Instrument Variable Selection in the Calibration Estimator under Nonresponse

Thomas Laitila

Örebro University and Statistics Sweden

Department of Statistics, Örebro University, SE-701 82 Örebro
Phone: +46 70 1905 713, E-mail: thomas.laitila@oru.se

Abstract

The calibration estimator suggested by [1] uses sample and/or population level information on a set of auxiliary variables to adjust the design weights in an effort to reduce nonresponse bias. The version of the estimator usually applied is the standard weight formulation where the instrument variable vector is defined as the auxiliary variable vector. In this paper the objective is to reduce bias by choosing an instrument variable vector different from the auxiliary variable vector. A condition on the instrument vector for approximately zero bias is derived. The condition gives a connection between the [1] calibration estimator and the procedure proposed by [2] for dealing with selection bias in regression analysis. This in turn suggests an estimator and instrument vectors giving approximately zero bias. Results from a simulation study illustrates the finite sample properties of the new estimator.

1. Introduction

Nonresponse in sample surveys is an increasing problem and different methods for adjusting weights in the estimation stage have been proposed (e.g. [3]). One part of the literature on calibration estimation is devoted to its potentials to adjust design weights for nonresponse bias ([1], [4-7]). Although the calibration approach can be used to define estimators which are approximately unbiased and consistent under mild conditions in the full sample case, calibration estimators are biased and inconsistent under nonresponse in general.

Consistency of calibration estimators can be obtained in particular cases. One is considered by [4-6] who study calibration estimators when the unit's response probabilities are known functions of a known model variable vector and an unknown parameter vector.

One choice of functional form is to use the inverse of the response probability function, interpreting the response set as an outcome of two-phase sampling.

This paper focuses on the consistency properties of the linear calibration estimator under the quasi-randomization framework ([8]). The linear calibration estimator implies a restrictive form of the response probability function, whereby interest here is on selection of appropriate instrument variables for consistency. A condition on the instrument vector for consistency is given and a modification of the linear calibration estimator is suggested.

Consistency of the new estimator discussed and, two examples of the estimator are presented and their finite sample properties are illustrated with results from a simulation study.

The modified linear calibration estimator is defined in the next section and a condition on the instrument vector for consistency is given. Section 3 contains results from the simulation study and final comments contained in the closing section.

2. The calibration estimator

Consider a fixed, finite population U of N units and a non-random, scalar study variable y_k, $k \in U$. A probability sample $s \subset U$ with expected sample size $n(N)$ is selected from the population, using a probability sampling design $p(s)$, with the purpose of estimating the population total $Y = \sum_U y_k$.

Due to nonresponse observations are only obtained for a subset of the sample $r \subset s$. Whether the sampled units respond or not are assumed results of a random trial beyond control of the researcher. The conditional probability of a response set r given a sample s is denoted $q(r|s)$. The first order inclusion probability of unit $k \in U$ is denoted π_k, $d_k = \pi_k^{-1}$ denotes the corresponding design weight and $\theta_k = \Pr(k \in r | k \in s, s)$ denotes the response probability.

Let x_k denote a column vector of non-random auxiliary variables satisfying the unity condition $\mu^t x_k = 1$ for some constant vector μ. This condition is satsisfied if e.g. x_k includes a constant term. The auxiliary vector x_k is assumed known for all units in r and its population total is denoted as $X = \sum_U x_k$. Here the population totals may either be known or estimated and the vector $\tilde{X}$ is used to denote the vector X where some or all elements are replaced by estimates. In addition to x_k an instrument vector z_k of the same dimension as x_k is assumed known for all units in the response set *r*. (Note $z_k = x_k$ is one option.)

With these definitions the population regression coefficient vector $B_U(x) = \left(\sum_U x_k x_k^t\right)^{-1} \sum_U x_k y_k$ yields the population regression errors $e_k = y_k - x_k^t B_U(x)$. These errors sum to zero and $Y = X^t B_U(x)$.

The calibration estimator suggested is

$$\hat{Y}_C = \sum_r w_k y_k \tag{1}$$

where the calibration weights w_k are defined by the system

$$\begin{aligned} w_k &= d_k v_k \\ v_k &= \lambda_r^t z_k \\ \tilde{X} &= \sum_r w_k x_k \end{aligned}$$

This system yields the calibration weights

$$w_k = \tilde{X}^t \left(\sum_r d_k z_k x_k^t\right)^{-1} d_k z_k$$

This calibration estimator equals the calibration estimator suggested by [1] when the instrument vector satisfies a unity condition $q^t z_k = 1$ for some vector *q*. This assumption is not made here since it restricts the prospects of finding an instrument vector yielding a consistent calibration estimator.

