Research and Development – Methodology reports from Statistics Sweden 2012:3

Statistiska centralbyrån Statistics Sweden

Consistent Calibration Estimation under Nonresponse

Thomas Laitila

The series entitled "**Research and Development** – **Methodology Reports from Statistics Sweden**" presents results from research activities within Statistics Sweden. The focus of the series is on development of methods and techniques for statistics production. Contributions from all departments of Statistics Sweden are published and papers can deal with a wide variety of methodological issues.

Previous publication:

2006:1 Quantifying the quality of macroeconomic variables

2006:2 Stochastic population projections for Sweden

2007:1 Jämförelse av röjanderiskmått för tabeller

2007:2 Assessing auxiliary vectors for control of nonresponse bias in the calibration estimator.

2007:3 Kartläggning av felkällor för bättre aktualitet

2008:1 Optimalt antal kontaktförsök i en telefonundersökning

2009:1 Design for estimation: Identifying auxiliary vectors to reduce nonresponse bias

2009:2 Demographic methods for the statistical office

2011:1 Three Factors to Signal Nonresponse Bias – With applications to Categorical Auxiliary Variables

2011:2 Quality assessment of administrative data

2012:1 Aspects of Responsive Design for the Swedish Living Conditions Survey

2012:2 Linking a system of time series – Recalculation of Labour Force Surveys 1987–1992

Research and Development – Methodology reports from Statistics Sweden 2012:3

Consistent Calibration Estimation under Nonresponse

Thomas Laitila

Statistiska centralbyrån
2012

Research and Development – Methodology reports from Statistics Sweden 2012:3

Consistent Calibration Estimation under Nonresponse

Statistics Sweden
2012

Producer Statistics Sweden, Research and Development Department
SE-701 89 ÖREBRO
+ 46 19 17 60 00

Enquiries Thomas Laitila, +46 (0) 70 38 62 42
Hassan.mirza@scb.se

Cover Ateljén, SCB

ISSN 1653-7149 (Online)
URN:NBN:SE:SCB-2012-X103BR1203_pdf

This publication is only published electronically on Statistics Sweden's website www.scb.se

Preface

Nonresponse is a threat to the reliability of sample survey statistics. Proper analysis and estimation methods have to be used to avoid misleading statistics due to nonresponse bias. Estimators based on the calibration approach have been shown to reduce nonresponse bias and is therefore used by statistical agencies. Improvement of the calibration approach to gain further insights into its properties and potentials are of great interest to Statistics Sweden. This research report deals with calibration estimation and presents new results useful for further development of the calibration approach.

Statistics Sweden, November 2012

Lilli Japec

Acknowledgement

The author is indebted to Carl-Erik Särndal and Martin Axelson for valuable comments on earlier versions of this paper.

Disclaimer

The series Research and Development – Methodology reports from Statistics Sweden is published by Statistics Sweden and includes results on development work concerning methods and techniques for statistics production. Contents and conclusions in these reports are those of the author(s).

Contents

Abstract

This paper considers the consistency of the calibration estimator suggested by Särndal and Lundström (2005) for estimation in sample surveys under nonresponse. Formulating the calibration estimator in terms of a GREG estimator, the estimator is shown generally inconsistent as the population regression coefficient vector is inconsistently estimated. The sources to the inconsistency are detailed upon and a modified calibration estimator is suggested, along with new types of instrument variable vectors. Consistency is derived for the modified estimator and results from a simulation study encourage further research on the modified calibration estimator.

1. Introduction

Calibration estimation in sample surveys has since its introduction by Deville and Särndal (1992) developed into an established theory and method for estimation of finite population parameters (see e.g. Särndal (2007) and Kim and Park (2010) for reviews). The core idea of the calibration estimation approach is to utilize auxiliary information by replacing the design weights in the Horwitz-Thompson (HT) estimator with weights replicating known population totals when attached to auxiliary variables. Although the generalized regression (GREG) estimator can be derived as a special case of a calibration estimator, the approaches of utilizing auxiliary information in estimation are conceptually different. The GREG approach is based on the assumption of a linear relationship between the study variable and the vector of auxiliary variables. An assumption of such an assisting model is not necessary in the calibration approach (Särndal, 2007).

Calibration against a small set of population totals does not uniquely determine the weights, and different approaches for reaching uniqueness have been considered. Deville and Särndal (1992) obtained uniqueness by minimizing a summary measure of the distances between design weights and calibrated weights. In the functional form approach, the weights are specified by a function of a variable vector, a function known up to an unknown parameter vector. Estevao and Särndal (2000; 2006) proposed the functional form approach and specified the calibration weight as a linear function of a variable vector. A more general functional form approach is considered by Kott (2006) who proposes a class of functions defined by the restriction to yield consistency of the calibration estimator. In the functional form approach the variable vector may include other variables then those included in the auxiliary variable vector. Here the term of an "instrument variable vector" has been adopted from the regression literature, and the functional form approach is often referred to as the "instrument vector method" (e.g. Särndal, 2007). The naming "functional form approach" is preferred here. The name "instrument vector method" does not reveal the essentiality in the method of specifying a known function.

One part of the theory developed on calibration estimation is devoted to its potentials to adjust for nonresponse bias in sample surveys (Lundström and Särndal, 1999; Särndal and Lundström, 2005; Kott, 2006). However, although the calibration approach can be used to define estimators which are approximately unbiased and consistent in the full sample case, calibration estimators are biased and inconsistent under nonresponse in general. Särndal and Lundström (2005) provide several approximate bias expressions, providing guidelines on how to reduce bias by appropriate selection of auxiliary variables. Similar attempts to develop indicators and algorithms for selection of appropriate sets of auxiliary variables are found in e.g. Schouten (2007) and Särndal and Lundström (2008).

Consistency of the calibration estimator is obtained in particular cases. One such case is considered by Kott (2006), Chang and Kott (2008) and Kott and Chang (2010) who study calibration estimators when the unit's response probabilities are known functions of an instrument variable vector and an unknown parameter vector. One choice of functional form is to use the inverse of the response probability function. The calibration equations then serve as "estimating equations" providing consistent estimates of the unknown parameters in the response probability functions. This in turn renders consistency to the calibration estimator.

