Regression Coefficient Estimation in Dual Frame Surveys

Yan Lu*

Abstract

Dual frame surveys, in which independent samples are selected from two frames to decrease survey costs or to improve coverage, can present challenges for regression coefficient estimation because of complex designs and unknown degree of overlap. In this research, we developed four regression coefficient estimators in dual frame surveys. Simulation results show that all the proposed methods work well.

Key Words: Dual frame surveys, cross validation, prediction error, regression coefficient, simulations

1. Introduction

Traditionally, large surveys use a single sampling frame from which the sample is selected. Let $\mathbf{x}$ be the matrix of explanatory values for the sample, $\mathbf{y}$ be the response vector of the sample observations. From design-based perspectives, the finite population quantities of interest $\mathbf{B}$ for regression are the least squares coefficients for the population that minimizes the residual sum of squares $\sum_{i=1}^{i=N}[y_i - \mathbf{x}_i^T\mathbf{B}]^2$. The estimator of $\mathbf{B}$ is

$$\hat{\mathbf{B}} = (\mathbf{x}^T\mathbf{w}\mathbf{x})^{-1}\mathbf{x}^T\mathbf{w}\mathbf{y}, \tag{1}$$

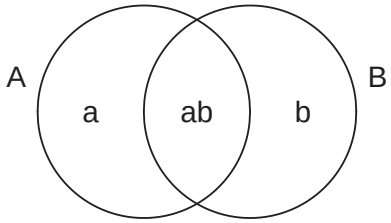
where $\mathbf{w}$ is a diagonal matrix of the sample weights w_i. Linearization, as shown in Shah and Folsom (1977), can be used to estimate the variance of $\hat{\mathbf{B}}$.

Shah and Folsom (1977) discussed regression inference in complex survey data. Holt et al. (1980) studied regression in complex surveys from a maximum likelihood perspective. Skinner and Coker (1996) extended the method of incoporating incomplete observations with missing values of a covariate into the fitting of a linear regression model by maximum likelihood methods to complex surveys. Zieschang (1990), Renssen and Nieuwenbroek (1997), Merkouris (2004) and other researchers also studied combining independent regression estimators from multiple surveys of the same population.

As the population and methods used to collect survey data change, single frame surveys may miss parts of the population. For example, random digit dialing is a popular sampling method. However, as mentioned in Keeter et al. (2010), "The number of Americans who rely solely or mostly on a cell phone has been growing for several years, posing an increasing likelihood that public opinion polls conducted only by landline telephone will be biased". In order to obtain better coverage of the population of interests and to decrease survey costs, there is an increasing interest of U.S. government to employ dual frame design, in which independent samples are taken from two overlapping sampling frames. In a general type of a dual frame survey, each frame can contain units the other frame does not have as well as units in common as depicted in Figure 1. For example, frame A can be a landline frame and frame B can be a cell phone frame. The overlap domain ab includes those people who have both landlines and cellphones. A dual frame survey presents additional challenges to those from a single frame survey because there are now two samples, each with a possibly complex sampling design and may have an unknown degree of overlap. Most research on dual frame surveys concentrate on estimating population totals.

*Yan Lu, Assistant Professor, Department of Mathematics and Statistics, University of New Mexico, Albuquerque, NM, 87131-0001

Figure 1: Frames A and B are both incomplete but overlapping.



In practice, we may want to discover a relationship between diastolic blood pressure as a function of age, gender and ethnicity from a dual frame survey (Metcalf and Soctt, 2009). For applications where prediction is the objective, such as imputing missing values, regression estimation provides a useful tool. This research is to study the regression coefficient estimation in a dual frame survey. One approach considered is to treat the union of two samples as a single sample by adjusted weights and perform regression analysis. Another approach is to consider the regression coefficients of the union as weighted average of the regression coefficients from the two independent samples. Traditional minimizing variance criterion and minimizing prediction error criterion are used to derive the estimators.

