Preliminary Testing Procedures for Regression with Survey Samples

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

Abstract

We examine preliminary testing 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. In a preliminary testing procedure, we first test for the importance of weights in estimation. If the null hypothesis is accepted, we use the unweighted estimator. Otherwise we incorporate the design weights into the estimation procedure. Pretest procedures we studied use the probability weighted estimator and alternative design consistent weighted estimator.

KEY WORDS: preliminary testing procedure, informative design

1 Introduction

In a simple random sample from a population, an unbiased estimator of the population parameter is the ordinary least squares estimator, and an estimator of its variance is easy to calculate. However, in many surveys, the elements enter the sample with unequal probabilities. In these cases, the sampling weights commonly are the inverses of the selection probabilities. These weights are used to construct the probability weighted estimator. In more complex analyses such as regression, the weighted estimator not only requires a more complicated calculation, but also often gives a larger variance than the unweighted version of the estimator. Therefore, one might question whether weights are necessary in the analysis.

Preliminary testing (pretest) procedures are procedures in which a test of a model assumption is used to decide between two estimation procedures. Bancroft (1944), Huntsberger (1955) and Mosteller (1948) provide details about pretest procedures. The pretest procedure is characterized by a test statistic, T, calculated from the data set. The test T serves the purpose of determining the estimation method. If T is statistically significant at some significance level chosen a priori, a given procedure will be used to estimate a parameter. Otherwise an alternative procedure will be used for calculating the parameter estimator.

We describe a test for the importance of weights and discuss an estimation strategy. If the test statistic is not significant, the unweighted estimator is used, if the test is significant, a weighted estimator is used. When the testing procedure indicates that the weighted analysis is preferred, we consider some consistent weighted estimators that have smaller variances than the probability weighted estimator. One estimator, based on a superpopulation model with error variances determined by values of a covariate, was suggested by Pfeffermann and Sverchkov (1999).

2 Regression Model

We assume the finite population to be generated by some random process, called the superpopulation. The N population values $y_1, y_2, \ldots, y_N$ of the study variable y are generated from 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 x_i with the model for the entire finite population written as

$$\begin{aligned} \boldsymbol{y}_U &= \boldsymbol{X}_U\boldsymbol{\beta} + \boldsymbol{e}_U, \\ \boldsymbol{e}_U &\sim (\boldsymbol{0}, \boldsymbol{I}\sigma^2), \end{aligned} \tag{1}$$

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

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

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

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 the error vector $\boldsymbol{e} = (e_1, e_2, \ldots, e_n)'$ is the n dimensional vector which is independent of $\boldsymbol{X}$.

3 Estimators for the Population Parameter

3.1 Ordinary Least Squares Estimator

On the basis of model (2), the ordinary least squares (OLS) estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}}_{ols} = \left(\sum_{i \in A} \boldsymbol{x}_i' \boldsymbol{x}_i \right)^{-1} \sum_{i \in A} \boldsymbol{x}_i' y_i = \left(\boldsymbol{X}'\boldsymbol{X} \right)^{-1} \boldsymbol{X}'\boldsymbol{y}, \tag{3}$$

where $\boldsymbol{X}$ is the $n \times k$ matrix of the explanatory variables and $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)'$ is the n dimensional column vector of observations. An estimator of variance of $\hat{\boldsymbol{\beta}}_{ols}$ is

$$\hat{V}(\hat{\boldsymbol{\beta}}_{ols}) = \left(\boldsymbol{X}'\boldsymbol{X} \right)^{-1} \hat{\sigma}^2_{ols}, \tag{4}$$

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 $\hat{e}_{i,ols} = y_i - \boldsymbol{x}_i \hat{\boldsymbol{\beta}}_{ols}$. On the basis of model (2), the OLS estimator is the best linear unbiased estimator (BLUE) of $\boldsymbol{\beta}$.

