Estimation of Regression Coefficients with Unequal Probability Samples

Yu Y. Wu, Wayne A. Fuller
Center for Survey Statistics and Methodology
Iowa State University

Abstract

We compare alternative estimators for regression coefficients estimated with data from a complex survey. The ordinary least squares estimator is a common choice of researchers, but under an informative design, the ordinary least squares estimator is biased. The probability weighted estimator is consistent but may have a large variance. Design consistent estimators based on instrumental variable procedures are compared theoretically and in a Monte Carlo study.

KEY WORDS: informative design, endogenous explanatory variable, instrumental variable

1 Introduction

In a simple random sample, an unbiased estimator of the population regression coefficient is the ordinary least squares (OLS) estimator, and an estimator of its variance is easy to calculate. In many surveys, the elements enter the sample with unequal probabilities. In these cases, the sampling weights, commonly the inverses of the selection probabilities, can be used to construct the probability weighted (PW) estimator. In complex analyses such as regression, the weighted estimator requires a more complicated calculation, and often gives a larger variance than the unweighted version of the estimator. The OLS estimator and PW estimator are straightforward procedures, but for complex sampling designs, the OLS estimator and the PW estimator do not always perform well.

It is known that the presence of errors of measurement in the explanatory variable and the presence of endogenous explanatory variables in the regression model make the OLS estimator inconsistent and biased. For such cases, additional information is needed to obtain consistent parameter estimators. A variable that is correlated with the explanatory variable but uncorrelated with the error is one type of additional information. If a variable meets these two requirements, we call this variable an instrumental variable (IV). The method of instrumental variables has been used for more than sixty years. See Reiersøl (1941, 1945). Sargan's (1958) work and the instrumental variable character of two-stage least squares (2SLS) have made instrumental variable estimation widely used.

In Section 2, the regression models are presented and two common estimators, OLS estimator and PW estimator, are given. In section 3 we introduce the instrumental variable estimator, describe some limiting properties, and describe a test for endogeneity in the instrumental variable procedure. In section 4, a Monte Carlo simulation study is constructed to compare the estimators.

2 Models and Common Estimators

2.1 Regression Model

We assume the finite population to be generated by a random process, called the superpopulation. We will use script $\mathcal{F}$ to denote the finite population, U to denote the set of indices of the finite population, and A to denote the set of indices of the sample. We assume that there is a function $p(\cdot)$ such that $p(A)$ gives the probability of selecting sample A from U.

Consider a regression model relating y_i to $\boldsymbol{x}_i$ as

$$y_i = \boldsymbol{x}_i\boldsymbol{\beta} + e_i, \tag{1}$$

where e_i are independent $(0, \sigma^2)$ random variables independent of x_j for all i and j. The model for the finite population can be written as

$$\begin{aligned} \boldsymbol{y}_N &= \boldsymbol{X}_N\boldsymbol{\beta} + \boldsymbol{e}_N, \\ \boldsymbol{e}_N &\sim (\boldsymbol{0}, \boldsymbol{I}_N\sigma^2), \end{aligned} \tag{2}$$

where $\boldsymbol{y}_N = (y_1, y_2, \ldots, y_N)'$ is the N dimensional vector of values for the dependent variable, $\boldsymbol{X}_N = (\boldsymbol{x}_1', \boldsymbol{x}_2', \ldots, \boldsymbol{x}_N')'$ is the $N \times k$ matrix of values of explanatory variables, and the error vector $\boldsymbol{e}_N = (e_1, e_2, \ldots, e_N)'$ is an N dimensional vector which is independent of $\boldsymbol{X}_N$.

Assume a simple random sample (SRS) of size n is selected from the finite population. Then we can write the model for the sample as

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \\ \boldsymbol{e} &\sim (\boldsymbol{0}, \boldsymbol{I}\sigma^2), \end{aligned} \tag{3}$$

where $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)'$ is the n dimensional column vector of observations, $\boldsymbol{X} = (\boldsymbol{x}_1', \boldsymbol{x}_2', \ldots, \boldsymbol{x}_n')'$ is the $n \times k$ matrix of observations on the explanatory variables, and $\boldsymbol{e} = (e_1, e_2, \ldots, e_n)'$ is the n dimensional error vector. Because the sample is a simple random sample, $\boldsymbol{e}$ is independent of $\boldsymbol{X}$.

2.2 Ordinary Least Squares Estimator

On the basis of model (3), the ordinary least squares (OLS) estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}}_{ols} = (\sum_{i\in A} \boldsymbol{x}_i \boldsymbol{x}_i')^{-1} \sum_{i\in A} \boldsymbol{x}_i y_i = (\boldsymbol{X}'\boldsymbol{X})^{-1} \boldsymbol{X}'\boldsymbol{y}. \tag{4}$$

An estimator of the variance of $\hat{\boldsymbol{\beta}}_{ols}$ is

$$\hat{V}(\hat{\boldsymbol{\beta}}_{ols}) = (\boldsymbol{X}'\boldsymbol{X})^{-1} \hat{\sigma}^2_{ols}, \tag{5}$$

where

$$\hat{\sigma}^2_{ols} = (n-k)^{-1} \sum_{i\in A} \hat{e}^2_{i,ols},$$

k is the dimension of $\boldsymbol{x}_i$ and $\hat{e}_{i,ols} = y_i - \boldsymbol{x}_i \hat{\boldsymbol{\beta}}_{ols}$. The OLS estimator is the best linear unbiased estimator (BLUE) of $\boldsymbol{\beta}$, given model (3).

