|\boldsymbol{X}_k\|\}$. Therefore, with probability tending to one, the left-hand side is smaller than $\varepsilon[1 + 2\mathbb{E}\{(1 + R_k)\|\boldsymbol{X}_k\|\}]$. Because ε was arbitrary, we have proved

$$\sup_{\boldsymbol{\lambda}\in\boldsymbol{\Lambda}} |\Psi_n(\boldsymbol{\lambda}) - \Psi(\boldsymbol{\lambda})| \xrightarrow{P} 0.$$

Hence, condition (i) in Theorem 5.9 of van der Vaart (2000) holds.

We now check condition (ii). First, by Assumptions 1 and 2,

$$\mathbb{E}\left[\left(R_k F(\boldsymbol{\lambda}_0^\top \boldsymbol{Z}_k) - 1\right) \boldsymbol{X}_k | \boldsymbol{Z}_k\right] = \left[F(\boldsymbol{\lambda}_0^\top \boldsymbol{Z}_k)\mathbb{E}\left(R_k|\boldsymbol{Z}_k\right) - 1\right] \mathbb{E}\left(\boldsymbol{X}_k|\boldsymbol{Z}_k\right) = 0.$$

Thus, $\boldsymbol{\psi}(\boldsymbol{\lambda}_0) = 0$ and $\Psi(\boldsymbol{\lambda}_0) = 0$. Suppose that there exists $\boldsymbol{\lambda}_1$ such that $\Psi(\boldsymbol{\lambda}_1) = 0$. Then $\boldsymbol{\psi}(\boldsymbol{\lambda}_1) = 0$ and because $\mathbb{E}(\boldsymbol{X}_k|R_k = 1, \boldsymbol{Z}_k) = \boldsymbol{\Gamma}\boldsymbol{Z}_k$, we obtain, by the law of iterated expectation,

$$\boldsymbol{\Gamma}\, \mathbb{E}\left[R_k \left\{F(\boldsymbol{\lambda}_0^\top \boldsymbol{Z}_k) - F(\boldsymbol{\lambda}_1^\top \boldsymbol{Z}_k)\right\} \boldsymbol{Z}_k\right] = 0.$$

Because the rank of $\boldsymbol{\Gamma}$ is equal to $\dim(\boldsymbol{Z}_k)$,

$$\mathbb{E}\left[R_k \left\{F(\boldsymbol{\lambda}_0^\top \boldsymbol{Z}_k) - F(\boldsymbol{\lambda}_1^\top \boldsymbol{Z}_k)\right\} \boldsymbol{Z}_k\right] = 0.$$

This, in turn, implies that

$$\mathbb{E}\left[R_k \left(F(\boldsymbol{\lambda}_0^\top \boldsymbol{Z}_k) - F(\boldsymbol{\lambda}_1^\top \boldsymbol{Z}_k)\right) \left(\boldsymbol{\lambda}_0^\top \boldsymbol{Z}_k - \boldsymbol{\lambda}_1^\top \boldsymbol{Z}_k\right)\right] = 0. \tag{27}$$

Now, because $F(.)$ is strictly increasing, we have

$$\left\{F(\boldsymbol{\lambda}_0^\top \boldsymbol{Z}_k) - F(\boldsymbol{\lambda}_1^\top \boldsymbol{Z}_k)\right\} \left(\boldsymbol{\lambda}_0^\top \boldsymbol{Z}_k - \boldsymbol{\lambda}_1^\top \boldsymbol{Z}_k\right) \geq 0 \text{ with equality iff } \boldsymbol{\lambda}_0^\top \boldsymbol{Z}_k - \boldsymbol{\lambda}_1^\top \boldsymbol{Z}_k = 0.$$

Hence, (27) implies that $(\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}_1)^\top \boldsymbol{Z}_k = 0$ almost surely. This and the fact that $\mathbb{E}(\boldsymbol{Z}_k \boldsymbol{Z}_k^\top)$ is nonsingular imply that $\boldsymbol{\lambda}_1 = \boldsymbol{\lambda}_0$.