3.2 Probability Weighted Estimator

Assume that a probability sample is selected with unequal probabilities π_i's. The probability weighted estimator, constructed with the inverses of the selection probabilities, is

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{\pi} &= \left(\sum_{i \in A} \boldsymbol{x}_i' \pi_i^{-1} \boldsymbol{x}_i \right)^{-1} \sum_{i \in A} \boldsymbol{x}_i' \pi_i^{-1} y_i \\ &= \left(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X} \right)^{-1} \boldsymbol{X}'\boldsymbol{W}\boldsymbol{y}, \end{aligned} \tag{5}$$

where

$$\begin{aligned} \boldsymbol{W} &= diag(\pi_1^{-1}, \pi_{2^{-1}}, \ldots, \pi_n^{-1}) \\ &=: diag(w_1, w_2, \ldots, w_n). \end{aligned}$$

An estimated covariance matrix of $\hat{\boldsymbol{\beta}}_{\pi}$ is

$$\hat{V}(\hat{\boldsymbol{\beta}}_{\pi}) = \left(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X} \right)^{-1} \boldsymbol{X}'\boldsymbol{W}\hat{\boldsymbol{D}}_{ee,\pi}\boldsymbol{W}\boldsymbol{X} \left(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X} \right)^{-1}, \tag{6}$$

where

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

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

The probability weighted regression coefficient $\hat{\boldsymbol{\beta}}_{\pi}$ 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. In most cases the variances of the probability weighted estimator are larger than the variances of the OLS estimator.

3.3 Generalized Least Squares Estimator

If π_i and e_i are independent in an unequal probability sample, the OLS estimator remains unbiased. If π_i and e_i are correlated, the OLS estimator is biased. The design is called informative if π_i and e_i are correlated. We will construct some consistent weighted estimators that are more efficient than the probability weighted estimator under an informative design by modifying the probability weighted estimator to reduce the variance. Pfeffermann et al (1998), Krieger and Pfeffermann (1997), and Pfeffermann and Sverchkov (1999) considered such approaches. To identify the connection to estimated generalized least squares, let a working model be

$$\pi_i^{-1/2} y_i = \pi_i^{-1/2} \boldsymbol{x}_i \boldsymbol{\beta} + \pi_i^{-1/2} e_i. \tag{7}$$

Under the assumption that is $\boldsymbol{e}$ independent of $\boldsymbol{X}$, regressing $h(\boldsymbol{x}_i)\pi_i^{-1/2} y_i$ on $h(\boldsymbol{x}_i)\pi_i^{-1/2}\boldsymbol{x}_i$ will provide a consistent estimator of $\boldsymbol{\beta}$ for any function $h(\boldsymbol{x}_i)$. We could develop a model for $Var(\pi_i^{-1/2} e_i)$ directly, but it may be convenient to approach estimation in steps, first identifying the portion of $w_i = \pi_i^{-1}$ that is related to $\boldsymbol{x}_i$, then developing a variance model for the modified equation.

Define the generalized least squares (GLS) estimator

$$\hat{\boldsymbol{\beta}}_g = \left(\boldsymbol{X}'\boldsymbol{Q}\boldsymbol{X} \right)^{-1} \boldsymbol{X}'\boldsymbol{Q}\boldsymbol{y}, \tag{8}$$

where

$$\boldsymbol{Q} = diag(q_1, q_2, \ldots, q_n)$$

and $q_i = w_i h(\boldsymbol{x}_i)$.

Pfeffermann and Sverchkov (1999) propose estimators for regression models fitted to survey data. One

estimator is obtained by a two-step procedure: (1) Estimate $\hat{w}(\boldsymbol{x}_i)$ by the regression of w_i on known functions of $\boldsymbol{x}_i$ using the sample measurements. The estimator of $\boldsymbol{\beta}$ is calculated in step (2) as

$$\hat{\boldsymbol{\beta}}_{PS} = \left(\sum_{i\in A} q_i \boldsymbol{x}_i' \boldsymbol{x}_i\right)^{-1} \sum_{i\in A} q_i \boldsymbol{x}_i' y_i = (\boldsymbol{X}'\boldsymbol{Q}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Q}\boldsymbol{y}, \quad (9)$$

where $q_i = w_i \hat{w}^{-1}(\boldsymbol{x}_i)$.