This paper is organized as follows. Section 2 gives a brief review of frame work and point estimators in dual frame surveys. Section 3 proposes four methods for regression coefficient estimation. Section 4 presents simulation studies. A discussion of the research is given in Section 5.

2. Background

In a dual frame survey, frame A and frame B together cover the population of interest. The union of these two frames is divided into three mutually exclusive domains, illustrated in Figure 1. Domain a includes the elements contained only in frame A. Domain b includes the elements contained only in frame B. The overlap domain ab includes the elements contained in both frame A and frame B. The population sizes for the frames and domains are denoted by N_A, N_B, N_a, N_b, and N_{ab}, where $N_A = N_a + N_{ab}$, and $N_B = N_b + N_{ab}$. The population size for the union of the two frames N is $N = N_A + N_B - N_{ab}$. Two independent samples $\mathcal{S}_A$ and $\mathcal{S}_B$ are taken from frame A and frame B respectively according to specified probability sampling designs. The probability of unit i being included in $\mathcal{S}_A$ is $\pi_i^A = p\{i \in \mathcal{S}_A\}$. The probability of unit i being included in $\mathcal{S}_B$ is $\pi_i^B = p\{i \in \mathcal{S}_B\}$. The sample sizes for the frames and domains are n_A, n_B, n_a, n_b, n_{ab}^A and n_{ab}^B, where n_{ab}^A and n_{ab}^B represent the sample sizes for the elements of domain ab that were originally taken from frames A and B respectively. So $n_A = n_a + n_{ab}^A$ and $n_B = n_b + n_{ab}^B$.

A number of researchers have proposed methods for combining the information from the two samples in a dual frame survey to estimate population quantities such as total, mean and gross flows, including Hartley (1962, 1974), Fuller and Burmeister (1972), Skinner (1991), Skinner and Rao (1996) and Lu and Lohr (2010) etc,. Lohr and Rao (2000) summarized estimators used for estimating population totals in cross-sectional dual frame surveys.

In the following, we review the pseudo-maximum likelihood (PML) estimator proposed by Skinner and Rao (1996), which we will use in our proposed method 1. Skinner and Rao (1996) considered estimators under complex designs where the same weights are used for all variables. They modified the maximum likelihood estimator for a simple random sample to obtain a PML estimator for complex designs and suggested the following estimator:

$$\hat{Y}_{PML} = \frac{N_A - \hat{N}_{ab,PML}}{\hat{N}_a}\hat{Y}_a + \frac{\hat{N}_{ab,PML}}{\hat{N}_{ab}}\hat{Y}_{ab} + \frac{N_B - \hat{N}_{ab,PML}}{\hat{N}_b}\hat{Y}_b, \tag{2}$$

where $\hat{N}_a$, $\hat{Y}_a$, $\hat{N}_b$ and $\hat{Y}_b$ are standard basic estimators, $\hat{Y}_{ab} = \theta\hat{Y}^A_{ab} + (1-\theta)\hat{Y}^B_{ab}$, and $\hat{N}_{ab} = \theta\hat{N}^A_{ab} + (1-\theta)\hat{N}^B_{ab}$. The estimator $\hat{N}_{ab,PML}$ is a function of $\hat{N}^A_{ab}$, $\hat{N}^B_{ab}$ and θ, and is the smaller root of the quadratic equation

$$\left[\frac{\theta}{N_B} + \frac{(1-\theta)}{N_A}\right]x^2 - \left[1 + \theta\frac{\hat{N}^A_{ab}}{N_B} + (1-\theta)\frac{\hat{N}^B_{ab}}{N_A}\right]x + \left[\theta\hat{N}^A_{ab} + (1-\theta)\hat{N}^B_{ab}\right] = 0, \tag{3}$$

where

$$\theta_P = \frac{\hat{N}_a N_B v(\hat{N}^B_{ab})}{\hat{N}_a N_B v(\hat{N}^B_{ab}) + \hat{N}_b N_A v(\hat{N}^A_{ab})} \tag{4}$$

is chosen to minimize the asymptotic variance of $\hat{N}_{ab,PML}(\theta)$ and $v(\cdot)$ is the variance of $(\cdot)$.