This paper focuses on the consistency properties of the calibration estimator proposed by Särndal and Lundström (2005). Their estimator is defined with weights being a linear function of an instrument variable vector, and the purpose here is to study how this functional form approach can be utilized to reach consistency of the calibration estimator. For this a condition on the instrument variables for consistency is presented. Based on this instrument variable condition a small modification of the calibration estimator is suggested along with a new kind of instrument variables. The resulting calibration estimator is shown to yield consistency under appropriate selection of instrument variables. Indications on the finite population properties of the modified calibration estimator are provided by means of simulation.

The calibration estimator proposed by Särndal and Lundström (2005) is defined in the next section and its consistency properties are discussed in Section 3. A modified calibration estimator is suggested in Section 4 along with an introduction of a new type of instrument vectors. The setup and results of a simulation study are reported in Section 5. A discussion of results and suggestions for future research are contained in the final section. Proofs of theorems are found in the Appendix.

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 $r \subset s$. The response set is assumed random, generated from a conditional response distribution $q(r|s)$. The response distribution is assumed to imply a probability of a response from unit k, given its selection to a sample, equaling $\theta_k = \Pr(k \in r | k \in s)$.

Let x_k denote a column vector of non-random auxiliary variables known for all units in r and with population totals $X = \sum_U x_k$ which are either known or estimated. The vector $\tilde{X}$ is used to denote the vector X where some or all elements may be replaced by estimates.

The first order inclusion probability of unit $k \in U$ is denoted π_k and $d_k = \pi_k^{-1}$ denotes the corresponding design weight. Also, let z_k ($k \in r$) denote a known instrument vector with the same dimension as x_k . Särndal and Lundström (2005) defines the calibration estimator

$$\hat{Y}_W = \sum_r w_k y_k \tag{2.1}$$

where the calibration weights are defined by the system

$$\begin{aligned} w_k &= d_k v_k \\ v_k &= 1 + \lambda_r z_k \\ \tilde{X} &= \sum_r w_k x_k \end{aligned} \tag{2.2}$$

This system yields $\lambda_r = \left(\tilde{X} - \sum_r d_k x_k\right)^t \left(\sum_r d_k z_k x_k^t\right)^{-1}$.

The calibration system (2.2) with $z_k = x_k$ is derived by Deville and Särndal (1992) in the full sample case by minimizing the Lagrange function

$$L(w,\lambda_s) = \sum_s \left(w_k - d_k\right)^2 \Big/ d_k + \lambda_s\left(\tilde{X} - \sum_s w_k x_k\right)$$

yielding the generalized regression (GREG) estimator. Following their results, implementing the weights defined by (2.2) in (2.1) yields

$$\hat{Y}_W = \tilde{X}^t \hat{B}_r + \sum_r d_k (y_k - x_k^t \hat{B}_r) \qquad (2.3)$$

where

$$\hat{B}_r = \left(\sum_r d_k z_k x_k^t\right)^{-1} \sum_r d_k z_k y_k \qquad (2.4)$$

Consistency of the calibration estimator (2.1) is here treated by studying the asymptotic properties of (2.4). The following definitions are used for the study of consistency.

Definition 1: (Sequence of populations)
The variable vector $t_k = \begin{pmatrix} y_k & x_k^t & z_k^t \end{pmatrix}^t$ is non-random, real-valued and defined for a bounded set such that $\left\| t_k \right\| < \kappa$ for some $0 < \kappa < \infty$. Assume existence of an infinite sequence $\{t_k\}_{k=1}^{\infty}$. The population U_N is defined as the index set for the first N units in the sequence $\{t_k\}_{k=1}^{\infty}$ with the associated set of variable vectors $\{t_1 \quad t_2 \quad \cdots \quad t_N\}$.

Definition 2: (Sequence of response sets)
A probability sample $s_N \subseteq U_N$ of expected size $n(N)$ is drawn from U_N using a probability sampling design $p_N(s)$, yielding positive inclusion probabilities $\pi_k > \varsigma > 0$ for all $k \in U_N$. Conditionally on the sample, observations of t_k are obtained for a subset of the sample, $r_N \subseteq s_N$ due to non-response. The response set r_N is generated

according to the response distribution $q_N(r_N|s_N)$ yielding response probabilities $\theta_k = \Pr(k \in r_N | k \in s_N)$.

Definition 3: (Consistency)
Let $\hat{\psi}$ denote an estimator of a population quantity ψ defined on the finite population U. For the sequences of populations, samples and response sets in Definition 2, let $\hat{\psi}_N$ and ψ_N $(N = 1,2,3,...)$ denote corresponding sequences of $\hat{\psi}$ and ψ, respectively. Then $\hat{\psi}$ is defined as consistent for ψ if

$$p\lim_{N\to\infty}(\hat{\psi}_N - \psi_N) = 0$$

A sequence of populations of the kind specified in Definition 1 is usually adapted when considering asymptotic properties of statistics in finite population settings (e.g. Särndal et al. 1992; Fuller, 2009). The assumption of having a sequence of populations where the increase in size is obtained by adding new units to the previous populations is not, however, necessary for the theoretical arguments to hold. The assumption of a real valued t_k means the elements can be real valued, but they can also take on values within a subset of the real numbers, e.g. the natural numbers. The limit statement in Definition 3 is in terms of a population size *N* approaching infinity, and not in terms of the sample size *n*. Due to the lower bound on the inclusion probability in Definition 2, the expected sample size will approach infinity when the population size does so.