4 A Test for an Informative Design

To determine whether weights should be incorporated into the estimation of the parameters, we consider a test of the null hypothesis:

$$H_0 : E[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}] = E[(\boldsymbol{X}'\boldsymbol{Q}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Q}\boldsymbol{y}]. \quad (10)$$

To test whether or not the two procedures have the same expectation, we can use a standard technique of adding to our basic model the variables for the competing model. We can test the hypothesis by testing the coefficient for $\boldsymbol{Z}$ of the expanded regression model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{e}, \quad (11)$$

where

$$\boldsymbol{Z} = \boldsymbol{Q}\boldsymbol{X}.$$

If OLS provides an unbiased estimator then the coefficient for the weighted vector will be a zero, that is $\boldsymbol{\gamma} = \boldsymbol{0}$. The regression coefficient vector for $\boldsymbol{Z}$ in the regression of $\boldsymbol{y}$ on $(\boldsymbol{X}, \boldsymbol{Z})$ is the regression coefficient for the regression of $\boldsymbol{y} - \boldsymbol{P}_{\boldsymbol{x}}\boldsymbol{y}$ on $\boldsymbol{Q}\boldsymbol{X} - \boldsymbol{P}_{\boldsymbol{x}}\boldsymbol{Q}\boldsymbol{X}$, where $\boldsymbol{P}_{\boldsymbol{x}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ is a projection matrix into the column space of $\boldsymbol{X}$. For details see DuMouchel and Duncan (1983) and Fuller (1984).

If the test indicates that two estimators are estimating different quantities, the usual first response in practice will be to search for subject matter variables to add to the model. If the inclusion of such variables results in a nonsignificant test statistic, the expanded model is accepted. If we can not find such variables, then it is necessary to incorporate the inclusion probabilities into the estimation procedure.

5 Simulation Design

5.1 Introduction

To illustrate the preliminary testing procedure, a simulation study was conducted. We create each sample in the simulation by the following selection procedure. A vector (e_i, x_i, a_i, u_i) is generated, where e_i is a normal $(0, 0.5)$ random variable, x_i is a normal $(0, 0.5)$ random variable, a_i is a normal $(0, 0.5)$ random variable and u_i is a uniform $(0, 1)$ random variable. The variables e_i, x_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,

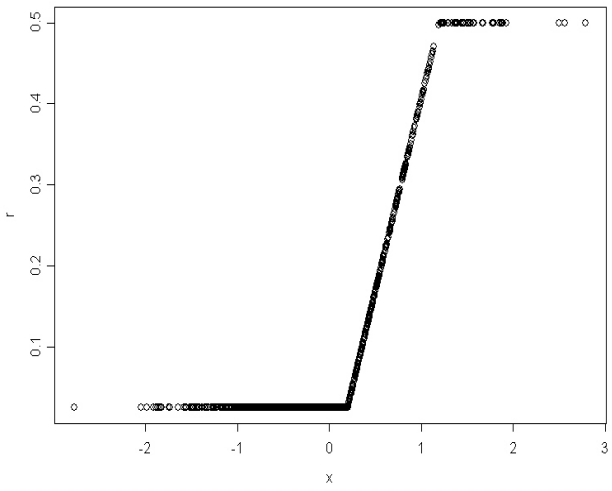
$$p_i(x_i, e_i, a_i) = r(x_i) + r([1-\psi]^{0.5}a_i + \psi^{0.5}e_i), \quad (12)$$

where

$$r(x) = \begin{cases} 0.025 & \text{if } x < 0.2 \\ 0.475(x - 0.20) + 0.025 & \text{if } 0.2 \le x \le 1.2 \\ 0.5 & \text{if } x > 1.2 \end{cases}, \quad (13)$$

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

Figure 1: Plot of r vs. x



From Figure 1, we see that the shape of the function r relative to x is approximately exponential. If $u_i > p_i$, we reject the vector (e_i, x_i, a_i, u_i). If $u_i \le p_i$, the vector (e_i, x_i, a_i, u_i) is accepted and y_i is defined by

$$y_i = 0.5 + x_i + e_i. \quad (14)$$

For each sample, we draw 1000 selections. This procedure gives an expected sample size of about 250. Results are reported for 10000 samples created in this way.

5.2 Pfeffermann-Sverchkov Estimator

In computing the PS estimators, estimated probabilites $\hat{p}_i$'s are constructed, where $\hat{p}_i$ is the predicted value from the regression of p_i on $(1, r(\boldsymbol{x}_i))$. The estimated weights $\hat{w}_i$'s are fitted values from the regression of w_i on $(1, \hat{p}_i^{-1})$. Then the PS estimator is

$$\hat{\boldsymbol{\beta}}_{PS} = (\boldsymbol{X'QX})^{-1}\boldsymbol{X'Qy}, \tag{15}$$

where

$$\boldsymbol{Q} = diag(q_1, q_2, \ldots, q_n)$$

and $q_i = w_i\hat{w}_i^{-1}$. An estimated covariance matrix is

$$\hat{V}(\hat{\boldsymbol{\beta}}_{PS}) = (\boldsymbol{X'QX})^{-1}\boldsymbol{X'Q}\hat{\boldsymbol{D}}_{ee,PS}\boldsymbol{QX}\,(\boldsymbol{X'QX})^{-1}, \tag{16}$$

where

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

and $\hat{e}_{i,PS} = y_i - \boldsymbol{x}_i\hat{\boldsymbol{\beta}}_{PS}$. The PS estimator is not strictly unbiased, but is consistent for the superpopulation parameter.