3. Estimators of Regression Coefficients

In this section, we propose four estimators for regression coefficient in dual frame surveys. We assume that the underlying regression models are the same in the three domains. Method 1 and method 2 consider the union of the two samples as a single sample using adjusted weights. Method 3 and method 4 consider a weighted average of independent regression coefficient estimates from sample A and sample B.

3.1 Method 1 and Method 2

Method 1 is a natural extension of the cross-sectional estimator suggested by Skinner and Rao (1996). The optimal variable θ_P in (4) is used to reweight the observations in the overlap domain in order to construct a pseudo sample. Let $\mathbf{y}$ be a sample of response values from $\mathcal{S}_A \cup \mathcal{S}_B$ with $\mathbf{y} = (t(\mathbf{y}^A), t(\mathbf{y}^B))^T$, $\mathbf{y}^A$ and $\mathbf{y}^B$ are the response vector of $\mathcal{S}_A$ and $\mathcal{S}_B$ respectively. Let $\mathbf{x}$ be the sample design matrix of $\mathcal{S}_A \cup \mathcal{S}_B$ defined as $(\mathbf{x}_A^T, \mathbf{x}_B^T)^T$, with $\mathbf{x}_A$ and $\mathbf{x}_B$ be the design matrix from $\mathcal{S}_A$ and $\mathcal{S}_B$ respectively. The fitted value of elements in $A \cup B$ for method 1 is

$$\hat{\mathbf{y}} = \mathbf{x}\hat{\mathbf{B}}. \tag{5}$$

Let $\mathbf{w}^*$ be the diagonal matrix of the modified sample weights w_i^*, with

$$w_i^* = \begin{cases} w_i, & \text{if} \quad i \in a, \\ \theta w_i, & \text{if} \quad i \in ab \text{ and } i \in \mathcal{S}_A, \\ (1-\theta)w_i, & \text{if} \quad i \in ab \text{ and } i \in \mathcal{S}_B, \\ w_i, & \text{if} \quad i \in b, \end{cases} \tag{6}$$

where $\theta = \theta_P$ for method 1. Apply (1), the proposed regression coefficient estimator is

$$\hat{\mathbf{B}} = (\mathbf{x}^T\mathbf{w}^*\mathbf{x})^{-1}\mathbf{x}^T\mathbf{w}^*\mathbf{y}. \tag{7}$$

In method 2, we use cross-validation (CV) to derive a fully data-driven θ selection procedure. Applying (5), the weighted prediction sum of squares is as follows

$$\begin{aligned} CV(\theta) &= \sum_{i\in\mathcal{S}_A\cup\mathcal{S}_B} w_i(y_i-\hat{y}_{(i)})^2 \\ &= \sum_{i=1}^{n_a} w_i(\frac{e_i}{1-h_{ii}})^2 + \sum_{i=1}^{n_{ab}^A}\theta w_i(\frac{e_i}{1-h_{ii}})^2 \\ &+ \sum_{j=1}^{n_{ab}^B}(1-\theta)w_j(\frac{e_j}{1-h_{jj}})^2 + \sum_{j=1}^{n_b} w_j(\frac{e_j}{1-h_{jj}})^2, \end{aligned}$$

where $\hat{y}_{(i)}$ is the estimate computed without using the ith observation $(\mathbf{x}_i, y_i)$, $e_i = y_i - \hat{y}_i$, $y_i - \hat{y}_{(i)}$ is the deleted residual d_i, which is equivalent to $e_i/(1-h_{ii})$ in complex surveys, h_{ii} is the ith diagonal element of the hat matrix $\mathbf{H}$ with $\mathbf{H} = \mathbf{x}(\mathbf{x}^T\mathbf{w}\mathbf{x})^{-1}\mathbf{x}^T\mathbf{w}$. The idea behind the cross validation method is that the ith observation is treated as an additional observation for prediction and $CV(\theta)$ measures the quality of predictions. In practice, we set up a grid between $(0, 1)$ to find the optimal θ that minimize the CV quantity.