Assume now that a probability sample is selected with unequal probabilities π_i's. Then, under the model,

$$E\{\hat{\boldsymbol{\beta}}_{ols} - \boldsymbol{\beta}\} \doteq E\{(\sum_{i\in U} \boldsymbol{x}_i \pi_i \boldsymbol{x}_i')^{-1} \sum_{i\in U} \boldsymbol{x}_i \pi_i e_i\}.$$

The approximate bias is zero if $\boldsymbol{x}_i \pi_i$ and e_i are uncorrelated. If $\boldsymbol{x}_i \pi_i$ and e_i are correlated, the expected value of $\boldsymbol{X}'\boldsymbol{e}$ is not zero and the OLS estimator is biased.

2.3 Probability Weighted Estimator

The probability weighted (PW) estimator, constructed with the inverses of the selection probabilities, is

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{PW} &= (\sum_{i\in A} \boldsymbol{x}_i \pi_i^{-1} \boldsymbol{x}_i')^{-1} \sum_{i\in A} \boldsymbol{x}_i \pi_i^{-1} y_i \\ &= (\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1} \boldsymbol{X}'\boldsymbol{W}\boldsymbol{y}, \end{aligned} \tag{6}$$

where

$$\begin{aligned} \boldsymbol{W} &= diag(\pi_1^{-1}, \pi_2^{-1}, \ldots, \pi_n^{-1}) \\ &=: diag(w_1, w_2, \ldots, w_n). \end{aligned}$$

Under the model,

$$E\{\hat{\boldsymbol{\beta}}_{PW} - \boldsymbol{\beta}\} \doteq E\{(\sum_{i\in U} \boldsymbol{x}_i \boldsymbol{x}_i')^{-1} \sum_{i\in U} \boldsymbol{x}_i e_i\}.$$

The probability weighted regression coefficient $\hat{\boldsymbol{\beta}}_{PW}$ is design consistent for the finite population parameter and is a consistent estimator of the superpopulation parameter $\boldsymbol{\beta}$, because $\boldsymbol{x}_i$ and e_i are independent under the superpopulation model.

If the selection is such that $y_i \pi_i^{-1}$ is uncorrelated with $y_j \pi_j^{-1}$ for $i \neq j$, an estimated covariance matrix of $\hat{\boldsymbol{\beta}}_{PW}$ is

$$\hat{V}(\hat{\boldsymbol{\beta}}_{PW}) = (\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1} \boldsymbol{X}'\boldsymbol{W}\hat{\boldsymbol{D}}_{ee,\pi}\boldsymbol{W}\boldsymbol{X} (\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}, \tag{7}$$

where

$$\hat{\boldsymbol{D}}_{ee,\pi} = diag(\hat{e}^2_{1,\pi}, \hat{e}^2_{2,\pi}, \ldots, \hat{e}^2_{n,\pi})$$

and $\hat{e}_{i,\pi} = y_i - \boldsymbol{x}_i \hat{\boldsymbol{\beta}}_{\pi}$. In most cases the variance of the PW estimator is larger than the variance of the OLS estimator.

3 Instrumental Variable Estimator

3.1 Introduction

In the regression model

$$y_i = \boldsymbol{x}_i \boldsymbol{\beta} + e_i, \tag{8}$$

assume some members of $\boldsymbol{x}_i$ are not independent of e_i. If an explanatory variable is correlated with the error term, this explanatory variable is sometimes called an endogenous explanatory variable. The OLS estimator is generally inconsistent when one or more explanatory variables are endogenous in a regression model.

Assume some additional variables, denoted by $\boldsymbol{r}_i$, are available with the superpopulation properties

$$E\{\boldsymbol{r}_i' e_i\} = \boldsymbol{0} \tag{9}$$

$$|E\{\boldsymbol{x}_i' \boldsymbol{r}_i \boldsymbol{r}_i' \boldsymbol{x}_i\}| \neq 0, \tag{10}$$

where $|\boldsymbol{C}|$ is the determinant of the matrix $\boldsymbol{C}$. Variables satisfying (9) and (10) are called instrumental

variables. Thus, an instrumental variable must have two properties: (1) it must be uncorrelated with the error term of the structural equation; (2) it must be correlated with the endogenous explanatory variable. For details see Wooldridge (2000).

3.2 Central Limit Theorem

In this section, we show that the IV estimator is consistent for the population parameter under mild assumptions and has a limiting normal distribution. We begin with Lemma 1 which is adapted from Schenker and Welsh (1988). See also Legg (2006). Let $E\{\cdot|\mathcal{F}_N\}$ be the expectation over repeated samples holding the particular finite population $\mathcal{F}_N$ fixed. Let $V\{\cdot|\mathcal{F}_N\}$ be the variance over repeated samples holding the particular finite population $\mathcal{F}_N$ fixed.

Lemma 1. *Let $\{\mathcal{F}_N\}$ be a sequence of finite populations and let $\boldsymbol{\theta}_N$ be a function on $\mathcal{R}^k$ of the elements of $\mathcal{F}_N$ such that*

$$N^{1/2}(\boldsymbol{\theta}_N - \boldsymbol{\theta}) \xrightarrow{\mathcal{L}} N_k(\mathbf{0}, \boldsymbol{V}_{11}). \tag{11}$$

Let a design and an estimator, $\hat{\boldsymbol{\theta}}_N$, and a sequence of conditional variance matrices $\boldsymbol{V}_{22,N}$ be such that

$$N^{1/2}(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_N)|\mathcal{F}_N \xrightarrow{\mathcal{L}} N_k(\mathbf{0}, \boldsymbol{V}_{22}) \text{ a.s.}, \tag{12}$$

$$\lim_{N\to\infty} \boldsymbol{V}_{22,N} = \boldsymbol{V}_{22} \text{ a.s.}, \tag{13}$$

where $\boldsymbol{V}_{11}+\boldsymbol{V}_{22,N}$ is positive definite for all N. Then

$$N^{1/2}(\boldsymbol{V}_{11}+\boldsymbol{V}_{22,N})^{-1/2}(\hat{\boldsymbol{\theta}}_N-\boldsymbol{\theta}) \xrightarrow{\mathcal{L}} N_k(\mathbf{0}, \boldsymbol{I}_k), \tag{14}$$

where $\boldsymbol{I}_k$ is the $k \times k$ identity matrix.