3. Consistency of the calibration estimator

Consider a census in which all units in the population are sampled. Then due to non-response the sums involved in the r.h.s. of (2.4) has expected values $E_{q(r|U)}\left(\sum_U z_k x_k^t\right) = \sum_U \theta_k z_k x_k^t$ and $E_{q(r|U)}\left(\sum_U z_k y_k\right) = \sum_U \theta_k z_k y_k$, respectively. These quantities define the parameter vector

$$B_\theta = \left(\sum_U \theta_k z_k x_k^t\right)^{-1} \sum_U \theta_k z_k y_k \qquad (3.1)$$

Under a general sampling design, complying with Definition 2, the following theorem states $\hat{B}_r$ to be consistent for B_θ.

Theorem 1: (Probability limit for $\hat{B}_r$)

Assume definitions 1 and 2 and let $\hat{B}_{Nr}$ and $B_{N\theta}$ denote corresponding sequences of $\hat{B}_r$ and B_θ, respectively. Also assume i) the sampling design yields second order inclusion probabilities $\pi_{kl} = \Pr(k \,\&\, l \in s)$ such that $N^{-2} \sum\sum_{U_N, k \neq l} |\pi_{kl} d_k d_l - 1| = o(1)$ and ii) $N^{-1} \sum_{U_N} \theta_k z_k x_k^t$ has a determinant bounded away from zero for $N > N_0$, then

$$p\lim_{N\to\infty}(\hat{B}_{Nr} - B_{N\theta}) = 0$$

A proof of Theorem 1 is given in the Appendix and it yields the following result for the calibration estimator $\hat{Y}_W$.

Corollary 1: (Consistency of $\hat{Y}_W$)

Under the assumptions of Theorem 1, the assumption $\mu^t z_k = 1$ for all $k \in U_N$ and some nonzero vector μ , and $p\lim_{N\to\infty}(\tilde{X}_N - X_N)/N = 0$, then

$$p\lim_{N\to\infty}(\hat{Y}_{NW} - Y_{N\theta})/N = 0 .$$

where $Y_{N\theta} = X_N B_{N\theta}$.

Corollary 1 follows from Theorem 1 and the consistency of $\tilde{X}$. The condition $\mu^t z_k = 1$ implies $\sum_r d_k (y_k - x_k^t \hat{B}_r) = 0$ in (2.3), and the corollary shows the calibration estimator (2.1) to be consistent for $Y_\theta = XB_\theta$, which is different from the population total $Y = \sum_U y_k$, in general. The assumption $\mu^t z_k = 1$ implies $Y = XB_U$ where $B_U = \left(\sum_U z_k x_k^t\right)^{-1} \sum_U z_k y_k$. The result then yields the approximate bias expression

$$Bias(\hat{Y}_W) \approx X\left(B_\theta - B_U\right)$$

which is one of the three equivalent nearbias expressions derived by Särndal and Lundström (2005, Corollary 9.1). Here Corollary 1 shows that the inconsistency of the calibration estimator can be interpreted as a consequence of the inconsistency of $\hat{B}_r$ as an estimator of B_U .

An interesting problem is the restrictions imposed by the condition $\mu^t z_k = 1$ on the consistency property of the calibration estimator. Consistency of $\hat{B}_r$ is obtained under the following instrument variable condition:

Instrument Variable Condition: $\sum_U \theta_k z_k e_k = 0 \qquad ;\ e_k = y_k - x_k^t B_U$

Adding the condition $\mu^t z_k = 1$ yields $\sum_U \theta_k e_k = 0$, i.e. the response probabilities are uncorrelated with the regression errors in the population. Thus, consistency of the calibration estimator (2.1) via consistent estimation of B_U cannot generally be achieved under the condition $\mu^t z_k = 1$. On the other hand, relaxing the condition does not improve the matter, since $Y = X^t B_U$ does not generally hold if $\mu^t z_k \neq 1$. Relaxing $\mu^t z_k = 1$ and with $\hat{B}_r$ consistent for B_U then (2.1) is consistent for

$$Y_{We} = X^t B_U + \sum_U \theta_k (y_k - x_k^t B_U)$$

This quantity is equal to Y if $\sum_U (1-\theta_k)e_k = 0$, where $\sum_U e_k \neq 0$ in general.

Consistency of $\hat{Y}_W$ can be obtained in cases with $B_\theta \neq B_U$. For instance, the result of zero "nearbias" of the calibration estimator under $\mu^t z_k = 1$ and the additional assumption $\phi_k = \theta_k^{-1} = \eta^t z_k$, for some vector of constants η, is derived by Särndal and Lundström (2005, Prop. 9.2). Consistency is obtained since B_θ satisfies $\sum_U \theta_k (y_k - x_k^t B_\theta) z_k = 0$ and multiplying from the left with η^t yields $\sum_U (y_k - x_k^t B_\theta) = Y - X^t B_\theta = 0$. Furthermore, by relaxing $\mu^t z_k = 1$ and assuming $\phi_k = 1 + \eta^t z_k$ yields $\sum_U \theta_k (y_k - x_k^t B_\theta)(\phi_k - 1) = 0$, implying $Y = X^t B_\theta + \sum_U \theta_k (y_k - x_k^t B_\theta)$ and consistency of $\hat{Y}_W$.

Consistency of $\hat{Y}_W$ under the assumed functions for ϕ_k can also be obtained by the result $p\lim_{N\to\infty} \lambda_r = \eta$ (Kott and Chang, 2010). This result implies v_k in (2.2) to be consistent estimates of ϕ_k. The linear form assumed for ϕ_k is restrictive and more general forms of functions are considered by Kott (2006), Chang and Kott (2008) and Kott and Chang (2010).

4. A modified calibration estimator

Schouten (2007) makes use of a super population model approach and shows zero bias of the GREG estimator under a MAR (Missing at Random) response mechanism. The MAR condition implies conditional independence of the response and the study variable, given a set of auxiliary variables. Treat z_k and e_k as realizations of the random variables *z* and *e*, respectively, and interpret the response indicator $R_k = 1(k \in r | k \in s)$ as a realization of the random indicator variable *R*. Then one interpretation of the MAR condition is $E(Rze) = E_z(zE(Re|z)) = 0$, which would imply zero expectation of the sum of *Rze* over a selected sample. The instrument variable condition $\sum_U \theta_k z_k e_k = E\left(\sum_U R_k z_k e_k\right) = 0$ can thus be seen as a MAR condition in the pseudo randomization setting. An important distinction is that the MAR condition puts restrictions on individual observations, while the instrument variable condition puts a restriction on a vector sum over the units in the population.