3.2 Method 3 and Method 4

Method 3 and method 4 consider a weighted average of independent regression coefficient estimates from $\mathcal{S}_A$ and $\mathcal{S}_B$. Denote the regression fit of elements in frame A by $\hat{\mathbf{y}}^A = \mathbf{x}^A\hat{\mathbf{B}}^A$ and regression fit of elements in frame B by $\hat{y}^B = \mathbf{x}^B\hat{\mathbf{B}}^B$. The regression fit for elements in $A \cup B$ is

$$\hat{\mathbf{y}} = \mathbf{x}\hat{\mathbf{B}} \text{ with } \hat{\mathbf{B}} = \lambda\hat{\mathbf{B}}^A + (1-\lambda)\hat{\mathbf{B}}^B. \tag{8}$$

Method 3 chooses λ to minimize $\sum_{i=0}^{k-1} v(\hat{B}_i)$, with $v(\hat{B}_i) = \lambda^2 v(\hat{B}_i^A) + (1-\lambda)^2 v(\hat{B}_i^B)$ and k be the number of predictor variables. $v(\hat{B}_i^A)$ can be estimated using linearization suggested by Shah and Folsom (1977) as follows:

$$\hat{v}(\hat{\mathbf{B}}) = (\mathbf{x}^T\mathbf{w}\mathbf{x})^{-1}\hat{v}(\sum_{i\in S} w_i\mathbf{q}_i)(\mathbf{x}^T\mathbf{w}\mathbf{x})^{-1}, \tag{9}$$

where $\mathbf{q}_i = \mathbf{x}_i^T(y_i - \mathbf{x}_i^T\hat{\mathbf{B}})$. Taking derivative of λ and set to zero, the optimal λ derived by method 3 is

$$\tilde{\lambda}_P = \frac{\sum_{i=0}^{k-1} v(B_i^B)}{\sum_{i=0}^{k-1} v(B_i^A) + \sum_{i=0}^{k-1} v(B_i^B)}. \tag{10}$$

Method 3 considers minimizing the variance of a linear combination of $\hat{\mathbf{B}}^A$ and $\hat{\mathbf{B}}^B$. The optimal λ found by method 3 may not be "optimal" from a prediction perspective.

In Method 4, we consider minimizing the prediction error instead of minimizing variance in Method 3. Using (8), the deleted residual d_i is

$$\begin{aligned} d_i &= y_i - \hat{y}_{i(i)} \\ &= y_i - \mathbf{x}\hat{\mathbf{B}}_{(i)} \\ &= y_i - \lambda\mathbf{x}\hat{\mathbf{B}}_{(i)}^A - (1-\lambda)\mathbf{x}\hat{\mathbf{B}}_{(i)}^B \\ &= \lambda(y_i - \mathbf{x}\hat{\mathbf{B}}_{(i)}^A) + (1-\lambda)(y_i - \mathbf{x}\hat{\mathbf{B}}_{(i)}^B) \\ &= \begin{cases} \lambda(d_i^A) + (1-\lambda)(e_i^B) & \text{if } i \in A \\ \lambda(e_i^A) + (1-\lambda)(d_i^B) & \text{if } i \in B \end{cases} \end{aligned}$$

where $\hat{B}_{(i)}$ is the coefficient computed without using the ith observation $(\mathbf{x}_i, y_i)$, $\hat{B}^F_{(i)}$, $F \in \{A, B\}$, is the coefficient computed without using the ith observation $(\mathbf{x}^F_i, y_i)$, $d^F_i = e^F_i/(1-h^F_{ii})$, h^F_{ii} is the ith diagonal element of the hat matrix related to sample F. The weighted prediction sum of squares is as follows

$$\begin{aligned} CV(\lambda) &= \sum_{i \in \mathcal{S}_A \cup \mathcal{S}_B} w_i (y_i - \hat{y}_{(i)})^2 \\ &= \sum_{i=1}^{n_A} w_i (\lambda \frac{e^A_i}{1-h^A_{ii}} + (1-\lambda) e^B_i)^2 + \sum_{j=1}^{n_B} w_j (\lambda e^A_j + (1-\lambda) \frac{e^B_j}{1-h^B_{jj}})^2. \end{aligned}$$