Thus, $\Psi(\boldsymbol{\lambda}) = 0$ implies that $\boldsymbol{\lambda} = \boldsymbol{\lambda}_0$. Second, by the same argument following (26), we have, for any $\boldsymbol{\lambda}, \boldsymbol{\lambda}'$ such that $\|\boldsymbol{\lambda} - \boldsymbol{\lambda}'\| < \delta/C$,

$$\|\Psi(\boldsymbol{\lambda}) - \Psi(\boldsymbol{\lambda}')\| \leq \|\boldsymbol{\psi}(\boldsymbol{\lambda}) - \boldsymbol{\psi}(\boldsymbol{\lambda}')\| \leq \varepsilon \mathbb{E}\left(\|\boldsymbol{X}_k\|\right).$$

Hence, Ψ is continuous. Thus, for any $\varepsilon' > 0$,

$$\inf_{\boldsymbol{\lambda} \in \boldsymbol{\Lambda}: \|\boldsymbol{\lambda} - \boldsymbol{\lambda}_0\| \geq \varepsilon} \|\Psi(\boldsymbol{\lambda})\| = \min_{\boldsymbol{\lambda} \in \boldsymbol{\Lambda}: \|\boldsymbol{\lambda} - \boldsymbol{\lambda}_0\| \geq \varepsilon} \|\Psi(\boldsymbol{\lambda})\| > 0 = \|\Psi(\boldsymbol{\lambda}_0)\|$$

and condition (ii) in Theorem 5.9 of van der Vaart (2000) holds.

As a result, both conditions of this theorem are satisfied, and $\widehat{\boldsymbol{\lambda}}$ is consistent.

Consistency of $\widehat{t}_C$.

First,

$$\begin{aligned}\left(\widehat{t}_C - t_y\right)/N = &\frac{1}{N}\sum_{k \in S} d_k R_k F(\boldsymbol{\lambda}_0^\top \boldsymbol{Z}_k) Y_k - \frac{1}{N}\sum_{k \in P} Y_k \\ &+ \frac{1}{N}\sum_{k \in S} d_k R_k \left\{F(\widehat{\boldsymbol{\lambda}}^\top \boldsymbol{Z}_k) - F(\boldsymbol{\lambda}_0^\top \boldsymbol{Z}_k)\right\} Y_k. \end{aligned} \tag{28}$$

We now show that both terms on the right-hand side of (28), denoted by A_1 and A_2 hereafter, tend to zero in probability. For A_1, we use arguments similar to those used for the pointwise consistency of $\boldsymbol{\psi}_n(\lambda)$. We have

$$\mathbb{E}(A_1|\mathcal{U}) = \frac{1}{N}\sum_{k \in P} \left\{R_k F(\boldsymbol{\lambda}_0^\top \boldsymbol{Z}_k) - 1\right\} Y_k.$$

Moreover, by the law of iterated expectation and (10),

$$\mathbb{E}\left\{\left[R_k F(\boldsymbol{\lambda}_0^\top \boldsymbol{Z}_k) - 1\right] Y_k\right\} = \mathbb{E}\left\{\left[\mathbb{E}(R_k \mid \boldsymbol{Z}_k) F(\boldsymbol{\lambda}_0^\top \boldsymbol{Z}_k) - 1\right] \mathbb{E}\left[Y_k | \boldsymbol{Z}_k\right]\right\} = 0.$$

Hence, $\mathbb{E}(A_1) = 0$ and $\mathbb{V}\{\mathbb{E}(A_1|\mathcal{U})\}$ converges to 0. Moreover,

$$\mathbb{E}\left\{\mathbb{V}(A_1|\mathcal{U})\right\} = \mathbb{E}\left\{\mathbb{V}\left(\frac{1}{N}\sum_{k \in S} d_k R_k F(\boldsymbol{\lambda}_0^\top \boldsymbol{Z}_k) Y_k \middle| \mathcal{U}\right)\right\},$$

and the right-hand side converges to 0 by Assumption 3-(i). Thus, $\mathbb{V}(A_1)$ tends to 0 and A_1 converges to 0 in probability.

Turning to A_2, note that $F(.)$ is uniformly continuous on the compact set. Set $\varepsilon > 0$ and δ as above. By consistency of $\widehat{\boldsymbol{\lambda}}$, $||\widehat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_0|| < \delta/C$ with probability close to one. Then, by the Cauchy-Schwarz inequality, $\max_k ||(\widehat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_0)^\top \boldsymbol{Z}_k|| < \delta$, implying that

$$\max_k \left| F(\widehat{\boldsymbol{\lambda}}^\top \boldsymbol{Z}_k) - F(\boldsymbol{\lambda}_0^\top \boldsymbol{Z}_k) \right| < \varepsilon,$$

with a probability close to one. Hence, with such a probability,

$$|A_2| < \varepsilon \left(\frac{1}{N} \sum_{k \in S} d_k |Y_k| \right).$$

By Assumption 3-(i), the term into parentheses converges to $\mathbb{E}(|Y_k|)$. Therefore, A_2 converges to zero, and the result follows. □