One way of defining an estimator $\hat{B}_r$ consistent for B_U, is to find an instrument vector satisfying the instrument variable condition. Unfortunately, this is not sufficient for consistency of the resulting calibration estimator, as shown in the previous section. To solve the dilemma a calibration estimator defined with two instrument variable vectors is suggested. The first is denoted as z_k and is assumed to satisfy the instrument variable condition, and the second is denoted as z_{Bk} assumed to satisfy the unity condition $\mu^t z_{Bk} = 1$. The second instrument vector is utilized for defining the population regression vector $B_U(z_B) = \left(\sum_U z_{Bk} x_k^t\right)^{-1} \sum_U z_{Bk} y_k$. (An argument ($z_B$) is introduced to stress the dependence on the chosen instrument vector.) The first instrument vector then yields an estimator $\hat{B}_r(z) = \left(\sum_r d_k z_k x_k^t\right)^{-1} \sum_r d_k z_k y_k$ which is consistent, since it is assumed to satisfy the instrument variable condition. The calibration estimator suggested is $\hat{Y}_{MW} = \tilde{X}^t \hat{B}_r(z)$ which in the weighted sum form equals

$$\hat{Y}_{WM}(z) = \sum_r w_{Mk}(z) y_k \tag{4.1}$$

with weights

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

The weights $w_{Mk}(z)$ are defined in Särndal and Lundström (2005, Sec. 6.8) and by them suggested to constitute the main part of the weights given by (2.2). They also notice that the weights given by (2.2) reduce to the weights $w_{Mk}(z)$ under the condition $\mu^t z_k = 1$. Here, the weights are suggested for the calibration estimator irrespective if the condition $\mu^t z_k = 1$ is satisfied or not. It can be noted that estimator (4.1) is a "synthetic" estimator, which is biased in general (e.g. Lehtonen and Pahkinen, 2004, p. 198). However, under appropriate selection of instrument variables, the synthetic estimator (4.1) is consistent for an unbiased instrument variable GREG estimator.

A simple example of the estimator is $z_{Bk} = x_k$, where the auxiliary vector includes a constant term, and $z_k = \phi_k x_k$. Then $Y = X^t B_U(x)$ and $\hat{B}_r(\phi x)$ is consistent for $B_U(x)$ by Theorem 1 and since $\sum_U \theta_k z_k e_k = \sum_U x_k e_k = 0$. Then $\hat{Y}_{WM} = X^t \hat{B}_r(\phi x)$ is consistent for Y by Corollary 1.

The two sets of instrument variable vectors considered are below named as *Theoretical instruments* (z_{Bk}) and *Operational instruments* (z_k). In the example above, the theoretical instruments were chosen as the known and observable vector x_k. In general the theoretical instruments can be unobservable, theoretical concepts.

In practice it is likely that the operational instruments are calculated from the data at hand, e.g. the response set or the sample, providing "approximate" instruments or estimated instruments. In the propensity scoring example above, i.e. $\hat{B}_r(\hat{\phi}x)$, the inverse response probability are replaced by estimates $\hat{\phi}_k$ yielding a propensity scoring estimator $\hat{B}_r(\hat{\phi}x)$ of $B_U(x)$, and not the estimator $\hat{B}_r(\phi x)$. Thus, the consistency property of the calibration estimator (4.1) has to be considered when it is defined with such "estimated" instruments. For this, the following definition and assumption are made.

Definition 1A: (Sequence of populations)
The variable vector $t_k = \begin{pmatrix} y_k & x_k^t & z_k^t & z_{Bk}^t \end{pmatrix}^t$ is non-random, real-valued and defined for a bounded set such that $\|t_k\| < \kappa$ for some $0 < \kappa < \infty$. Assume the existence of the infinite sequence $\{t_k\}_{k=1}^{\infty}$. The population U_N is then defined as the index set for the first *N* units in the sequence $\{t_k\}_{k=1}^{\infty}$ with the associated set of variable vectors $\{t_1 \quad t_2 \quad \cdots \quad t_N\}$.

Assumption 1: (Uniform convergence of estimated operational instruments $\tilde{z}_k$)
Let $\tilde{z}_k \ (k \in r)$ denote estimates of the corresponding unknown vectors $z_k \ (k \in r)$. For sufficiently large *N*, $\|\tilde{z}_k - z_k\| < \mathrm{M}\,\omega_N$ where M is a finite positive constant, and $\omega_N = o_p(1)$ is a random scalar term.

If $\hat{B}_r(\tilde{z})$ is consistent for $B_\theta(z)$ and $B_\theta(z)$ equals $B_U(z_B)$, then $\hat{B}_r(\tilde{z})$ is consistent for $B_U(z_B)$ whereby consistency of the calibration estimator (4.1) is obtained. The probability limit of $\hat{B}_r(\tilde{z})$ is stated in the following theorem.

Theorem 2: (Consistency of $\hat{B}_r(\tilde{z})$ for $B_\theta(z)$)
Assume definitions 1A and 2, and let $\hat{B}_{Nr}(\tilde{z})$ and $B_{NU}(z_B)$ denote corresponding sequences of $\hat{B}_r(\tilde{z})$ and $B_U(z_B)$. Assume Assumption 1, i) the sampling design yields second order inclusion

probabilities $\pi_{kl} = \Pr(k \,\&\, l \in s)$ such that $N^{-2} \sum \sum_{U_N, k \neq l} \left| \pi_{kl} d_k d_l - 1 \right| = o(1)$, and ii) the determinant of $\sum_{U_N} \theta_k z_k x_k^t$ is bounded away from zero for all $N > N_0$, then

$$p \lim_{N \to \infty} (\hat{B}_{Nr}(\tilde{z}) - B_{N\theta}(z)) = 0$$

A proof of Theorem 2 is given in the Appendix. Theorem 2 gives the following corollary.