Taking derivative of λ and set to zero, the optimal λ from method 4 is

$$\hat{\lambda}_P = \frac{-\left(\sum_{i=1}^{n_A} w_i e^B_i (\frac{e^A_i}{1-h^A_{ii}} - e^B_i) + \sum_{j=1}^{n_B} w_j \frac{e^B_j}{1-h^B_{jj}} (e^A_j - \frac{e^B_j}{1-h^B_{jj}}) \right)}{\sum_{i=1}^{n_A} w_i (\frac{e^A_i}{1-h^A_{ii}} - e^B_i)^2 + \sum_{j=1}^{n_B} w_j (e^A_j - \frac{e^B_j}{1-h^B_{jj}})^2}. \tag{11}$$

4. Simulations

In this section, a small simulation study has been conducted to investigate the finite sample properties of the four proposed regression coefficient estimators. The simulation set up is similar as Harms and Duchesne (2010).

4.1 Comparison of the Four Methods

The following equation is used to generate the population

$$y_i = \beta_0 + \beta_1 t_i + \epsilon_i, \; i = 1, ..., 1000, \tag{12}$$

where each population has $N = 1000$ values of t_i which is equally spaced in the interval $[0, 1]$ and random errors are from the normal distribution with mean 0 and constant variance σ^2. First, we generate population of $A \cup B$ by setting $t \in [0, 1]$. Frame A is defined by setting $t \in [0, 0.7]$ and frame B is defined by setting $t \in [0.3, 1]$. Note, When $t \in [0.3, 0.7]$, frame A and frame B overlapped.

The simulation study was performed with factors: (1) σ : 1 and 0.4; (2) Sampling rate f: $5\%, 10\%$ and 20%; (3) Sampling plan: Poisson sampling scheme (unequal probability design). The sampling weights w_i of poisson sampling scheme have been chosen such that the weights are proportional to the auxiliary variable $z_i = (y_i + 2)(t_i + 2)$ and $\sum_{A \cup B} 1/w_i = E(n_s) = N * f$. Note, elements in the overlap domain have the same weights; (4) Method: method 1 to 4.

Simulation does $L = 1000$ times for each setting. Each time, we generate a population based on model (12), then use Poisson sampling to draw two samples from frame A and frame B respectively. The regression coefficient estimates using the four methods and variance estimates using (9) are calculated. In Table (1) and Table (2), $\hat{\beta}_0$ is the average value of the estimates of β_0 from the 1000 replications; SE($\hat{\beta}_0$) is the sample standard error and is considered as the true standard deviation of $\hat{\beta}_0$; $\sqrt{\widehat{V(\beta_0)}}$ is the average standard error of $\hat{\beta}_0$ using (9) from the 1000 replications; Numbers in parenthesis are the sample standard error. Similarly interpret the quantities related to β_1. Table (1) and Table (2) report the performance of the proposed estimators under Poisson sampling scheme for different settings.