Theorem 1. *Let $\{(y_i, \boldsymbol{x}_i, \boldsymbol{r}_i)\}$ be a sequence of independent identically distributed random variables with bounded eighth moment. Let $\{U_N, \mathcal{F}_N : N = k+3, k+4, \ldots\}$ be a sequence of finite populations, where U_N is the set of indices identifying the elements and $\mathcal{F}_N = ((y_1, \boldsymbol{x}_1, \boldsymbol{r}_1), \ldots, (y_N, \boldsymbol{x}_N, \boldsymbol{r}_N))$. In the superpopulation y_i is related to $\boldsymbol{x}_i$ through a regression model, that, for the finite population, can be written as*

$$\begin{aligned} \boldsymbol{y}_N &= \boldsymbol{X}_N\boldsymbol{\beta} + \boldsymbol{e}_N, \\ \boldsymbol{e}_N &\sim (\mathbf{0}, \boldsymbol{I}_N\sigma^2). \end{aligned} \tag{15}$$

Assume $\boldsymbol{r}_i$ is independent of e_i and assume that $E\{(\boldsymbol{R}_N\boldsymbol{\Gamma}_N)'\boldsymbol{R}_N\boldsymbol{\Gamma}_N\}$ is nonsingular, where $\boldsymbol{R}_N$ is the $N \times r$ matrix of observations on $\boldsymbol{r}_i$ and $\boldsymbol{\Gamma}_N = \{E(\boldsymbol{R}'_N\boldsymbol{R}_N)\}^{-1}E(\boldsymbol{R}'_N\boldsymbol{X}_N)$. Let $\boldsymbol{t}_j = (y_j, \boldsymbol{x}_j, \boldsymbol{z}_j)$, let

$$\boldsymbol{M}_{T\pi T,N} = n_N^{-1}\boldsymbol{T}'_N\boldsymbol{D}_{\pi,N}\boldsymbol{T}_N \tag{16}$$

and

$$\boldsymbol{M}_{T\pi T} = E\{\boldsymbol{M}_{T\pi T,N}\}, \tag{17}$$

where $\boldsymbol{z}_j = N^{-1}n_N\pi_i^{-1}\boldsymbol{r}_i$, $\boldsymbol{D}_{\pi,N} = diag(\pi_1, \pi_2, \ldots, \pi_N)$, π_i is the inclusion probability for element i, and $\boldsymbol{T}_N = (\boldsymbol{t}'_1, \boldsymbol{t}'_2, \ldots, \boldsymbol{t}'_N)'$. Assume $K_L < Nn_N^{-1}\pi_i < K_N$ for some positive K_L and K_N.

Let $\boldsymbol{d}_j = (y_j, \boldsymbol{z}_j)$. Assume the sequence of sample designs is such that for any $\boldsymbol{d}$ with bounded fourth moments

$$\begin{aligned} \lim_{N\to\infty} \bar{\boldsymbol{d}}_{HT} &= E\{y_j, \boldsymbol{z}_j\} \text{ a.s.}, \\ \lim_{N\to\infty} n_N V\{\bar{\boldsymbol{d}}_{HT} - \bar{\boldsymbol{d}}_N|\mathcal{F}_N\} &= \boldsymbol{V}_{\infty,\tilde{d}\tilde{d}} \text{ a.s.}, \end{aligned} \tag{18}$$

where

$$\bar{\boldsymbol{d}}_{HT} = N^{-1}\sum_{i\in A}\pi_i^{-1}\boldsymbol{d}_i,$$

$\bar{\boldsymbol{d}}_{HT}$ is the Horvitz-Thompson mean of $\boldsymbol{d}$, $\bar{\boldsymbol{d}}_N$ is the finite population mean of $\boldsymbol{d}$, and $V\{\bar{\boldsymbol{d}}_{HT} - \bar{\boldsymbol{d}}_N|\mathcal{F}_N\}$ and $\boldsymbol{V}_{\infty,\tilde{d}\tilde{d}}$ are positive definite. Assume

$$[V\{\bar{\boldsymbol{d}}_{HT}-\bar{\boldsymbol{d}}_N|\mathcal{F}_N\}]^{-1/2}(\bar{\boldsymbol{d}}_{HT}-\bar{\boldsymbol{d}}_N)|\mathcal{F}_N \xrightarrow{\mathcal{L}} N(\mathbf{0}, \boldsymbol{I}) \text{ a.s.}. \tag{19}$$

Let $\hat{V}\{\bar{\boldsymbol{d}}_{HT}\}$ be the Horvitz-Thompson variance estimator of $V\{\bar{\boldsymbol{d}}_{HT}|\mathcal{F}_N\}$, and assume

$$\hat{V}\{\bar{\boldsymbol{d}}_{HT}\} - V\{\bar{\boldsymbol{d}}_{HT}|\mathcal{F}_N\} = o_p(n_N^{-1}) \tag{20}$$

for any $\boldsymbol{d}$ with bounded fourth moments.