5.3 Preliminary Testing Procedures

We constructed two pretest estimators. One is based on the ordinary least squares estimator and the probability weighted estimator and the other one is based on the ordinary least squares estimator and the PS estimator.

The preliminary test based on probability weighted estimator is obtained from two regressions: the regression of y_i on $(1, x_i, w_i, w_ix_i)$ (full model) and the regression of y_i on $(1, x_i)$ (reduced model). The F -statistic

$$F^2_{n-4} = \frac{(SSE_{red} - SSE_{full})/2}{MSE_{full}} \tag{17}$$

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 square error for the full model. If F^2_{n-4} is not statistically significant, we use $\hat{\boldsymbol{\beta}}_{ols}$, otherwise we use the probability weighted estimator $\hat{\boldsymbol{\beta}}_{\pi}$. Thus the pretest estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}}_{pre,\pi} = \begin{cases} \hat{\boldsymbol{\beta}}_{ols} & \text{if } F < F^2_{n-4}(\alpha) \\ \hat{\boldsymbol{\beta}}_{\pi} & \text{if } F \geq F^2_{n-4}(\alpha) \end{cases}, \tag{18}$$

where $F^2_{n-4}(\alpha)$ is the $1-\alpha$ quantile of F distribution. $\alpha = 0.05$ and $\alpha = 0.25$ were used in the simulation.

We can compute a standard error for $\hat{\boldsymbol{\beta}}_{pre,\pi}$ using the variance estimation procedure appropriate for the estimator chosen. Thus

$$\hat{V}(\hat{\boldsymbol{\beta}}_{pre,\pi}) = \begin{cases} \hat{V}(\hat{\boldsymbol{\beta}}_{ols}) & \text{if } F < F^2_{n-4}(\alpha) \\ \hat{V}(\hat{\boldsymbol{\beta}}_{\pi}) & \text{if } F \geq F^2_{n-4}(\alpha) \end{cases}, \tag{19}$$

where $\hat{V}(\hat{\boldsymbol{\beta}}_{ols})$ is defined in (5) and $\hat{V}(\hat{\boldsymbol{\beta}}_{\pi})$ is defined in (8). We call the statistic

$$t_{\hat{\boldsymbol{\beta}}} = [\hat{V}(\hat{\boldsymbol{\beta}}_{pre,\pi})]^{-1/2}(\hat{\boldsymbol{\beta}}_{pre,\pi} - \boldsymbol{\beta})$$

the t-statistic for $\hat{\boldsymbol{\beta}}_{pre,\pi}$.

The preliminary test based on PS estimator is similarly obtained from two regressions: the regression of y_i on $(1, x_i, q_i, q_ix_i)$ (full model) and the regression of y_i on $(1, x_i)$ (reduced model). The F -statistic is

$$F^2_{n-4} = \frac{(SSE_{red} - SSE_{full})/2}{MSE_{full}}. \tag{20}$$

The pretest estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}}_{pre,PS} = \begin{cases} \hat{\boldsymbol{\beta}}_{ols} & \text{if } F < F^2_{n-4}(\alpha) \\ \hat{\boldsymbol{\beta}}_{PS} & \text{if } F \geq F^2_{n-4}(\alpha) \end{cases}. \tag{21}$$

The estimated covariance matrix for $\hat{\boldsymbol{\beta}}_{pre,PS}$ is

$$\hat{V}(\hat{\boldsymbol{\beta}}_{pre,PS}) = \begin{cases} \hat{V}(\hat{\boldsymbol{\beta}}_{ols}) & \text{if } F < F^2_{n-4}(\alpha) \\ \hat{V}(\hat{\boldsymbol{\beta}}_{PS}) & \text{if } F \geq F^2_{n-4}(\alpha) \end{cases}, \tag{22}$$

where $\hat{V}(\hat{\boldsymbol{\beta}}_{ols})$ is defined in (5) and $\hat{V}(\hat{\boldsymbol{\beta}}_{PS})$ is defined in (19). The t-statistic for $\hat{\boldsymbol{\beta}}_{pre,PS}$ is

$$t_{\hat{\boldsymbol{\beta}}} = [\hat{V}(\hat{\boldsymbol{\beta}}_{pre,PS})]^{-1/2}(\hat{\boldsymbol{\beta}}_{pre,PS} - \boldsymbol{\beta}).$$