2. **Linearization of $\widehat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_0$.**

Let us define $\widehat{\mathbf{G}}(\boldsymbol{\lambda}) = \frac{1}{N} \sum_{k \in S} d_k R_k F'(\boldsymbol{\lambda}^\top \boldsymbol{Z}_k) \boldsymbol{X}_k \boldsymbol{Z}_k^\top$. Then, by the first-order condition of (5) and the mean value theorem,

$$\begin{aligned} 0 &= \frac{\widehat{\mathbf{G}}(\widehat{\boldsymbol{\lambda}})^\top}{N} \left\{ \sum_{k \in S} d_k R_k F(\widehat{\boldsymbol{\lambda}}^\top \boldsymbol{Z}_k) \boldsymbol{X}_k - \sum_{k \in P} \boldsymbol{X}_k \right\} \\ &= \frac{\widehat{\mathbf{G}}(\widehat{\boldsymbol{\lambda}})^\top}{N} \left\{ \sum_{k \in S} d_k R_k F(\boldsymbol{\lambda}_0^\top \boldsymbol{Z}_k) \boldsymbol{X}_k - \sum_{k \in P} \boldsymbol{X}_k \right\} + \widehat{\mathbf{G}}(\widehat{\boldsymbol{\lambda}})^\top \widehat{\mathbf{G}}(\widetilde{\boldsymbol{\lambda}})(\widehat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_0), \end{aligned} \tag{29}$$

where $\widetilde{\boldsymbol{\lambda}} = \widetilde{t}\boldsymbol{\lambda}_0 + (1 - \widetilde{t})\widehat{\boldsymbol{\lambda}}$ for some $\widetilde{t} \in [0, 1]$.

Because $F'(\boldsymbol{\lambda}_0^\top \boldsymbol{Z}_k)$ is bounded, we have, by the same arguments as when showing $\boldsymbol{\psi}_n(\boldsymbol{\lambda}) \xrightarrow{P} \boldsymbol{\psi}(\boldsymbol{\lambda})$,

$$\widehat{\mathbf{G}}(\boldsymbol{\lambda}_0) \xrightarrow{P} \mathbf{G}.$$

Now fix $\varepsilon > 0$. F' is continuous, and therefore uniformly continuous on the interval I defined above. Thus, there exists $\delta_2 > 0$ such that for any $(a, b) \in I^2$, $|a - b| < \delta_2$ implies $|F'(a) - F'(b)| < \varepsilon$. By Step 1, with a large probability, $\|\widehat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_0\| < \delta_2/C$

and $\|\widetilde{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_0\| < \delta_2/C$. Then, by the triangular and Cauchy-Schwarz inequality, with a large probability,

$$\left\|\widehat{\mathbf{G}}(\widehat{\boldsymbol{\lambda}}) - \widehat{\mathbf{G}}(\boldsymbol{\lambda}_0)\right\| \leq \frac{\varepsilon}{N}\sum_{k\in S} d_k R_k \|\boldsymbol{X}_k \boldsymbol{Z}_k^\top\|.$$

The same holds if $\widehat{\boldsymbol{\lambda}}$ is replaced with $\widetilde{\boldsymbol{\lambda}}$. Because $\boldsymbol{Z}_k$ is bounded, the right-hand side converges to 0 by Assumption 3-(i). Hence, both $\widehat{\mathbf{G}}(\widehat{\boldsymbol{\lambda}})$ and $\widehat{\mathbf{G}}(\widetilde{\boldsymbol{\lambda}})$ converge in probability to $\mathbf{G}$. This and (29) imply that

$$\widehat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_0 = -\frac{(\mathbf{G}^\top\mathbf{G})^{-1}\mathbf{G}^\top}{N}\left\{\sum_{k\in S} d_k R_k F(\boldsymbol{\lambda}_0^\top \boldsymbol{Z}_k)\boldsymbol{X}_k - \sum_{k\in P}\boldsymbol{X}_k\right\}\{1 + o_P(1)\}.$$

Asymptotic normality of $\widehat{t}_C$.