Let the instrumental variable estimator be

$$\hat{\boldsymbol{\beta}}_{IV} = \hat{\boldsymbol{L}}_{XZ}n_N^{-1}\boldsymbol{Z}'\boldsymbol{y}, \tag{21}$$

where

$$\hat{\boldsymbol{L}}_{XZ} = [n_N^{-1}\boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X}]^{-1}\boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}.$$

Then

$$\hat{\boldsymbol{\beta}}_{IV} - \boldsymbol{\beta} = \boldsymbol{L}_{XZ}\bar{\boldsymbol{b}} + O_p(n_N^{-1}), \tag{22}$$

where

$$\boldsymbol{L}_{XZ} = [\boldsymbol{M}_{X\pi Z}\boldsymbol{M}_{Z\pi Z}^{-1}\boldsymbol{M}_{Z\pi X}]^{-1}\boldsymbol{M}_{X\pi Z}\boldsymbol{M}_{Z\pi Z}^{-1},$$

$$\bar{\boldsymbol{b}} = n_N^{-1}\sum_{i\in A}\boldsymbol{b}_i,$$

and $\boldsymbol{b}_i = \boldsymbol{z}_i e_i$.

Then

$$[\hat{V}\{\hat{\boldsymbol{\beta}}_{IV}\}]^{-1/2}[\hat{\boldsymbol{\beta}}_{IV}-\boldsymbol{\beta}] \xrightarrow{\mathcal{L}} N(\mathbf{0},\boldsymbol{I}), \quad (23)$$

where

$$\hat{V}\{\hat{\boldsymbol{\beta}}_{IV}\} = \hat{\boldsymbol{L}}_{XZ}[\hat{V}\{\bar{\boldsymbol{b}}\}+n_N^{-2}\boldsymbol{Z}'\boldsymbol{D}_\pi\hat{\boldsymbol{D}}_{ee}\boldsymbol{Z}]\hat{\boldsymbol{L}}'_{XZ},$$

$\hat{V}\{\bar{\boldsymbol{b}}\}$ *is the Horvitz-Thompson estimated variance of* $\bar{\boldsymbol{b}}$ *calculated with* $\hat{\boldsymbol{b}}_i = \boldsymbol{z}_i\hat{e}_i$, $\hat{\boldsymbol{D}}_{ee} = diag(\hat{e}_i^2)$ *and* $\hat{e}_i = y_i - \boldsymbol{x}_i\hat{\boldsymbol{\beta}}_{IV}$.

Proof. Omitted □

3.3 A Test for Endogeneity

In this section, we describe a test for endogeneity in the context of instrumental variable estimation. The general idea of using a pretest to determine an estimation procedure is discussed by Bancroft (1944), Mosteller (1948) and Huntsberger (1955). Suppose we have a regression model written as

$$\begin{aligned}\boldsymbol{y} &= \boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{e}, \quad (24)\\ \boldsymbol{e} &\sim (\mathbf{0},\boldsymbol{I}\sigma^2).\end{aligned}$$

The $\boldsymbol{Z}$ is a known instrumental variable for $\boldsymbol{X}$. For example, in the survey situation, a possible $\boldsymbol{Z}$ is $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{X}$. The two-stage least squares form of the IV estimator constructed using $\boldsymbol{Z}$ can be written as

$$\hat{\boldsymbol{\beta}}_{IV} = (\hat{\boldsymbol{X}}'\hat{\boldsymbol{X}})^{-1}\hat{\boldsymbol{X}}'\boldsymbol{y}, \quad (25)$$

where

$$\hat{\boldsymbol{X}} = \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X}.$$

Wooldridge (2000) gives a test for exogeneity based on the two-stage least squares estimator. We extend the test to the complex survey case. If the design is such that $\pi_i^{-1}e_i$ is independent of $\pi_j^{-1}e_j$, $i \neq j$, an estimated covariance matrix of $\hat{\boldsymbol{\beta}}_{IV}$ is

$$\hat{V}(\hat{\boldsymbol{\beta}}_{IV}) = (\hat{\boldsymbol{X}}'\hat{\boldsymbol{X}})^{-1}\hat{\boldsymbol{X}}'\hat{\boldsymbol{D}}_{ee,IV}\hat{\boldsymbol{X}}(\hat{\boldsymbol{X}}'\hat{\boldsymbol{X}})^{-1}, \quad (26)$$

where

$$\hat{\boldsymbol{X}}'\hat{\boldsymbol{D}}_{ee,IV}\hat{\boldsymbol{X}} = \hat{V}\{n\bar{\boldsymbol{b}}\},$$

$$\hat{\boldsymbol{D}}_{ee,IV} = diag(\hat{e}_{1,IV}^2,\hat{e}_{2,IV}^2,\ldots,\hat{e}_{n,IV}^2),$$

$\bar{\boldsymbol{b}}$ is defined in (22) and $\hat{e}_{i,IV} = y_i - \boldsymbol{x}_i\hat{\boldsymbol{\beta}}_{IV}$.