5.4 Simulation Results

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

ψ	$\hat{\beta}_{ols,0}$	$\hat{\beta}_{\pi,0}$	$\hat{\beta}_{PS,0}$	$\hat{\beta}_{pre,\pi,0}$ $\alpha = 0.25$	$\hat{\beta}_{pre,PS,0}$ $\alpha = 0.25$
0	2.35	4.36	3.20	3.53	2.84
.01	3.52	4.42	3.24	4.37	3.52
.02	4.57	4.39	3.22	4.81	3.75
.05	8.09	4.33	3.20	5.17	3.59
.07	10.42	4.24	3.15	4.94	3.35
.10	13.73	4.32	3.22	4.78	3.29
.20	25.28	4.23	3.09	4.25	3.09
.30	36.79	4.16	3.11	4.16	3.11
.50	59.87	4.02	3.04	4.02	3.04

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

ψ	$\hat{\beta}_{ols,1}$	$\hat{\beta}_{\pi,1}$	$\hat{\beta}_{PS,1}$	$\hat{\beta}_{pre,\pi,1}$ $\alpha = 0.25$	$\hat{\beta}_{pre,PS,1}$ $\alpha = 0.25$
0	3.97	8.38	5.95	6.51	5.16
.01	4.48	8.52	5.98	7.40	5.66
.02	5.01	8.47	5.98	7.78	5.90
.05	6.68	8.29	5.96	8.31	6.07
.07	7.68	8.50	5.98	8.62	6.06
.10	9.41	8.31	5.81	8.43	5.83
.20	15.12	8.17	5.82	8.18	5.82
.30	20.69	8.09	5.78	8.09	5.78
.50	31.63	7.68	5.51	7.68	5.51

Table 1 contains the mean square errors of $\hat{\beta}_0$. Table 2 contains the mean square errors of $\hat{\beta}_1$. The pretest estimators are for $\alpha = 0.25$. The mean square errors 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 π_i and e_i. When the correlation between π_i and e_i increases, the mean square errors of $\hat{\beta}_{ols,0}$ and $\hat{\beta}_{ols,1}$ increase because of the squared bias. The estimators $\hat{\beta}_{PS,0}$ and $\hat{\beta}_{PS,1}$ are more efficient than $\hat{\beta}_{\pi,0}$ and $\hat{\beta}_{\pi,1}$ respectively, because the selection probability is a function of x. The pretest estimators based on the probability weighted estimators $\hat{\beta}_{pre,\pi,0}$ and $\hat{\beta}_{pre,\pi,1}$ are uniformly inferior to the pretest estimators based on the PS estimators $\hat{\beta}_{pre,PS,0}$ and $\hat{\beta}_{pre,PS,0}$ in terms of mean square error. When ψ get larger, the mean square errors of $\hat{\beta}_{pre,\pi,0}$ and $\hat{\beta}_{pre,\pi,1}$ are closer to the mean square errors of $\hat{\beta}_{\pi,0}$ and $\hat{\beta}_{\pi,1}$, respectively. The pretest estimators based on the PS estimators and PS estimators have the same tendency. The reason for this trend is that the pretest procedure rejects the null hypothesis more frequently when the correlation between π_i and e_i increases.