Table 1: Performance of the Four Methods: Simulation Result 1

$\sigma = 1, \beta_0 = 5, \beta_1 = 1, N = 1000$							
Method	Sampling rate	$\hat{\beta}_0$	SE($\hat{\beta}_0$)	$\sqrt{\widehat{V(\beta_0)}}$	$\hat{\beta}_1$	SE($\hat{\beta}_1$)	$\sqrt{\widehat{V(\beta_1)}}$
Method 1	5%	4.9970	.2552	$.2418_{(.0454)}$	1.0063	.4272	$.4126_{(.0669)}$
	10%	5.0046	.1749	$.1731_{(.0241)}$	.9859	.2947	$.2947_{(.03515)}$
	20%	5.0092	.1267	$.1235_{(.0114)}$	.9913	.2166	$.2095_{(.0166)}$
Method 2	5%	5.0046	.2469	$.2373_{(.0442)}$	1.0016	.4251	$.4054_{(.0638)}$
	10%	5.0099	.1783	$.1737_{(.0245)}$	.9852	.3023	$.2954_{(.0351)}$
	20%	5.0091	.1282	$.1238_{(.0120)}$	.9857	.2165	$.2098_{(.0177)}$
Method 3	5%	5.0260	.3122	$.2688_{(.0427)}$	.9650	.5605	$.4954_{(.0653)}$
	10%	5.0051	.2020	$.1948_{(.0224)}$	.9904	.3618	$.3572_{(.0345)}$
	20%	5.0031	.1451	$.1402_{(.0115)}$	.9936	.2547	$.2554_{(.0172)}$
Method 4	5%	5.0072	.2652	$.3304_{(.1714)}$	1.0013	.4591	$.5902_{(.2811)}$
	10%	5.0032	.2031	$.2421_{(.1508)}$	1.0037	.3487	$.4296_{(.2466)}$
	20%	4.9965	.1388	$.1695_{(.0786)}$	1.0082	.2375	$.3009_{(.1245)}$

From Table (1) and Table (2), we see that the point estimates and variance estimates from the four methods are all very close to the true value, indicating that our estimators perform well. On the other hand, we observe that Method 1, 2 and 3 give smaller variance estimates than the true value, while method 4 gives a little larger variance estimates.

4.2 Assumption of regression function in domains

In the four proposed Methods, we assume that the regression function in the three domains are the same. Therefore, by combining the information from both frame A and frame B, we would have more degrees of freedom in estimating the regression coefficients. If different domain has different underlying regression function, combining the information from two frames to derive a unified regression function is not appropriate. In such case, the residual plot would present a pattern related to domains. In the following, a simulated data was used to study this issue.

Assume $\epsilon \sim N(0, .16)$, we generate a data using $y = 3 + 5t + \epsilon$ in domain a by setting $t \in (0, .3)$, $y = 3 + 8t + \epsilon$ in domain ab by setting $t \in (.3, .7)$, and $y = 3 + 5t + \epsilon$ in domain b by setting $t \in (.7, 1)$. The fitted regression line by using method 2 is

$$\hat{y} = 3.6269 + 4.9116t. \tag{13}$$

Figure 2 presents the scatterplot together with the fitted regression line. Figure 3 presents the residual plot. From Figure 2 and Figure 3, we see an obvious pattern related to domains that most of residuals in domain ab are positive and most of residuals in domain a and domain b are negative. This suggests that the assumption of same regression model in the three domains is violated. Therefore, the proposed methods are not appropriate. In such situation, we would suggest fit the regression lines by different domains.

5. Discussion

It is becoming more difficult, for a single sampling frame to include the entire population of interest and to be inexpensive to sample. As a result, dual frame surveys are becoming