We describe a test that a set of variables can be used as instrumental variables, given a set that is known to be exogenous. We partition the $\boldsymbol{Z}$ as $(\boldsymbol{Z}_2,\boldsymbol{Z}_3)$. The $\boldsymbol{Z}_2$ is a set of variables known to be exogenous and $\boldsymbol{Z}_3$ is a set for which we wish to test

$$H_0 : E\{\boldsymbol{Z}_3'\boldsymbol{e}\} = \mathbf{0}. \quad (27)$$

The test is the test that $H_0 : \boldsymbol{\delta} = \mathbf{0}$ in the representation

$$\boldsymbol{y} = \hat{\boldsymbol{X}}\boldsymbol{\beta}+(\boldsymbol{Z}_3-\hat{\boldsymbol{Z}}_3)\boldsymbol{\delta}+\boldsymbol{e}^*, \quad (28)$$

where

$$\hat{\boldsymbol{Z}}_3 = \boldsymbol{Z}_2(\boldsymbol{Z}_2'\boldsymbol{Z}_2)^{-1}\boldsymbol{Z}_2'\boldsymbol{Z}_3,$$

$$\hat{\boldsymbol{X}} = \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X},$$

and $\boldsymbol{Z} = (\boldsymbol{Z}_2,\boldsymbol{Z}_3)$. We compute

$$\begin{pmatrix}\hat{\boldsymbol{\beta}}\\ \hat{\boldsymbol{\delta}}\end{pmatrix} = (\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}'\boldsymbol{y}, \quad (29)$$

where $\tilde{\boldsymbol{x}}_i = (\hat{\boldsymbol{x}}_i, \boldsymbol{z}_{3i}-\hat{\boldsymbol{z}}_{3i})$. An estimated covariance matrix is

$$\hat{V}\begin{pmatrix}\hat{\boldsymbol{\beta}}\\ \hat{\boldsymbol{\delta}}\end{pmatrix} = (\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}})^{-1}\hat{V}(\tilde{\boldsymbol{X}}'\boldsymbol{e})(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}})^{-1}, \quad (30)$$

$\hat{V}(\tilde{\boldsymbol{X}}'\boldsymbol{e})$ is a variance estimator calculated with $\tilde{\boldsymbol{X}}'\hat{\boldsymbol{e}}$, where $\hat{e}_i = y_i - (\boldsymbol{x}_i,\boldsymbol{z}_{3i}-\hat{\boldsymbol{z}}_{3i})(\hat{\boldsymbol{\beta}}',\hat{\boldsymbol{\delta}}')'$. The null hypothesis is $H_0 : \boldsymbol{\delta} = \mathbf{0}$ and the test statistic is

$$\hat{\boldsymbol{\delta}}'[\hat{V}(\hat{\boldsymbol{\delta}})]^{-1}\hat{\boldsymbol{\delta}}. \quad (31)$$

If the test is statistically significant at the chosen significance level, we reject the hypothesis that $\boldsymbol{Z}_3$ can be used as an instrumental variable. Under the null model, the distribution of the test statistic (31) is approximately a Chi square with degrees of freedom equal to the dimension of $\boldsymbol{Z}_3$.

4 Monte Carlo Study

4.1 Introduction

To illustrate the instrumental variable procedure, a simulation study was conducted. We create each sample in the simulation by the following selection procedure. Let (x_i,e_i,a_i,u_i) be a vector, where x_i is a normal $(0,0.5)$ random variable, e_i is a normal $(0,0.5)$ random variable, a_i is a normal $(0,0.5)$ random variable, u_i is a uniform $(0,1)$ random variable, and the variables x_i, e_i, a_i, and u_i are mutually independent. Let the selection probability p_i be a function of x_i, e_i and a_i,

$$p_i(x_i,e_i,a_i) = 0.25r(x_i)+1.75r(\psi^{0.5}e_i+[1-\psi]^{0.5}a_i), \quad (32)$$

where

$$r(x) = \begin{cases}0.025 & \text{if } x < 0.2\\ 0.475(x-0.20)+0.025 & \text{if } 0.2 \le x \le 1.2\\ 0.5 & \text{if } x > 1.2\end{cases} \quad (33)$$

and ψ is a parameter that is varied in the experiment. The parameter ψ determines the correlation between π_i and e_i.

If $u_i > p_i$, we reject the vector (x_i, e_i, a_i, u_i). If $u_i \leq p_i$, the vector (x_i, e_i, a_i, u_i) is accepted and y_i is defined by

$$y_i = 0.5 + x_i + e_i. \tag{34}$$

For each sample, we draw 1000 vectors. This procedure gave an expected sample size of about 220. Results are reported for 10000 samples created in this way.

4.2 Instrumental Variable Estimator

Under the regression model, $E\{\sum_{i\in U} e_i\} = 0$ and $E\{\sum_{i\in U} x_i e_i\} = 0$. Thus w_i and $w_i x_i$ are possible instrumental variables. We consider x_i as a potential instrumental variable.

We construct two instrumental variable (IV) estimators. In computing the IV estimators, estimated probabilites $\hat{p}_i$'s are constructed, where $\hat{p}_i$ is the predicted value from the regression of p_i on $(1, r(x_i))$. This procedure is suggested by Pfeffermann and Sverchkov (1999). The first IV estimator is based on four instrumental variables, w_i, $w_i(x_i - \bar{x}_n)$, $w_i\hat{p}_i$, and $w_i\hat{p}_i(x_i - \bar{x}_n)$. The second IV estimator is based on five instrumental variables, w_i, $w_i(x_i - \bar{x}_n)$, $w_i\hat{p}_i$, $w_i\hat{p}_i(x_i - \bar{x}_n)$, and x_i.