One example of the estimator (1) is $z_k = \phi_k x_k$ where $\phi_k = \theta_k^{-1}$. Then $\sum_U \theta_k z_k e_k = \sum_U x_k e_k = 0$ by definition of e_k and the calibration estimator (1) equals

$$\hat{Y}_C = \sum_r \tilde{X}^t \left(\sum_r \phi_k x_k x_k^t\right)^{-1} \phi_k x_k y_k = \tilde{X}^t \hat{B}_r(\phi x)$$

where $\hat{B}_r(\phi x) = \left(\sum_r \phi_k x_k x_k^t\right)^{-1} \sum_r \phi_k x_k y_k$. This example corresponds to the response propensity GREG estimator.

Another example is obtained by adapting Heckman's sample selection model ([2]). For the reponse set, consider an assisting model of the form $y_k = x_k^t B_U + \eta_k + \omega_k$, where η_k is defined as a "systematic" component such that $\sum_U \theta_k x_k \eta_k = \sum_U \theta_k x_k e_k$, and ω_k is an "irregular" component such that $\sum_U \theta_k x_k \omega_k = 0$ and $\sum_U \theta_k \eta_k \omega_k = 0$. For the population, define the instrument $z_k = x_k - \eta_k \delta$, where $\delta = \left(\sum_U \theta_k \eta_k^2\right)^{-1} \sum_U \theta_k \eta_k x_k$ is a vector with population fits of θ weighted, through the origin, LS regressions of elements in x_k on η_k. (Note here that δ is a vector with slope coefficients, one for each element in x_k when regressed on η_k.) These instruments have the property $\sum_U \theta_k z_k(\eta_k + \omega_k) = 0$ whereby $\hat{B}_r(z) = \left(\sum_r z_k x_k^t\right)^{-1} \sum_r z_k y_k$ is consistent for $B_U(x)$.

If the instrument vector z_k satisfies the asymptotic condition $p\lim_{N\to\infty}(1/N)\sum_{U_N} \theta_k z_k e_k = 0$ for a sequence of populations $(U_1 \subset U_2 \subset U_3 \cdots)$ the calibration estimator (1) can be shown consistent under some additional mild conditions. However, in practice the instrument vectors z_k are usually not known and replaced by estimates $\tilde{z}_k$. Consistency of the estimator (1) with estimated instruments can be shown under the assumption $\max_{k\in U}\left\|\tilde{z}_k - z_k\right\| < \mathrm{M}\omega$ where M is a finite positive constant, and $\omega = o_p(1)$, as $N \to \infty$, is a random scalar term.

Let $\hat{Y}_C(\tilde{z})$ denote the estimator (1) with estimated instruments. Then we have the following proposition:

Proposition (Consistency of $\hat{Y}_C(\tilde{z})$ *)*

Assume $p\lim_{N\to\infty}(1/N)\sum_{U_N} \theta_k z_k e_k = 0$, $\max_{k\in U}\left\|\tilde{z}_k - z_k\right\| < \mathrm{M}\omega$ where $0 < \mathrm{M} < \kappa < \infty$ and $p\lim_{N\to\infty} \omega = 0$, and $p\lim_{N\to\infty}(\tilde{X} - X)/N = 0$. Adding appropriate conditions on $p(s)$ and $q(r|s)$, and with bounded values on y_k x_k and z_k, the calibration estimator (1) is consistent for Y, i.e. $p\lim_{N\to\infty}(\hat{Y}_C(\tilde{z}) - Y)/N = 0$.

5. Numerical Illustration

The calibration estimator (1) is here illustrated with two examples of instrument vectors. The first is the response propensity GREG estimator with the response probability specified by a normal cdf $\theta_k = \Phi(u_k^t\alpha)$. Here u_k is a real valued bounded vector ($\left\|u_k\right\| < \kappa < \infty$) of variables and α is a corresponding real valued vector of unknown, fixed parameters. For application the estimated instruments $\tilde{z}_k = \Phi(u_k^t\tilde{\alpha})^{-1} x_k$ are calculated with the Probit ML estimate $\tilde{\alpha}$.

In the second example the instruments are specified with $z_k = x_k - \eta_k\delta$ and $\eta_k = f(u_k^t\alpha)/\Phi(u_k^t\alpha)$, the ratio of the standard normal pdf to its cdf. Replacing for the Probit

ML estimator $\tilde{\alpha}$ and the LS estimates $\tilde{\delta} = \left(\sum_r \eta_k^2\right)^{-1} \sum_r \eta_k x_k$ yields the instruments $\tilde{z}_k = x_k - \eta_k \tilde{\delta}$. The estimator $\hat{B}_r(\tilde{z}) = \left(\sum_r \tilde{z}_k x_k^t\right)^{-1} \sum_r \tilde{z}_k y_k$ is then equal to the two-step estimator suggested by [2].