Table 2: Performance of the Four Methods: Simulation Result 2

$\sigma = .4, \beta_0 = 5, \beta_1 = 1, N = 1000$							
Method	Sampling rate	$\hat{\beta}_0$	SE($\hat{\beta}_0$)	$\sqrt{\widehat{V(\beta_0)}}$	$\hat{\beta}_1$	SE($\hat{\beta}_1$)	$\sqrt{\widehat{V(\beta_1)}}$
Method 1	5%	4.9999	.0964	$.0949_{(.0164)}$	.9992	.1615	$.1620_{(.0242)}$
	10%	4.9969	.0690	$.0677_{(.0084)}$	1.0089	.1191	$.1150_{(.0121)}$
	20%	5.0016	.0498	$.0485_{(.0041)}$	.9975	.0817	$.0821_{(.0058)}$
Method 2	5%	5.0044	.1015	$.0949_{(.0159)}$	.9954	.1710	$.1631_{(.0237)}$
	10%	5.0042	.0706	$.0679_{(.0084)}$	.9935	.1196	$.1152_{(.0118)}$
	20%	5.0017	.0482	$.0485_{(.0039)}$	.9978	.0816	$.0822_{(.0057)}$
Method 3	5%	5.0058	.1174	$.1061_{(.0156)}$	.9911	.2070	$.1951_{(.0239)}$
	10%	5.0042	.0847	$.0762_{(.0084)}$	.9913	.1528	$.1401_{(.0127)}$
	20%	5.0032	.0577	$.0547_{(.0039)}$	.9938	.1022	$.0999_{(.0059)}$
Method 4	5%	5.0003	.1076	$.1347_{(.0849)}$	.9973	.1846	$.2403_{(.1415)}$
	10%	5.0046	.0745	$.0910_{(.0513)}$	.9935	.1286	$.1624_{(.0900)}$
	20%	5.0035	.0517	$.0663_{(.0461)}$	.9963	.0889	$.1177_{(.0757)}$

Figure 2: Scatterplot of the data from $A \cup B$ with the fitted regression line

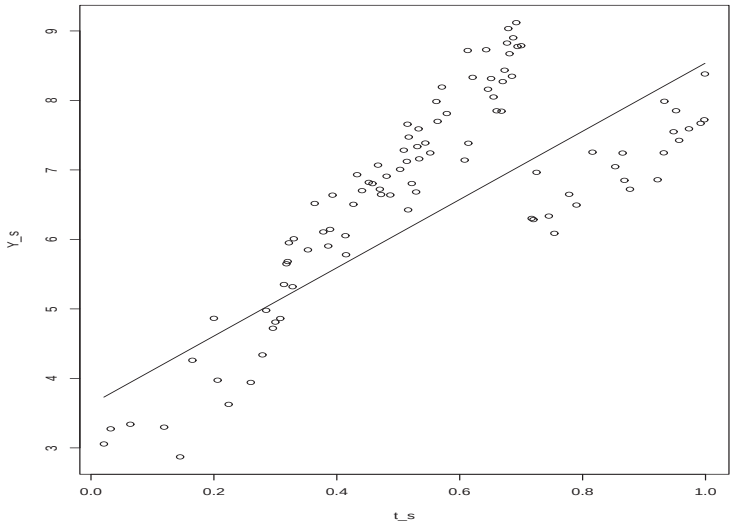
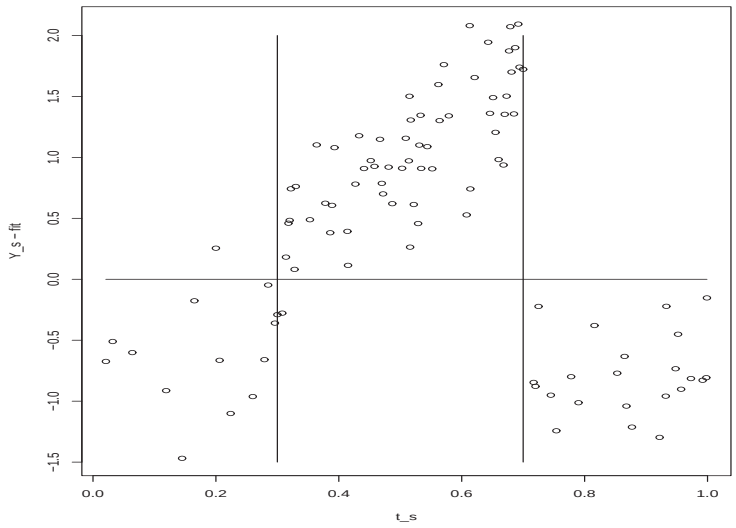


Figure 3: Plot of residual v.s predictor variable t, the vertical lines divide the union of two frames into three nonoverlap domains



more common. Such surveys require new methods for analyzing the regression aspects of the data.