The first IV estimator is

$$\hat{\boldsymbol{\beta}}_{IV1} = (\hat{\boldsymbol{X}}_2'\hat{\boldsymbol{X}}_2)^{-1}\hat{\boldsymbol{X}}_2'\boldsymbol{y}, \tag{35}$$

where $\boldsymbol{z}_{2,i} = (w_i, w_i(x_i - \bar{x}_n), w_i\hat{p}_i, w_i\hat{p}_i(x_i - \bar{x}_n))$, $\boldsymbol{Z}_2 = (\boldsymbol{z}_{2,1}, \boldsymbol{z}_{2,2}, \ldots, \boldsymbol{z}_{2,n})'$ be $n \times 4$ matrix, and

$$\hat{\boldsymbol{X}}_2 = \boldsymbol{Z}_2(\boldsymbol{Z}_2'\boldsymbol{Z})_2^{-1}\boldsymbol{Z}_2'\boldsymbol{X}.$$

Because the design is such that $\pi_i^{-1}e_i$ is independent of $\pi_j^{-1}e_j$, $i \neq j$, an estimated covariance matrix of $\hat{\boldsymbol{\beta}}_{IV1}$ is

$$\hat{V}(\hat{\boldsymbol{\beta}}_{IV1}) = (\hat{\boldsymbol{X}}_2'\hat{\boldsymbol{X}}_2)^{-1}\hat{\boldsymbol{X}}_2'\hat{\boldsymbol{D}}_{IV1}\hat{\boldsymbol{X}}_2(\hat{\boldsymbol{X}}_2'\hat{\boldsymbol{X}}_2)^{-1}, \tag{36}$$

where

$$\hat{\boldsymbol{D}}_{IV1} = diag(\hat{e}_{1,IV1}^2, \hat{e}_{2,IV1}^2, \ldots, \hat{e}_{n,IV1}^2),$$

and $\hat{e}_{i,IV1} = y_i - \boldsymbol{x}_i\hat{\boldsymbol{\beta}}_{IV1}$.

$\boldsymbol{Z}_2$ is a set of variables known to be exogenous and if $z_{3,i} = x_i$ is also exogenous, we can construct the second IV estimator of $\boldsymbol{\beta}$

$$\hat{\boldsymbol{\beta}}_{IV2} = (\hat{\boldsymbol{X}}'\hat{\boldsymbol{X}})^{-1}\hat{\boldsymbol{X}}'\boldsymbol{y}, \tag{37}$$

where $\boldsymbol{z}_i = (w_i, w_i(x_i - \bar{x}_n), w_i\hat{p}_i, w_i\hat{p}_i(x_i - \bar{x}_n), z_{3,i})$, $z_{3,i} = x_i$, $\boldsymbol{z}_3 = (x_1, x_2, \ldots, x_n)'$, $\boldsymbol{Z} = (\boldsymbol{Z}_2, \boldsymbol{z}_3)$ is an $n \times 5$ matrix, and

$$\hat{\boldsymbol{X}} = \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X}.$$

An estimated covariance matrix for $\hat{\boldsymbol{\beta}}_{IV2}$ is

$$\hat{V}(\hat{\boldsymbol{\beta}}_{IV2}) = (\hat{\boldsymbol{X}}'\hat{\boldsymbol{X}})^{-1}\hat{\boldsymbol{X}}'\hat{\boldsymbol{D}}_{IV2}\hat{\boldsymbol{X}}(\hat{\boldsymbol{X}}'\hat{\boldsymbol{X}})^{-1}, \tag{38}$$

where

$$\hat{\boldsymbol{D}}_{IV2} = diag(\hat{e}_{1,IV2}^2, \hat{e}_{2,IV2}^2, \ldots, \hat{e}_{n,IV2}^2)$$

and $\hat{e}_{i,IV} = y_i - \boldsymbol{x}_i\hat{\boldsymbol{\beta}}_{IV2}$.

4.3 Preliminary Testing Procedure

We constructed a pretest estimator based on the OLS estimator and the two IV estimators. The pretest procedure is a two-step testing approach. The first step test is a test for importance of weights. This test is based on two regressions: the regression of y_i on $(1, x_i, w_i, w_i x_i)$ (full model) and the regression of y_i on $(1, x_i)$ (reduced model). The F-statistic

$$F_{n-4}^2 = \frac{(SSE_{red} - SSE_{full})/2}{MSE_{full}} \tag{39}$$

is computed, where SSE_{full} and SSE_{red} are error sum of squares for the full model and the reduced model respectively, and MSE_{full} is mean squared error for the full model. If F_{n-4}^2 is not statistically significant, we use $\hat{\boldsymbol{\beta}}_{ols}$, otherwise we proceed to the second test.

The second test is a test for endogeneity. We compute the OLS regression of y_i on $(\tilde{x}_0, \tilde{x}_i, x_i - \hat{x}_i)$ as defined in (29) where $\tilde{x}_0$ is the predicted value from the regression of 1 on $[w_i, w_i(x_i - \bar{x}_n), w_i\hat{p}_i, w_i\hat{p}_i(x_i - \bar{x}_n), x_i]$, $\tilde{x}_i$ is the predicted value from the regression of x_i on $[w_i, w_i(x_i - \bar{x}_n), w_i\hat{p}_i, w_i\hat{p}_i(x_i - \bar{x}_n), x_i]$, and $\hat{x}_i$ is the predicted value from the regression of x_i on $[w_i, w_i(x_i - \bar{x}_n), w_i\hat{p}_i, w_i\hat{p}_i(x_i - \bar{x}_n)]$. The t-statistic for the hypothesis that $H_0 : \delta = 0$ is

$$t = \hat{\delta}/\hat{v}(\hat{\delta}), \tag{40}$$

where $\hat{\delta}$ is the ordinary least squares coefficient for $x_i - \hat{x}_i$ in the regression of (29). Under the null model, the distribution of the t statistic (40) is approximately a normal distribution. If t is not statistically significant, we conclude that x_i can be used as an instrumental variable. Thus the pretest estimator $\hat{\boldsymbol{\beta}}_{pre} = \hat{\boldsymbol{\beta}}_{ols}$, if $F < F_{2,n-4}(\alpha)$. If $F \geq F_{2,n-4}(\alpha)$, the pretest estimator is

$$\hat{\boldsymbol{\beta}}_{pre} = \begin{cases} \hat{\boldsymbol{\beta}}_{IV2} & \text{if } |t| < Z(\alpha/2) \\ \hat{\boldsymbol{\beta}}_{IV1} & \text{if } |t| \geq Z(\alpha/2), \end{cases}$$

where α is the size of the test.