A small simulation study is used to illustrate the empirical properties of the estimator (1) based on the instruments defined in this section. For comparison the Särndal and Lundström [1] estimator is also included. Population data is simulated from the following model:

$$y = \beta_0 + \beta_1 x + \varepsilon$$
$$R^* = \alpha_0 + \alpha_1 y + \alpha_2 u + \varphi$$

where y represents the study variable, x and u are auxiliary variables with known population totals and R^* is a variable generating a response if $R^* > 0$ and a non-response if not. The variables x and u are independently generated from uniform distributions in $(0,2\sqrt{3})$, yielding $E(x) = E(u) = \sqrt{3}$ and $V(x) = V(u) = 1$. The variables ε and φ are independently generated from distributions with zero means and variances 1. The parameter α_1 is set to control the correlation ρ between $\chi = \alpha_1 \varepsilon + \varphi$ and ε, governing the strength of the relation between the study variable and the nonresponse mechanism. The parameters β_1 and α_2 are used to control the population R^2s in the regression model for the study variable and the regression of R^* on x and u. Finally, $\beta_0 = 5$ while α_0 is used to control the response rate. The following two sets of population models are used in the simulations.

i) *Population model Norm(ε)/Norm(φ):* The population R^2s in the regression model for the study variable and in the regression R^* on x and u, are both 0.5. The variables ε and φ are both generated from normal distributions. Expected response rates are 60%.

ii) *Population model BinU(ε)/U(φ):* As in *i)* with ε and φ generated from uniform distributions and with a dichotomized study variable *1*(*y*>6,5). Response rates are around 58%.

For each of the two population models, finite populations $U_1 \subset U_2 \subset U_3$ of sizes *N*=2000, 8000, 15000 units are generated. For the generated populations, samples of sizes *n*=*N*/10 are drawn with SRS, without replacement, and response sets are defined using the generated values on R^*. The study variable and the auxiliary variables are kept fixed in the population, from which the sample is drawn, while the response indicator is randomly generated in each replication for the units drawn to the sample. For each generated population, samples and response sets are replicated 1000 times.

The linear calibration estimator ([1]) is applied with standard weihgts (i.e. $z_k = x_k$) in two versions; one using the auxiliary vector $(1, x)$ (denoted as $\hat{Y}_W(x1)$ below) and one using the auxiliary vector $(1, x, u)$ ($\hat{Y}_W(x2)$). For the estimator (1), the auxiliary vector $(1, x)$ is used and the instruments defined above are calculated using $\hat{\phi} = 1/\Phi(\hat{\alpha}_{s0} + \hat{\alpha}_{s1}x + \hat{\alpha}_{s2}u)$ and $\hat{\eta} = \eta(\hat{\alpha}_{s0} + \hat{\alpha}_{s1}x + \hat{\alpha}_{s2}u)$, respectively. These estimators are below denoted as $\hat{Y}_C(z1)$ and $\hat{Y}_C(z2)$. For all estimators the variance is estimated using

$$\hat{V} = \sum\sum_r (d_k d_l - d_{kl}) v_k(\tilde{z}) v_l(\tilde{z}) \hat{e}_k \hat{e}_l + \sum_r d_k v_k(\tilde{z})(v_k(\tilde{z}) - 1)\hat{e}^2 \qquad (2)$$

where $\hat{e}_k = y_k - x_k^t \hat{B}_r(\tilde{z})$ and $v_k(\tilde{z}) = \lambda^t \tilde{z}_k$.

Table 1 includes simulation results under the Norm(ε)/Norm(φ) model, meaning that the assumptions of the model considered by [2] are satisfied. The calibration estimator $\hat{Y}_C(z2)$ is therefore expected to perform well with regard to bias.

In the case of $\rho = 0$, implying independence between the response indicator and the study variable, all estimators in Table 1 have biases tending to zero when the sample size increases. The standard deviations for all the estimators also decrease with the sample size. The results observed for all estimators when $\rho = 0$ are expected patterns of consistent estimators.

Also as expected, when $\rho = 0.3$ this "consistency pattern" is only observed for the estimator $\hat{Y}_C(z2)$. For the other estimators, the standard deviations decrease with the sample size, but the biases do not. Estimated biases for the $\hat{Y}_W(x1)$ estimator are around 2 percent of the population mean. Note that the $\hat{Y}_W(x2)$ estimator, which is utilizing both auxiliary variables are associated with larger bias estimates than $\hat{Y}_W(x1)$.