Table 3: Monto Carlo Probability that $|t_{\hat{\beta}_0}| > t_{.025}$ (10,000 samples)

ψ	$\hat{\beta}_{ols,0}$	$\hat{\beta}_{\pi,0}$	$\hat{\beta}_{PS,0}$	$\hat{\beta}_{pre,\pi,0}$ $\alpha = 0.25$	$\hat{\beta}_{pre,PS,0}$ $\alpha = 0.25$
0	0.050	0.055	0.053	0.067	0.060
.01	0.110	0.058	0.056	0.103	0.089
.02	0.168	0.058	0.055	0.121	0.099
.05	0.347	0.056	0.056	0.127	0.085
.07	0.457	0.055	0.053	0.107	0.066
.10	0.591	0.055	0.058	0.087	0.062
.20	0.869	0.058	0.053	0.059	0.053
.30	0.964	0.057	0.056	0.057	0.056
.50	0.998	0.059	0.057	0.059	0.057

Table 4: Monto Carlo Probability that $|t_{\hat{\beta}_1}| > t_{.025}$ (10,000 samples)

ψ	$\hat{\beta}_{ols,1}$	$\hat{\beta}_{\pi,1}$	$\hat{\beta}_{PS,1}$	$\hat{\beta}_{pre,\pi,1}$ $\alpha = 0.25$	$\hat{\beta}_{pre,PS,1}$ $\alpha = 0.25$
0	0.053	0.074	0.063	0.078	0.068
.01	0.067	0.075	0.062	0.087	0.072
.02	0.083	0.075	0.061	0.093	0.074
.05	0.133	0.072	0.064	0.090	0.073
.07	0.164	0.078	0.065	0.094	0.070
.10	0.218	0.075	0.065	0.084	0.066
.20	0.383	0.073	0.063	0.074	0.063
.30	0.524	0.077	0.064	0.077	0.064
.50	0.733	0.081	0.066	0.081	0.066

As the simulation results of Table 3 illustrates, for $\alpha = 0.25$, the statistics $t_{\hat{\beta}_{ols,0}}$, $t_{\hat{\beta}_{\pi,0}}$, $t_{\hat{\beta}_{PS,0}}$, $t_{\hat{\beta}_{pre,\pi,0}}$ and $t_{\hat{\beta}_{pre,PS,0}}$ exceed the tabular $t_{.025}$ for Student's t by more than the nominal fraction. As ψ increases, the probabilities of $P(|t_{\hat{\beta}_{pre,\pi,0}}| > t_{.025})$ are closer to the probabilities of $P(|t_{\hat{\beta}_{\pi,0}}| > t_{.025})$. $P(|t_{\hat{\beta}_{pre,PS,0}}| > t_{.025})$ and $P(|t_{\hat{\beta}_{PS,0}}| > t_{.025})$ show the same trend. Table 4 gives the probabilities of the statistics $t_{\hat{\beta}_{ols,1}}$, $t_{\hat{\beta}_{\pi,1}}$, $t_{\hat{\beta}_{PS,1}}$, $t_{\hat{\beta}_{pre,\pi,1}}$ and $t_{\hat{\beta}_{pre,PS,1}}$ exceeding the tabular $t_{.025}$. We can see the same tendency in Table 4.

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

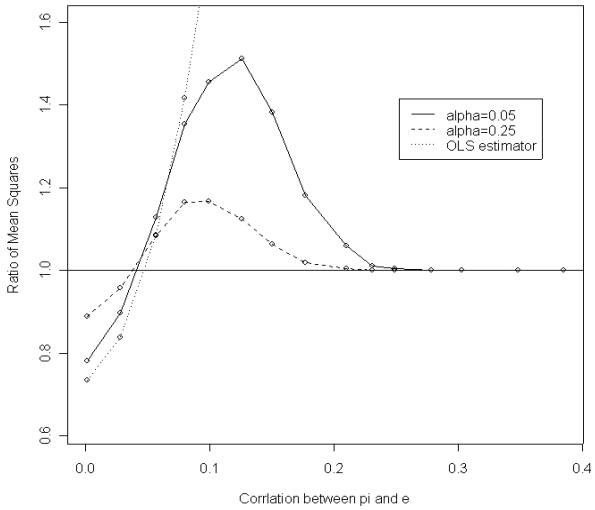
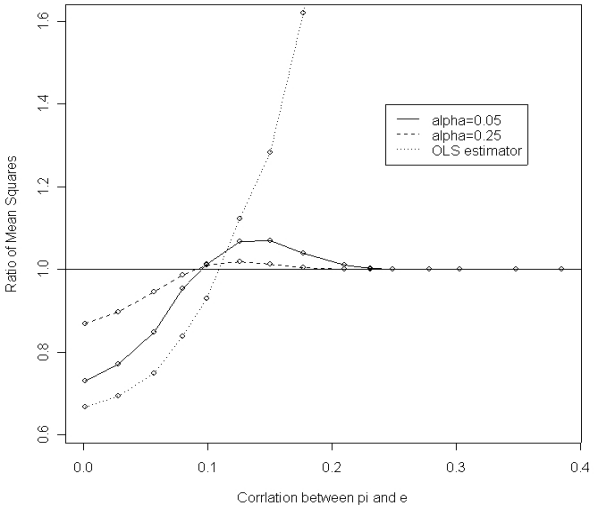