Corollary 2: (Consistency of $\hat{Y}_{WM}$)

If $B_\theta(z) = B_U(z_B)$ and the assumptions of Theorem 2 and the assumption $p \lim_{N \to \infty} (\tilde{X}_N - X_N) / N = 0$ holds, then

$$p \lim_{N \to \infty} (\hat{Y}_{WMN} - Y_N) / N = 0 .$$

where $\hat{Y}_{WMN} = \tilde{X}^t \hat{B}_{Nr}(\tilde{z})$.

Corollary 2 establishes the consistency of the calibration estimator (4.1) as an estimator of the population total Y.

5. Examples

The modified calibration estimator suggested in Section 4 is here illustrated with two examples of operational instrument vectors. The first is the case considered above using $z_k = \hat{\phi}_k x_k$. The second uses instruments derived from a specification of an assisting model for the response set data. For both cases the theoretical instrument vector is defined as $z_{Bk} = x_k$. A numerical illustration of the performance of the estimators is presented in the end of this section.

5.1 Instruments by weighting auxiliary variables

Consider the response probability function $\theta_k = G(u_k^t \alpha_0)$ where u_k is a real valued bounded vector ($\|u_k\| < \kappa < \infty$), known for units in the sample, and $\alpha_0 \in A$ is a corresponding real valued vector of unknown, fixed parameters, taking on a value within a bounded parameter space A. The length of u_k is assumed fixed and $G(u_k^t \alpha)$, $\alpha \in A$, is assumed bounded away from zero and one, i.e. $0 < \tau_0 < G(u_k^t \alpha) < \tau_1 < 1$, for all $k \in U$.

With a sampling design, consider the design weighted log likelihood function

$$l_s(\alpha) = N^{-1} \sum\nolimits_s \left(d_k R_k \log(G(u_k^t \alpha)) + d_k (1 - R_k) \log(1 - G(u_k^t \alpha)) \right)$$

and the design weighted maximum likelihood estimator

$$\hat{\alpha}_s = \arg\max l_s(\alpha)$$

Consistency of $\hat{\alpha}_s$ implies Assumption 1 since $\left| G(u_k^t \hat{\alpha}_s)^{-1} - G(u_k^t \alpha_0)^{-1} \right| < \tau_1^{-2} \|u_k\| \cdot \|\hat{\alpha}_s - \alpha_0\|$. A formal proof of the consistency of $\hat{\alpha}_s$, is out of the scope of this paper. However, consistency can be derived by showing uniform convergence in probability of $l_s(\alpha)$ to

$$\bar{l}_U(\alpha) = E(l_s(\alpha)) = N^{-1}\sum_U \theta_k \log(G(u_k^t\alpha)) + (1-\theta_k)\log(1-G(u_k^t\alpha))$$

a function that is uniquely maximized over A at $\alpha = \alpha_0$ for sufficiently large N. Consistency of $\hat{\alpha}_s$ can then be established by similar arguments as in the proof of Lemma 2.2 in White (1980).

5.2 Assisting model for response set data

The assisting model underlying the GREG estimator is of the form $y_k = x_k^t B_U + e_k$, where the properties of the "errors" e_k are defined by the definition of B_U. With B_U being the population least squares (LS) solution, the corresponding design weighted LS solution for a sample under full response is consistent for B_U.

For observations obtained in a response set r, consider the assisting model $y_k = x_k^t B_U + \eta_k + \omega_k$. Here, 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, consider 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, least squares regressions of the elements in x_k on η_k. Note here that δ is a vector with a slope coefficient for each element in x_k. These instruments have the property $\sum_U \theta_k z_k (\eta_k + \omega_k) = 0$ whereby $\hat{B}_r(z)$ is consistent for B_U.

Unfortunately η_k is not observable. However, assume η_k can be described by a function known up to a finite set of coefficients, e.g. $\eta_k = \eta(u_k, \alpha_0)$, where the arguments are defined as in subsection 5.1. Also, assume there exists a consistent estimator $\hat{\alpha}_s$ of α_0. Then instruments can be formed by $\tilde{z}_k = x_k - \hat{\eta}_k \hat{\delta}$ where $\hat{\eta}_k = \eta(u_k, \hat{\alpha}_s)$ and $\hat{\delta} = \left(\sum_r \theta_k \hat{\eta}_k^2\right)^{-1} \sum_U \theta_k \hat{\eta}_k x_k$. Under suitable assumptions, e.g.

boundedness of $\eta_k = \eta(u_k, \alpha_0)$ and existence of $\sum_U \theta_k \eta_k^2$, $p\lim_{N\to\infty} \hat{\delta}_N = \delta$ and Assumption 1 holds for the instrument vector $\tilde{z}_k = x_k - \hat{\eta}_k \hat{\delta}$. $\hat{B}_r(\tilde{z})$ is then consistent for B_U and the calibration estimator (4.1) is consistent for the population total Y.

One alternative for the specification of the function $\eta_k = \eta(u_k, \alpha_0)$ is provided by Heckman (1979) where $\eta(u_k, \alpha) = f(u_k^t \alpha)/\Phi(u_k^t \alpha)$, the ratio of the standard normal pdf to its cdf. This function is derived from assuming a probit model for the response probability function, e.g. $\theta_k = \Phi(u_k^t \alpha_0) = E1(u_k^t \alpha_0 + \varphi_k > 0 | u_k)$ with $\varphi \sim N(0,1)$, and e_k assumed to be a realization of a normal distributed random variable. The normal distribution assumption for e_k can be replaced by the moment assumption $E(e_k | R_k = 1, \varphi_k) = \rho \varphi_k$ (Wooldridge, 2002). Estimates $\hat{\alpha}_s$ can be obtained by applying the design weighted ML estimator defined in the previous subsection.