We can compute a standard error for $\hat{\boldsymbol{\beta}}_{pre}$ using the variance estimation procedure appropriate for the estimator chosen. Then an estimated covariance matrix is $\hat{V}(\hat{\boldsymbol{\beta}}_{pre}) = \hat{V}(\hat{\boldsymbol{\beta}}_{ols})$, if $F < F_{2,n-4}(\alpha)$. If $F \geq F_{2,n-4}(\alpha)$,

$$\hat{V}(\hat{\boldsymbol{\beta}}_{pre}) = \begin{cases} \hat{V}(\hat{\boldsymbol{\beta}}_{IV2}) & \text{if } |t| < Z(\alpha/2) \\ \hat{V}(\hat{\boldsymbol{\beta}}_{IV1}) & \text{if } |t| \geq Z(\alpha/2), \end{cases}$$

where $\hat{V}(\hat{\boldsymbol{\beta}}_{IV1})$ is defined in (36) and $\hat{V}(\hat{\boldsymbol{\beta}}_{IV2})$ is defined in (38).

4.4 Simulation Results

Table 1: Monte Carlo Mean Squared Error (×1000) for estimators of β_0 (10,000 samples)

ψ	$\hat{\beta}_{ols,0}$	$\hat{\beta}_{\pi,0}$	$\hat{\beta}_{IV1,0}$	$\hat{\beta}_{IV2,0}$	$\hat{\beta}_{pre,0}$ $\alpha = 0.10$
0	2.33	5.92	5.71	5.33	3.39
.0025	3.35	5.82	5.65	5.18	4.38
.01	6.77	5.71	5.55	5.14	6.97
.02	10.82	5.75	5.53	5.10	8.94
.03	15.16	5.58	5.44	5.08	9.61
.05	23.94	5.60	5.41	4.99	9.35
.07	32.45	5.65	5.47	5.02	8.01
.10	45.11	5.58	5.42	5.06	6.55
.14	62.13	5.60	5.45	5.12	5.58
.17	75.90	5.65	5.53	5.22	5.47
.20	88.22	5.67	5.55	5.18	5.41
.25	109.28	5.42	5.31	4.99	5.17
.30	131.22	5.44	5.34	4.89	5.11
.40	174.09	5.32	5.25	4.89	5.07
.50	217.28	5.26	5.23	4.88	5.07

Table 1 contains the mean squared error for estimators of β_0. Table 2 contains the mean squared error for estimators of β_1. The pretest estimator is for $\alpha = 0.10$. The mean squared error of $\hat{\beta}_{ols,0}$ and $\hat{\beta}_{ols,1}$ are the smallest among estimators of β_0 and β_1, respectively, when $\psi = 0$, that is, when there is no correlation between p_i and e_i. When the correlation between p_i and e_i increases, the mean squared error of $\hat{\beta}_{ols,0}$ and $\hat{\beta}_{ols,1}$ increase because of the squared bias. The IV estimators are more efficient than the PW estimator, because the selection probability p_i is a function of x_i. The pretest estimator is a compromise between alternative IV estimators. As ψ gets larger, the mean squared error of the pretest estimator becomes closer to the mean squared error of the IV estimator. The reason for this is that the pretest procedure rejects the null hypothesis more frequently when the correlation between p_i and e_i increases.

Table 2: Monte Carlo Mean Squared Error (×1000) for estimators of β_1 (10,000 samples)

ψ	$\hat{\beta}_{ols,1}$	$\hat{\beta}_{\pi,1}$	$\hat{\beta}_{IV1,1}$	$\hat{\beta}_{IV2,1}$	$\hat{\beta}_{pre,1}$ $\alpha = 0.10$
0	4.16	9.62	8.53	4.29	5.12
.0025	4.22	9.82	8.71	4.31	5.22
.01	4.30	9.87	8.61	4.32	5.61
.02	4.41	9.71	8.63	4.32	5.93
.03	4.62	9.74	8.64	4.45	6.16
.05	4.66	9.54	8.49	4.34	6.18
.07	4.94	9.80	8.64	4.46	6.49
.10	5.32	9.69	8.57	4.58	6.52
.14	5.92	9.62	8.57	4.69	6.58
.17	6.21	9.41	8.32	4.80	6.42
.20	6.47	9.48	8.39	4.84	6.56
.25	7.04	9.47	8.37	4.96	6.63
.30	7.91	9.30	8.25	5.20	6.66
.40	8.85	8.95	8.05	5.43	6.70
.50	10.29	9.10	8.25	5.76	6.97

Figure 1 is the plot of the mean squared errors of $\hat{\beta}_{ols,0}$, $\hat{\beta}_{IV2,0}$ and $\hat{\beta}_{pre,0}$ relative to the mean squared error of $\hat{\beta}_{IV1,0}$ as a function of the correlation between p_i and e_i. The shape of the mean squared error of the pretest estimator is typical of pretest procedures. In Figure 1 the solid line always equal to one is the mean squared error efficiency of $\hat{\beta}_{IV1,0}$ relative to itself. The $\hat{\beta}_{ols,0}$ is the best if p_i and e_i are independent, but has very poor performance when the correlation between p_i and e_i is large. The IV2 estimator is always better than the IV1 estimator. The pretest estimator has mean squared error that is between that of the OLS estimator and that of the IV estimators. The pretest estimator is never the best, nor the worst, so it is a compromise in terms of mean squared error.