Figure 2 is the plot of mean square error ratios of $\hat{\beta}_{pre,PS,0}$ and $\hat{\beta}_{ols,0}$ relative to $\hat{\beta}_{PS,0}$ as a function of correlation between π_i and e_i for $\alpha = 0.05$ and $\alpha = 0.25$. The shapes are typical of preliminary testing procedures. In Figure 2 the solid line always

Figure 3: Plot of MSE ratios relative to $\hat{\beta}_{PS,1}$



equal to one is the mean square error efficiency of $\hat{\beta}_{pre,PS,0}$ relative to $\hat{\beta}_{PS,0}$ when $\alpha = 1$ and we always reject $\hat{\beta}_{ols,0}$ in the preliminary testing procedures. Since $0.05 < 0.25 < 1$, the curve for the mean square error efficiency of $\hat{\beta}_{pre,PS,0}$ relative to $\hat{\beta}_{PS,0}$ with $\alpha = 0.25$ are generally between the curve of mean square error efficiency of $\hat{\beta}_{pre,PS,0}$ relative to $\hat{\beta}_{PS,0}$ with $\alpha = 0.05$ and the horizontal solid line.

The dotted line is the mean square error ratio of $\hat{\beta}_{ols,0}$ relative to $\hat{\beta}_{PS,0}$ as a function of correlation between π_i and e_i. The $\hat{\beta}_{ols,0}$ is the best if π_i and e_i are independent, but has very poor performance when the correlation between π_i and e_i is large. The pretest estimator $\hat{\beta}_{pre,PS,0}$ is better than $\hat{\beta}_{PS,0}$, but worse than $\hat{\beta}_{ols,0}$ when the correlation is low. But when the correlation gets larger, $\hat{\beta}_{pre,PS,0}$ is worse than $\hat{\beta}_{PS,0}$, but better than $\hat{\beta}_{ols,0}$. The pretest estimator $\hat{\beta}_{pre,PS,0}$ is never the best, nor the worst, so it is a compromise in terms of mean square error.

Figure 3 is the similar plot of the mean square error ratios of $\hat{\beta}_{pre,PS,1}$ and $\hat{\beta}_{ols,1}$ relative to $\hat{\beta}_{PS,1}$ for $\alpha = 0.05$ and 0.25. The curves for the mean square error efficiency of $\hat{\beta}_{pre,PS,1}$ relative to $\hat{\beta}_{PS,1}$ are similar to the curves of the mean square error efficiencyof $\hat{\beta}_{pre,PS,0}$ relative to $\hat{\beta}_{PS,0}$, but the former are smoother than the latter. One reason for this difference is that the bias of $\hat{\beta}_{pre,PS,0}$ relative to the standard error is larger than that of $\hat{\beta}_{pre,PS,1}$.

Acknowledgement

This research was supported in part by the USDA Natural Resources Conservation Service cooperative agreement NRCS-683A754122.

References

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

[2] W. H. DuMouchel and G. J. Duncan. Using sample survey weights in multiple regression analyses of stratified samples. *Journal of the American Statistical Association*, 78:535–543, 1983.

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

[4] A. M. Krieger and D. Pfeffermann. Testing of distribution functions from complex sample surveys. *Journal of Official Statistics*, 13:123–142, 1997.

[5] Frederick Mosteller. On pooling data. *Journal of the American Statistical Association*, 43:231–242, 1948.

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

[7] Krieger A. M. Pfeffermann, D. and Y. Rinott. Parametric distributions of complex survey data under informative probability sampling. *Statistica Sinica*, 8:1087–1114, 1998.

[8] J. N. K. Rao. Alternative estimators in pps sampling for multiple characteristics. *Sankhyā Series A*, 28:47–60, 1966.

[9] C. J. Skinner. Sample models and weights. In *In Proceedings of the Section on Survey Research Methods, American Statistical Association*, pages 133–142, 1994.

[10] Swensson B. Srndal, C. and J. Wretman. *Model Assisted Survey Sampling*. New York: Springer-Verlag, 1992.