5.3 Numerical illustration

A small simulation study is used to illustrate the empirical properties of the estimator (4.1) based on the instruments considered in subsections 5.1 and 5.2. The instruments are derived using $\theta_k = \Phi(u_k^t \alpha)$ and $\eta(u_k, \alpha) = f(u_k^t \alpha)/\Phi(u_k^t \alpha)$, respectively. For both cases the design weighted probit ML estimator $\hat{\alpha}_s$ is used.

Population data is simulated from the following model:

$$
\begin{aligned}
y &= \beta_0 + \beta_1 x + \varepsilon \\
R^* &= \alpha_0 + \alpha_1 y + \alpha_2 u + \varphi
\end{aligned}
\tag{5.1}
$$

where y represents the variable under study, 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 between $v = \alpha_1 \varepsilon + \varphi$ 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 Unif(* ε *)/Unif(u):* As in *i)* but with ε and φ generated from uniform distributions. Response rates are around 58%.

For each of the two population models, finite populations $U_1 \subset U_2 \subset U_3 \subset U_4$ of sizes *N*=2000, 5000, 10000, 20000 units are generated. For the generated populations, samples of sizes *n*=*N*/10 are drawn with SRS, without replacement, and response sets are generated using the generated values on R^*. For each generated population, samples and response sets are replicated 1000 times.

For comparison, the calibration estimator (2.1) is applied in two versions; one using the auxiliary vector $(1, x)$ and one using the auxiliary vector $(1, x, u)$. These calibration estimators are applied with standard weights, i.e. $z_k = x_k$. For the estimator (4.1), the auxiliary vector $(1, x)$ is used and the instruments defined in subsection (5.1) and (5.2) are constructed 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.

Table 1:
Simulated Bias and St.dev (in parenthesis) of the population mean calibration estimators $\hat{\bar{Y}}_W = \hat{Y}_W / N$ and $\hat{\bar{Y}}_{WM} = \hat{Y}_{WM}(z) / N$ (c.f. equations (2.1) and (4.1), respectively). Population generated from the population model Norm(ε)/Norm(u), for different sample sizes *n* and correlations $\rho = Corr(v, \varepsilon)$ (c.f. model (5.1)). Results based on 1000 replications

		Sample/Population size (n/N)			
Estimator[a)]	ρ	200/2000	500/5000	1000/10000	2000/20000
$\hat{Y}_W(x1)$	0	-.034 (.090)	.017 (.058)	-.004 (.039)	.007 (.028)
	0.1	.013 (.090)	.067 (.058)	.043 (.039)	.050 (.028)
	0.3	.115 (.092)	.164 (.058)	.150 (.038)	.152 (.028)
$\hat{Y}_W(x2)$	0	-.026 (.107)	.008 (.072)	.004 (.045)	.010 (.033)
	0.1	.048 (.111)	.079 (.071)	.074 (.044)	.073 (.033)
	0.3	.198 (.108)	.218 (.065)	.212 (.042)	.212 (.031)
$\hat{Y}_{WM}(z1)$	0	-.035 (.126)	.002 (.086)	.010 (.047)	.009 (.037)
	0.1	.035 (.125)	.075 (.083)	.079 (.048)	.072 (.037)
	0.3	.197 (.120)	.226 (.074)	.218 (.046)	.226 (.037)
$\hat{Y}_{WM}(z2)$	0	-.048 (.139)	.035 (.084)	-.020 (.060)	.002 (.042)
	0.1	-.046 (.142)	.049 (.086)	-.016 (.060)	.010 (.042)
	0.3	-.078 (.158)	.052 (.095)	-.011 (.067)	.003 (.047)

[a)] Auxiliary/Instrument vectors: $x1=(1\ x)^t$, $x2=(1\ x\ u)^t$, $z1=\hat{\phi}\,x1$, $z2=x1-\hat{\delta}\hat{\eta}$.

Table 1 includes simulation results under the Norm(ε)/Norm(φ) model, meaning that the assumptions of the model considered by Heckman (1979) are satisfied. The calibration estimator (4.1) based on the instrument variables considered in subsection 5.2 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, which is also observed for the other values of ρ. The results observed for all estimators are compatible with an expected pattern of consistent estimators when $\rho = 0$.

Also as expected, when $\rho = 0.1$ or $\rho = 0.3$ this "consistency pattern" is only observed for the estimator $\hat{Y}_{WM}(z2)$. For the other estimators, the standard deviations decrease with the sample size, but the biases do not. The observation is also that the biases are increasing with ρ for these estimators.

The consistency property of the estimator $\hat{Y}_{WM}(z2)$ comes at the price of a larger variance then the other estimators. The RMSE of $\hat{Y}_{WM}(z2)$ is around 50% higher in comparison with $\hat{Y}_{W}(x1)$ for $\rho = 0$. For $\rho = 0.3$ the RMSE is smaller for $\hat{Y}_{WM}(z2)$ compared to $\hat{Y}_{W}(x1)$ with the exception of the *n*=200 case, where the RMSEs are about the same. For the two smallest sample sizes when $\rho = 0.1$, $\hat{Y}_{W}(x1)$ has smaller MSE then $\hat{Y}_{WM}(z2)$, while it is about the same for *n*=1000 and smaller for $\hat{Y}_{WM}(z2)$ when *n*=2000.