In this research, we propose four regression coefficient estimators in dual frame surveys. Simulation results show that all the four proposed methods work well. Method 1, using the PML value of θ_P, uses the same value of θ for each responses. While, in Method 2, 3 and 4, the value of θ or λ depends on the response variable y, therefore will be differ for each regression model. Thus, Method 1 has the advantage that the same set of weights is used for every response and every model. We also observe that Method 1, 2 and 3 give smaller variance estimates than the true value, while method 4 gives a little larger variance estimates. Method 1 and Method 2 consider a pseudo sample by adjusting weights in the overlap domain. However, by treating θ_P as a constant when constructing the pseudo sample, we miss the part of additional variation by estimating $\hat{\theta}_P$, which is difficult to estimate. Method 1 and Method 3 consider minimizing variance. From the above reasons, Method 1, 2 and 3 tend to have smaller variance. Method 4 considers a weighted combination of independent regression coefficient estimators from the two samples and uses minimizing prediction error criterion. From the limited simulation result, we observe that method 4 provides a little larger variance estimates. Although all the four methods work well, we recommend method 2 and method 4 for prediction purpose in practice.

Our research is done in the context of survey sampling, but they also apply to other settings in which data could be combined from two independent sources and could be extended to more than two surveys.

References

Fuller, W. A. and Burmeister, L. F. (1972), "Estimators for Samples Selected from Two Overlapping Frames," in *ASA Proceedings of the Social Statistics Section*, American Statistical Association, pp. 245–249.

Harms, T. and Duchesne, P. (2010), "On Kernel Nonparametric Regression Designed for Complex Survey Data," *Metrika*, 72, 111–138.

Hartley, H. O. (1962), "Multiple Frame Surveys," in *ASA Proceedings of the Social Statistics Section*, American Statistical Association, pp. 203–206.

— (1974), "Multiple Frame Methodology and Selected Applications," *Sankhyā, Series C*, 36, 99–118.

Holt, D., Smith, T. M. F., and Winter, P. D. (1980), "Regression Analysis of Data from Complex Surveys," *Journal of the Royal Statistical Society. Series A (General)*, 143, 474–487.

Keeter, S., Dimock, M., and Christian, L. (2010), "The Growing Gap between Landline and Dual Frame Election Polls," *available at pewresearch.org)*, November 22.

Lohr, S. L. and Rao, J. N. K. (2000), "Inference from Dual Frame Surveys," *Journal of the American Statistical Association*, 95, 271–280.

Lu, Y. and Lohr, S. L. (2010), "Gross Flow Estimation in Dual Frame Surveys," *Survey Methodology*, 36, 13–22.

Merkouris, T. (2004), "Combining Independent Regression Estimators from Multiple Surveys," *Journal of the American Statistical Association,*, 99, 1131–1139.

Metcalf, P. and Soctt, A. (2009), "Using multiple frames in health surveys," *Statistics in Medicine*, 28, 1512–1523.

Renssen, R. H. and Nieuwenbroek, N. J. (1997), "Aligning Estimates for Com mon Variables in Two or More Sample Surveys," *Journal of the American Statistical Association*, 92.

Shah, B. V., M. M. H. and Folsom, R. E. (1977), "Inference about regression models from sample survey data," *Bulletin of the international statistical institute*, 47, 43–57.

Skinner, C. J. (1991), "On the Efficiency of Raking Ratio Estimation for Multiple Frame Surveys," *Journal of the American Statistical Association*, 86, 779–784.

Skinner, C. J. and Coker, O. (1996), "Regression Analysis for Complex Survey Data with Missing Values of a Covariate," *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 159, 265–274.

Skinner, C. J. and Rao, J. N. K. (1996), "Estimation in Dual Frame Surveys with Complex Designs," *Journal of the American Statistical Association*, 91, 349–356.

Zieschang, K. D. (1990), "Sample Weighting Methods and Estimation of Totals in the Consumer Expenditures Survey," *Journal of the American Statistical Association*, 85, 986–1001.