First, we have

$$\begin{aligned}
\frac{\widehat{t}_C}{N} &= \frac{1}{N}\sum_{k\in S} d_k R_k F(\boldsymbol{\lambda}_0^\top \boldsymbol{Z}_k) Y_k + \frac{1}{N}\sum_{k\in S} d_k R_k \left\{F(\widehat{\boldsymbol{\lambda}}^\top \boldsymbol{Z}_k) - F(\boldsymbol{\lambda}_0^\top \boldsymbol{Z}_k)\right\} Y_k \\
&= \frac{1}{N}\sum_{k\in S} d_k R_k F(\boldsymbol{\lambda}_0^\top \boldsymbol{Z}_k) Y_k - \left\{\frac{1}{N}\sum_{k\in S} d_k R_k Y_k F'(\overline{\boldsymbol{\lambda}}^\top \boldsymbol{Z}_k)\boldsymbol{Z}_k^\top\right\} \\
&\quad \times \frac{(\mathbf{G}^\top\mathbf{G})^{-1}\mathbf{G}^\top}{N}\left\{\sum_{k\in S} d_k R_k F(\boldsymbol{\lambda}_0^\top \boldsymbol{Z}_k)\boldsymbol{X}_k - \sum_{k\in P}\boldsymbol{X}_k\right\}\{1 + o_P(1)\},
\end{aligned}$$

where $\overline{\boldsymbol{\lambda}} = \bar{t}\boldsymbol{\lambda}_0 + (1-\bar{t})\widehat{\boldsymbol{\lambda}}$ for some $\bar{t} \in [0,1]$. By the same argument as above,

$$\frac{1}{N}\sum_{k\in S} d_k R_k Y_k F'(\overline{\boldsymbol{\lambda}}^\top \boldsymbol{Z}_k)\boldsymbol{Z}_k^\top \xrightarrow{P} \mathbb{E}(\rho_k Y_k \boldsymbol{Z}_k^\top).$$

Hence, since $\boldsymbol{\gamma} = \mathbf{G}(\mathbf{G}^\top\mathbf{G})^{-1}\mathbb{E}(\rho_k Y_k \boldsymbol{Z}_k^\top)$ and $W_k = (R_k F(\boldsymbol{\lambda}_0^\top \boldsymbol{Z}_k) - 1)(Y_k - \boldsymbol{\gamma}^\top \boldsymbol{X}_k)$, we get

$$\widehat{t}_C - t_y = \left\{\sum_{k\in S} d_k(W_k + Y_k - \boldsymbol{\gamma}^\top \boldsymbol{X}_k) - \sum_{k\in P} W_k + Y_k - \boldsymbol{\gamma}^\top \boldsymbol{X}_k + \sum_{k\in P} W_k\right\}\{1 + o_P(1)\}. \tag{30}$$

To prove the result, we now check the conditions of Theorem 2 in Chen and Rao (2007). Note first that $\mathcal{U}$ plays the role of $\mathcal{B}_n$ in their theorem, $\sum_{k\in S} d_k(W_k + Y_k - \boldsymbol{\gamma}^\top \boldsymbol{X}_k) - \sum_{k\in P} W_k + Y_k - \boldsymbol{\gamma}^\top \boldsymbol{X}_k$ corresponds to U_n, $\sum_{k\in P} W_k$ corresponds to V_n, $\sigma_{2N} \equiv$

$\mathbb{V}\left\{\sum_{k\in S} d_k(W_k + Y_k - \boldsymbol{\gamma}^\top \boldsymbol{X}_k)\middle|\mathcal{U}\right\}^{1/2}$ corresponds to σ_{2n} and $\sigma_{1N} \equiv \mathbb{V}\left(\sum_{k\in P} W_k\right)^{1/2}$ corresponds to σ_{1n}.

Then, by Assumptions 1 and 2,

$$\mathbb{E}(W_k|\boldsymbol{Z}_k) = \left[\mathbb{E}(R_k|\boldsymbol{Z}_k)F(\boldsymbol{\lambda}_0^\top \boldsymbol{Z}_k) - 1\right]\mathbb{E}\left(Y_k - \boldsymbol{\gamma}^\top \boldsymbol{X}_k|\boldsymbol{Z}_k\right) = 0.$$

Hence, $\mathbb{E}(W_k) = 0$ and by the central limit theorem,

$$\sigma_{1N}^{-1}\left(\sum_{k\in P} W_k\right) \xrightarrow{d} \mathcal{N}(0,1).$$

Also, $\sum_{k\in P} W_k$ is $\mathcal{U}$-measurable. Hence, their condition 1 holds.

Next,

$$\mathbb{E}\left\{\sum_{k\in S} d_k(W_k + Y_k - \boldsymbol{\gamma}^\top \boldsymbol{X}_k) - \sum_{k\in P} W_k + Y_k - \boldsymbol{\gamma}^\top \boldsymbol{X}_k\middle|\mathcal{U}\right\} = 0$$

and Condition (1) in Chen and Rao (2007) holds by Assumption 3-(iii) and Polya's theorem (see the remark below Theorem 2 in Chen and Rao, 2007). Thus, their condition 2 holds. Finally their condition 3 holds by Assumption 3-(ii).