Table 1: Simulated bias and st.dev (in parenthesis) of the calibration estimators $\hat{\bar{Y}}_W = \hat{Y}_W / N$ and $\hat{\bar{Y}}_C = \hat{Y}_C N$. Population model Norm($\varepsilon$)/Norm($\varphi$) with populations means between 6.71 – 6.73. (1000 replications.)

		Sample/Population size (*n/N*)		
Estimator[a)]	ρ [b)]	200/2000	800/8000	1500/15000
$\hat{Y}_W(x1)$	0 0.3	-.013 (.092) .124 (.092)	.009 (.045) .153 (.045)	.001 (.032) .144 (.033)
$\hat{Y}_W(x2)$	0 0.3	.001 (.111) .204 (.107)	-.001 (.055) .216 (.053)	.007 (.040) .217 (.040)
$\hat{Y}_C(z1)$	0 0.3	.002 (.129) .223 (.125)	.003 (.062) .230 (.061)	.000 (.045) .229 (.045)
$\hat{Y}_C(z2)$	0 0.3	-.041 (.140) -.043 (.147)	.024 (.066) .026 (.069)	-.006 (.046) -.008 (.048)

[a)] Auxiliary/Instrument vectors: *x*1=(1 *x*)^t, *x2*=(1 *x u*)^t, *z*1=$\hat{\phi}x1$, *z*2= *x*1-$\hat{\delta}\hat{\eta}$. [b)] $\rho = Corr(\chi, \varepsilon)$.

Table 2: Simulated bias and st.dev (in parenthesis) of the calibration estimators $\hat{\bar{Y}}_W = \hat{Y}_W / N$ and $\hat{\bar{Y}}_C = \hat{Y}_C N$. Population model BinU($\varepsilon$)/U($\varphi$) with population means between 0.55 – 0.57. (1000 replications.)

		Sample/Population size (*n/N*)		
Estimator[a)]	ρ [b)]	200/2000	800/8000	1500/15000
$\hat{Y}_W(x1)$	0 0.3	-.005 (.036) .038 (.038)	-.000 (.018) .044 (.019)	-.001 (.013) .043 (.014)
$\hat{Y}_W(x2)$	0 0.3	-.008 (.041) .050 (.044)	.002 (.021) .062 (.022)	-.001 (.015) .060 (.016)
$\hat{Y}_C(z1)$	0 0.3	-.003 (.046) .066 (.053)	-.000 (.023) .071 (.026)	-.000 (.016) .071 (.019)
$\hat{Y}_C(z2)$	0 0.3	-.002 (.055) .002 (.061)	-.003 (.028) -.001 (.031)	-.002 (.020) -.000 (.022)

[a-b)] See Table 1.

The consistency property of the estimator $\hat{Y}_C(z2)$ comes at the price of a larger variance then the other estimators. The RMSE (not reported) of $\hat{Y}_C(z2)$ is decreasing with *N* but is around 50% higher in comparison with $\hat{Y}_W(x1)$ when $\rho = 0$. For $\rho = 0.3$ the picture is different and $\hat{Y}_C(z2)$ is associated with the smallest RMSE estimates for all sample sizes.

Table 3: Simulated bias of variance estimators and coverage rates of 95 percent confidence intervals (% in parenthesis). Population model Norm(ε)/Norm(φ). (1000 replications.)

Estimator of variance for[a]	ρ [b]	Sample/Population size (*n/N*)		
		200/2000	800/8000	1500/15000
$\hat{Y}_W(x1)$	0	-.027 (94)	.026 (94)	.062 (96)
	0.3	-.023 (71)	.019 (8)	.028 (1)
$\hat{Y}_W(x2)$	0	-.051(94)	-.002 (96)	.006 (95)
	0.3	-.039 (50)	.020 (2)	-.039 (0)
$\hat{Y}_C(z1)$	0	-.103 (93)	-.048 (94)	-.048 (95)
	0.3	-.058 (51)	-.012 (3)	-.039 (0)
$\hat{Y}_C(z2)$	0	.055 (94)	.096 (93)	.194 (96)
	0.3	.087 (95)	.099 (94)	.216 (96)

[a-b] See Table 1.

Table 4: Simulated bias of variance estimators and coverage rates of 95 percent confidence intervals (% in parenthesis). Population model BinU(ε)/U(φ). (1000 replications.)