Figure 2 is the plot of the mean squared errors of $\hat{\beta}_{ols,1}$, $\hat{\beta}_{IV2,1}$ and $\hat{\beta}_{pre,1}$ relative to the mean squared error of $\hat{\beta}_{IV1,1}$ for $\alpha = 0.10$. The pretest estimator is always superior to the IV1 estimator because $Cov(p_i x_i, e_i) = 0$ for all parameter sets represented in this plot. If we changed the x-axis to be the correlation between $p_i x_i$ and e_i, we would see the typical

Figure 1: Plot of MSE ratios relative to $\hat{\beta}_{IV0,0}$

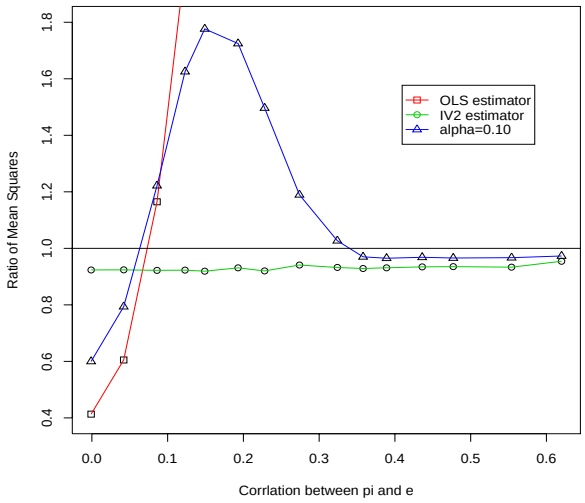
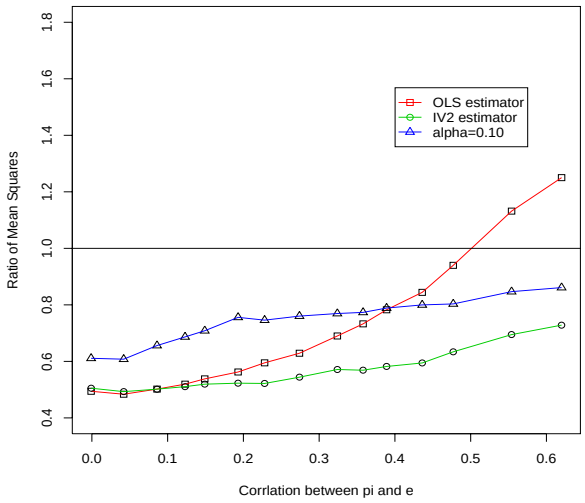


Figure 2: Plot of MSE ratios relative to $\hat{\beta}_{IV1,0}$



pretest procedure shape.

Acknowledgement

This research was supported in part by the USDA Natural Resources Conservation Service cooperative agreement NRCS-68-3A75-4-122.

References

Bancroft, T. A. (1944). On biases in estimation due to the use of preliminary tests of significance. *Annals of Mathematical Statistics*, 15:190–204.

Carter, R. L. and Fuller, W. A. (1980). Instrumental variable estimation of the simple errors-in-variables model. *Journal of the American Statistical Association*, 75:687–692.

Feldstein, M. (1974). Errors in variables: A consistent estimator with smaller mse in finite samples. *Journal of the American Statistical Association*, 69:990–996.

Fuller, W. A. (2006). *Sampling statistics*. Unpublished.

Huntsberger, D. V. (1955). A generalization of a preliminary testing procedure for pooling data. *Annals of Mathematical Statistics*, 26:734–743.

Legg, J. C. (2006). *Estimation for two-phase longitudinal surveys with application to the National Resources Inventory*. PhD dissertation, Iowa State University.

Mosteller, F. (1948). On pooling data. *Journal of the American Statistical Association*, 43:231–242.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review*, 61:317–337.

Pfeffermann, D. and Sverchkov, M. (1999). Parametric and semi-parametric estimation of regression models fitted to survey data. *Sankhya: The Indian Journal of Statistics*, 61:166–186.

Reiersøl, O. (1941). Confluence analysis by means of lag moments and other methods of confluence analysis. *Econometrica*, 9:1–23.

Reiersøl, O. (1945). Confluence analysis by means of instrumental sets of variables. *Arkiv für Matematik, Astronomi och Fysik*, 32A:1–119.

Richardson, D. H. and Wu, D. (1970). Least squares and grouping method estimators in the errors in variables model. *Journal of the American Statistical Association*, 65:724–748.

Richardson, D. H. and Wu, D. (1971). A note on the comparison of ordinary and two-stage least squares estimators. *Econometrica*, 39:973–981.

Sargan, J. D. (1958). The estimation of economic relationships using instrumental variables. *Econometrica*, 26:393–415.

Schenker, N. and Welsh, A. H. (1988). Asymptotic results for multiple imputation. *Annals of Statistics*, 16:1550–1566.

Summers, R. (1965). A capital intensive approach to the small sample properties of various simultaneous equation estimators. *Econometrica*, 33:1–41.

Wooldridge, J. M. (2000). *Introductory Econometrics: A Modern Approach*. South-Western Educational Publishing.

Wu, Y. and Fuller, A. W. (2005). Preliminary testing procedures for regression with survey samples. In *Proceedings of the Survey Research Methods Section, American Statistical Association*.