Therefore,

$$\frac{\sum_{k\in S} d_k(W_k + Y_k - \boldsymbol{\gamma}^\top \boldsymbol{X}_k) - \sum_{k\in P} W_k + Y_k - \boldsymbol{\gamma}^\top \boldsymbol{X}_k + \sum_{k\in P} W_k}{\sqrt{\sigma_{1N}^2 + \sigma_{2N}^2}} \xrightarrow{d} \mathcal{N}(0,1).$$

The result follows by (30) and Slutsky's lemma. □

A.2 Proof of Theorem 2

1. By assumption, a solution to (14) exists. By the same reasoning as the one used to show $\boldsymbol{\lambda}_0 = \boldsymbol{\lambda}_1$ in the proof of Theorem 1, the solution is unique. Moreover, still reasoning as in the first step of the proof of Theorem 1, we have $\widehat{\boldsymbol{\lambda}} \xrightarrow{P} \boldsymbol{\lambda}_\infty$. □

2. We can decompose the total error of estimation of $\widehat{t}_C$ as in (13). Using Assumption 3 and the same arguments as in the second step of the proof of Theorem 1, the first three terms of the right-hand side converge to 0 in probability. On the other hand, the

fourth term on the right-hand side does not tend to zero in probability. By the law of large number and Assumption 3-(i), it converges towards

$$\mathbb{E}\left[R_k \ \left\{F(\boldsymbol{\lambda}_\infty^\top \boldsymbol{Z}_k) - F(\boldsymbol{\lambda}_0^\top \boldsymbol{Z}_k)\right\} \ Y_k\right].$$

This in turn can be rewritten as:

$$\mathbb{E}\left\{R_k \ \left\{F(\boldsymbol{\lambda}_\infty^\top \boldsymbol{Z}_k) - F(\boldsymbol{\lambda}_0^\top \boldsymbol{Z}_k)\right\} \ Y_k\right\} \ = \ \mathbb{E}\left(f_k \ R_k \ Y_k \ \boldsymbol{Z}_k^\top\right) \ (\boldsymbol{\lambda}_\infty - \boldsymbol{\lambda}_0),$$

where f_k is defined in (16). Next, we prove (15). We have

$$\begin{aligned} -\mathbb{E}\left[f_k F(\boldsymbol{\lambda}_0^\top \boldsymbol{Z}_k)^{-1} X_k \boldsymbol{Z}_k^\top (\boldsymbol{\lambda}_\infty - \boldsymbol{\lambda}_0)\right] &= \mathbb{E}\left\{\left(1 - \frac{F(\boldsymbol{\lambda}_\infty^\top \boldsymbol{Z}_k)}{F(\boldsymbol{\lambda}_0^\top \boldsymbol{Z}_k)}\right) X_k\right\} \\ &= \mathbb{E}\left\{F(\boldsymbol{\lambda}_\infty^\top \boldsymbol{Z}_k)\,(R_k X_k - \mathbb{E}(R_k|Z_k) X_k)\right\} \\ &= \mathbb{E}\left\{F(\boldsymbol{\lambda}_\infty^\top \boldsymbol{Z}_k)\,\mathbb{C}ov\left(\boldsymbol{X}_k, \ R_k \mid \boldsymbol{Z}_k\right)\right\}, \end{aligned}$$

where the second equality comes from the nonresponse model and (14), and the third equality from the law of iterated expectation. This shows (15), which in turn implies that

$$\begin{aligned} & \mathbb{E}\left\{R_k \ \left\{F(\boldsymbol{\lambda}_\infty^\top \boldsymbol{Z}_k) - F(\boldsymbol{\lambda}_0^\top \boldsymbol{Z}_k)\right\} \ Y_k\right\} \\ = \ & - \left\{\left\{\mathbb{E}(f_k R_k \boldsymbol{Z}_k \boldsymbol{X}_k^\top)\right\}^{-1} \ \mathbb{E}(f_k R_k \boldsymbol{Z}_k Y_k)\right\}^\top \ \mathbb{E}\left\{F(\boldsymbol{\lambda}_\infty^\top \boldsymbol{Z}_k)\,\mathbb{C}ov\left(\boldsymbol{X}_k, \ R_k \mid \boldsymbol{Z}_k\right)\right\}. \ \square \end{aligned}$$