Estimator of variance for[a]	ρ [b]	Sample/Population size (*n/N*)		
		200/2000	800/8000	1500/15000
$\hat{Y}_W(x1)$	0	-.015 (95)	.006 (95)	.006 (95)
	0.3	-.015 (84)	-.013 (38)	-.045 (12)
$\hat{Y}_W(x2)$	0	-.021 (94)	.003 (94)	-.014 (94)
	0.3	-.038 (78)	-.042 (19)	-.067 (4)
$\hat{Y}_C(z1)$	0	-.093 (94)	-.026 (95)	.005 (95)
	0.3	-.136 (74(	-.036 (18)	-.040 (3)
$\hat{Y}_C(z2)$	0	.075 (96)	.058 (96)	.071 (96)
	0.3	.115 (96)	.092 (97)	.091 (96)

[a-b] See Table 1.

Although the assumptions in the model considered by [2] are not satisfied in the BivU(ε)/U(φ) case considered in Table 2, the results show a similar pattern as the one in Table 1. All estimators have small and negligible bias estimates in the $\rho = 0$ case. For $\rho = 0.3$ however, only the calibration estimator $\hat{Y}_C(z2)$ has negligible bias estimates.

Tables 3 and 4 contain relative bias estimates of the variance estimator (2) for the different calibration estimators considered in Tables 1 and 2. The tables also contain estimated coverage rates of 95 percent confidence intervals of the population total. Regarding relative bias of the variance estimators, results show no explicit pattern except for the $\hat{Y}_C(z2)$ estimator whose variance is overestimated in general. The bias is particularly large in the *n/N*=1500/15000 case.

Although the variance of $\hat{Y}_C(z2)$ is overestimated by the variance estimator proposed, calculated confidence intervals have an appropriate coverage level in all cases considered in tables 3 and 4. This is not observed for the other estimators, where the coverage rates are heavily distorted in the $\rho = 0.3$ case.

6. Final comments

One important feature of the linear calibration estimator, which makes it a popular alternative, is its simplicity in calculation. Not only are the weights easily calculated, they are also the same for different study variables if the response sets are the same. Given calculated instrument vectors, the calibration estimator (1) has the same features of simplicity in calculation and with weights being the same for different study variables.

The implementation of the modified calibration estimator using the sample selection model for deriving instrument variables gives a link between calibration estimation and the Heckman two-step estimator ([2]). Using the Frisch-Waugh-Lowell Theorem in [9, p. 19], the Heckman two-step estimator can be rewritten as an instrument variable regression estimator, where the instruments are formed as in the numerical illustrations. Thus, with those instruments the modified calibration estimator can be implemented by calculating the Heckman two-step estimator using design weighted LS regression in the second step.

Instruments based on the sample selection model showed some robustness against departure from the assumptions underlying the instrument calculations. Further developments are desired where either alternative approaches for construction of instruments are considered or where instruments are derived from e.g. semi-parametric sample selection models (e.g. [10]). In particular it is of interest to develop instruments in cases with categorical response variables.

References

[1] Särndal, C.E. and S. Lundström (2005). *Estimation in Surveys with Nonresponse*. Wiley, New York.

[2] Heckman, J.J. (1979). Sample selection as a specification error, *Econometrica* **47:1**, 153-161.

[3] Brick, J.M. (2013). Unit nonresponse and weighting adjustments: A critical review, *Journal of Official Statistics*, **29:3**, 329-353.

[4] Lundström, S. and C.E. Särndal (1999). Calibration as a standard method for treatment of nonresponse, *Journal of Official Statistics*, **15:2**, 305-327.

[5] Kott, P.S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors, *Survey Methodology*, **32:2**, 133-142.

[6] Chang, T. and P.S. Kott (2008). Using calibration weighting to adjust for nonresponse under a plausible model, *Biometrika*, **95:3**, 555-571.

[7] Kott, P.S. and T. Chang (2010). Using calibration weighting to adjust for nonignorable unit nonresponse, *Journal of the American Statistical Association*, **105:491**, 1265-1275.

[8] Oh, H.L. and F.J. Scheuren (1983). Weighting adjustment for unit nonresponse. In: Madow, W.G, Olkin, I. and D.B. Rubin (Eds.), *Incomplete Data in Sample Surveys:Vol 2*, Academic Press, New York, pp. 143 – 184.

[9] Davidson, R. and J.G. MacKinnon (1993). *Estimation and Inference in Econometrics*, Oxford University Press, New York.

[10] Martins, M.F.O. (2001). Parametric and semiparametric estimation of sample selection models: An empirical application to the female labour force in Portugal, *Journal of Applied Econometrics*, **16**, 23-39.