Table 2:
Simulated Bias and St.dev (in parenthesis) of the population mean calibration estimators $\hat{\bar{Y}}_W = \hat{Y}_W / N$ and $\hat{\bar{Y}}_{WM} = \hat{Y}_{WM}(z) / N$ (c.f. equations (2.1) and (4.1), respectively). Population generated from the population model Uniform(ε)/Uniform(u), for different sample sizes *n* and correlations $\rho = Corr(v, \varepsilon)$ (c.f. model (5.1)). Results based on 1000 replications

		Sample/Population size (*n/N*)			
Estimator[a)]	ρ	200/2000	500/5000	1000/10000	2000/20000
$\hat{Y}_W(x1)$	0	-.010 (.086)	.012 (.054)	.017 (.039)	.001 (.029)
	0.1	.039 (.086)	.059 (.054)	.059 (.039)	.047 (.029)
	0.3	.116 (.086)	.155 (.054)	.154 (.039)	.152 (.028)
$\hat{Y}_W(x2)$	0	-.032 (.098)	.007 (.061)	.018 (.044)	-.002 (.033)
	0.1	.044 (.097)	.071 (.061)	.078 (.044)	.060 (.032)
	0.3	.136 (.094)	.193 (.060)	.194 (.042)	.190 (.031)
$\hat{Y}_{WM}(z1)$	0	-.036 (.106)	.005 (.067)	.019 (.048)	-.002 (.035)
	0.1	.047 (.102)	.077 (.067)	.082 (.047)	.064 (.035)
	0.3	.154 (.096)	.206 (.064)	.209 (.043)	.207 (.031)
$\hat{Y}_{WM}(z2)$	0	.046 (.136)	.026 (.086)	.013 (.060)	.007 (.045)
	0.1	.026 (.143)	.028 (.087)	.017 (.061)	.017 (.045)
	0.3	.033 (.165)	.032 (.099)	.033 (.067)	.025 (.051)

[a)] Auxiliary/Instrument vectors: $x1=(1\ x)^t$, $x2=(1\ x\ u)^t$, $z1=\hat{\phi}\,x1$, $z2=x1-\hat{\delta}\hat{\eta}$.

Although the assumptions in the model considered by Heckman (1979) are not satisfied in the Uniform(ε)/Uniform(φ) case considered in Table 2, the results show a similar pattern as the one in Table 1. All estimators have decreasing biases and standard deviations as the sample size increase in the $\rho = 0$ case. For the other values of ρ this is only observed for $\hat{Y}_{WM}(z2)$. However, $\hat{Y}_{WM}(z2)$ is not consistent in the Uniform(ε)/Uniform(φ) case, a fact indicated by the slight increase in bias for increasing ρ. Still, estimates of the bias are much smaller than those of the other estimators. The pattern of RMSE for $\hat{Y}_{WM}(z2)$ in comparison with $\hat{Y}_{W}(x1)$ is the same in Table 2 as in Table 1.

6. Summary

This paper is concerned with the consistency properties of the calibration estimator suggested by Lundström and Särndal (1999) and Särndal and Lundström (2005). It is shown that the calibration estimator is generally inconsistent. This does not come as a surprise since the estimator is already known to have a non-zero approximate bias. Efforts have also been made in the literature to find ways of selecting auxiliary information in order to reduce bias as much as possible. In these contributions, the calibration estimator considered has been based on "standard weights" where the auxiliary vector is used as the instrument vector.

The prospect of defining a generally consistent calibration estimator based on a standard weights formulation is limited. The auxiliary variable vector does not generally satisfy the instrument variable condition, and the success of reducing bias in the standard weight version of the calibration estimator depends on how well the auxiliary variables i) "explains" the variation in the study variable, and ii) how well a linear combination of them approximates the inverses of response probabilities (Särndal and Lundström, 2005). Similar results are presented by Schouten (2007), who derive bounds on the bias of the GREG estimator in terms of correlations between auxiliary variables and the study variable, and between auxiliary variables and the response indicator.

Instead of the standard weight formulation, it is here suggested to place focus on the potentials of defining an instrument variable vector different from the auxiliary variable vector. Rewriting the calibration estimator in terms of a GREG estimator formulation, the non response bias can be interpreted in terms of a correlation between auxiliary variables (regressors) and the regression error term. Adapting the idea of instrument variable estimation suggests finding an appropriate instrument vector for the calibration estimator. However, for this approach to work the formulation of the calibration estimator must be modified as suggested.

Two examples of instrument vectors for the proposed calibration estimator are given in the simulation presented. The purpose of the simulation study is to provide an indication of the potentials of the suggested estimator. The instrument vectors considered are mere illustrations, not suggestions for general use. More work is needed for defining instrument vectors which satisfies the instrument variable condition under more general population and response distribution settings. An encouraging fact is that the instrument variable condition imposes restriction on a fixed and small number of weighted population covariances, and may be satisfied without consistent estimation of parameters for each unit in the response set.

Appendix

The proofs of theorems 1 and 2 utilize the following lemma.

Lemma 1:
Under definitions 1 and 2, and the assumptions in Theorem 1, consider the statistic $\hat{\psi}_N = N^{-1}\sum_{r_N} d_k c_k$, where $c_k \ (k = 1,..., N)$ are non-random real valued scalars, bounded by $|c_k| < \kappa < \infty$. Then $p\lim_{N\to\infty}(\hat{\psi}_N - \psi_{N\theta}) = 0$ where $\psi_{N\theta} = N^{-1}\sum_{U_N} \theta_k c_k$.

Proof of Lemma 1:
The design weight d_k is bounded by assumption whereby $\hat{\psi}_N$ exists. Using the sample inclusion indicators $I_k = 1(k \in s)$ and the response indicators $R_k = 1(k \in r | k \in s)$, the expected value of $\hat{\psi}_N$ can be expressed as $E(\hat{\psi}_N) = N^{-1}\sum_{U_N} E(I_k R_k) d_k c_k = N^{-1}\sum_{U_N} \theta_k c_k = \psi_{N\theta}$. For the variance of $\hat{\psi}_N$

$$V(\hat{\psi}_N) = N^{-2}\left(\sum_{U_N} \pi_k^{-1}\theta_k c_k^2 + \sum\sum_{U_N, k\neq l} \pi_{kl}(\pi_k\pi_l)^{-1}\theta_k\theta_l c_k c_l - \sum\sum_{U_N} \theta_k\theta_l c_k c_l\right)$$
$$= N^{-2}\left(\sum_{U_N} \theta_k(d_k - \theta_k)c_k^2 + \sum\sum_{U_N, k\neq l}\left(\pi_{kl}(\pi_k\pi_l)^{-1} - 1\right)\theta_k\theta_l c_k c_l\right)$$
$$\leq N^{-1}\varsigma^{-1}\kappa^2 + \kappa^2 N^{-2}\sum\sum_{U_N, k\neq l}|\pi_{kl}d_k d_l - 1| = o(1)$$

Thus, $V(\hat{\psi}_N) \to 0$ when $N \to \infty$. Chebychev's inequality then provides with the result in the lemma.

Proof of Theorem 1.
Note that $E\left(N^{-1}\sum_{r_n} d_k z_k x_k^t\right) = N^{-1}\sum_{U_N} \theta_k z_k x_k^t$ has a determinant bounded away from zero for $N > N_0$. This implies that $B_{N\theta}$ exists and is bounded. The regression error $e_{\theta k} = y_k - x_k^t B_{N\theta}$ is then also bounded since t_k is assumed bounded. Lemma 1 gives $p\lim_{N\to\infty}\left(N^{-1}\sum_{r_n} d_k z_k x_k^t - N^{-1}\sum_{U_N} \theta_k z_k x_k^t\right) = 0$ whereby the estimator $\hat{B}_{Nr}$ will also exist for *N* sufficiently large. The estimator can be written as

$$\hat{B}_{Nr} = B_{N\theta} + \left(N^{-1}\sum\nolimits_{r_N} d_k z_k x_k^t\right)^{-1} N^{-1}\sum\nolimits_{r_N} d_k z_k e_{\theta k}$$

Here $E\left(N^{-1}\sum\nolimits_{r_n} d_k z_k e_{\theta k}\right) = N^{-1}\sum\nolimits_{U_N} \theta_k z_k e_{\theta k} = 0$, since $B_{N\theta}$ is the solution to the moment condition $\sum\nolimits_{U_N} \theta_k z_k (y_k - x_k^t B_{N\theta}) = 0$. Again using Lemma 1, $p\lim_{N\to\infty} N^{-1}\sum\nolimits_{r_N} d_k z_k e_{\theta k} = 0$. Proposition 2.30 in White (2000) then yields $p\lim_{N\to\infty}\left(N^{-1}\sum\nolimits_{r_N} d_k z_k x_k^t\right)^{-1} N^{-1}\sum\nolimits_{r_N} d_k z_k e_{\theta k} = 0$ and the statement in the theorem is obtained.

Proof of Theorem 2.
The estimated operational instruments are uniformly bounded $\left\|\tilde{z}_k - z_k\right\| < M\cdot\omega_N$ by assumption. Consider $A_{Nrj} = N^{-1}\left\|\sum\nolimits_{r_n} d_k x_{kj}(\tilde{z}_k - z_k)\right\| \le N^{-1}\sum\nolimits_{r_n} d_k \left|x_{kj}\right| \cdot M\cdot\omega_N = o_p(1)$, where x_{kj} denotes the *j*th element in x_k. Thus $N^{-1}\sum\nolimits_{r_n} d_k \tilde{z}_{Hk} x_k^t - N^{-1}\sum\nolimits_{r_n} d_k z_{Hk} x_k^t = o_p(1)$. By applying Lemma 1 $N^{-1}\sum\nolimits_{r_n} d_k z_k x_k^t - N^{-1}\sum\nolimits_{U_N} \theta_k z_k x_k^t = o_p(1)$. Combining these two results in the triangular inequality yields $N^{-1}\sum\nolimits_{r_n} d_k \tilde{z}_k x_k^t - N^{-1}\sum\nolimits_{U_N} \theta_k z_k x_k^t = o_p(1)$. Using similar derivations the result $N^{-1}\sum\nolimits_{r_n} d_k \tilde{z}_k y_k - N^{-1}\sum\nolimits_{U_N} \theta_k z_k y_k = o_p(1)$ is obtained. The inverse of $N^{-1}\sum\nolimits_{U_N} \theta_k z_k x_k^t$ is assumed to exist for $N > N_0$, implying the existence of the solution $B_{N\theta}(z)$. Since $N^{-1}\sum\nolimits_{r_n} d_k \tilde{z}_k x_k^t - N^{-1}\sum\nolimits_{U_N} \theta_k z_k x_k^t = o_p(1)$ this implies that $\hat{B}_{r_N}(\tilde{z})$ also exists for all *N* sufficiently large. Finally, using Corollary 2.30 in White (2000) on the results derived yields the result $p\lim_{N\to\infty}(\hat{B}_{Nr}(\tilde{z}) - B_{N\theta}(z)) = 0$.

References

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

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

Estevao, V.M. and C.-E. Särndal (2000). A functional form approach to calibration, *Journal of Official Statistics*, **16:4**, 379–399.

Estevao, V.M. and C.-E. Särndal (2006). Survey estimates by calibration on complex auxiliary information, *International Statistical Review*, **74:2**, 127–147.

Fuller, W.A. (2009). *Sampling Statistics*. Wiley, Hoboken, New Jersey.

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

Kim, J.K. and M. Park (2010). Calibration estimation in surveys, *International Statistical Review*, **78:1**, 21–39.

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

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.

Lehtonen, R. and E. Pahkinen (2004). *Practical methods for design and analysis of complex surveys*, Wiley, Chichester.

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

Schouten, B. (2007). A selection strategy for weighting variables under a not-missing-at-random assumption. *Journal of Official Statistics,* **23**, 51–68.

Särndal, C.E. (2007). The calibration approach in survey theory and practice, *Survey Methodology*, **33:2**, 99–119.

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

Särndal, C.E. and S. Lundström (2008). Assessing auxiliary vectors for control of nonresponse bias in the calibration estimator, *Journal of Official Statistics*, **24:2**,
167–191.

Särndal, C.E, Swensson, B. and J. Wretman (1992). *Model Assisted Survey Sampling*, Springer, New York.

White, H. (1980). Nonlinear Regression on Cross-Section Data, *Econometrica*, **48:3**, 721–746.

White, H. (2000). *Asymptotic theory for econometricians*, Academic Press, New York.

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

ISSN 1653-7149 (Online)

All official statistics can be found at: www.scb.se
Statistics Service, phone +46 8 506 948 01