Multiple Regression Analysis with Data from Complex Survey

Esher Hsu [1], Chiu-Hui Lee [2], Chien-Ming Chen [3]

[1]Associate Professor, Department of Statistics, National Taipei University, Taipei, Taiwan

[2]Graduate Student, Department of Statistics, National Taipei University, Taipei, Taiwan

[3]Graduate Student, Department of Statistics, National Taipei University, Taipei, Taiwan

Abstract

This study explores multiple regression analysis with complex survey data. Four methods of multiple regression analysis, namely, ordinary least squares, weighted least squares, probability weighted least squares, and Quasi-Aitken probability weighted least squares are proposed for comparison by Monte Carlo approach to compare their efficiency based upon bias, variance, and MSE. The data from "Taiwan Social Change Survey 2007" collected under a stratified unequal probability sampling were used for empirical analysis to compare four proposed methods based upon the estimated regression coefficients and RMSE. The simulation results show that probability weighted least squares estimator and Quasi-Aitken weighted least square estimator perform better than others under the unequal probability design. The empirical results consist with the simulation results. The empirical results show that the education years of respondents in Taiwan has significant negative relationship with their age but has positive relationship with their parents' education years.

Keywords: Multiple Regression Analysis; Stratified weighted least squares estimator; Probability weighted least squares estimator; Quasi-Aitken weighted least square estimator; Complex Survey; Social Change

1. Introduction

The sampling design is getting more important along with the increasing demand of precise data for making a better decision which thus has boosted the use of complex

survey in practice and has also raised the importance of unequal probability sampling as well. In principle, the statistical analysis has to be adjusted along with the sampling design to obtain a better statistical inference. In order to simplify the process of statistical inference, the mechanism of the sampling design is usually ignored. That may cause biased estimation or obtain a wrong conclusion. It has occurred frequently, such as, the estimator with simple random sampling used for the data collected under unequal probability sampling. Recently, regression analysis with complex surveys has become popular. For regression analysis, traditional estimators, such as least squares estimator, used with data collected under complex survey may reduce the accuracy of the statistical analysis.

Fuller and Wu (2005) proposed a regression analysis with survey samples. Fuller and Wu (2005a) proposed an estimation of regression coefficients with unequal probability samples. The study results of Fuller and Wu (2005a) show that the least squares method would obtain a biased estimator with unequal probability samples as the variance is not homogeneity. Hot, Smith and Winter (1980) proposed a weighted least squares method with complex survey under equal probability sampling. DuMouchel and Duncan (1983) proposed a weighted least squares estimator for multiple regression analyses of stratified samples. The weighted least squares estimator could reduce the bias, but enlarge the variance of estimation. Cragg (1989) proposed a Quasi-Aitken weighted least square estimator to reduce the variance of estimation. White (1980) proposed an Eicker-white variance-covariance estimator (E-W VCE) to solve the estimator inconsistency under heteroskedasticity of variance.

This paper aims to compare the estimators of regression coefficients under stratified sampling with unequal probability based upon a Monte Carlo approach and proposed proper estimators for a further empirical study. Four methods of multiple regression analysis proposed by this study, namely, ordinary least squares (OLS), weighted least squares (WLS), probability weighted least squares (PWLS), and Quasi-Aitken probability weighted least squares (Q-A PWLS) are used in this study for comparison analysis to see their performance under data collected with stratified unequal probability sampling. An empirical study is conducted to see how the estimators work in practice.

2. Methodology

In this study, a Monte Carlo simulation experiment is conducted to compare the performance of the estimators of regression coefficients under stratified sampling with unequal probability based upon their biases and variances. The estimators include ordinary least squares, weighted least squares, probability weighted least squares, and Quasi-Aitken probability weighted least squares. The simulation study follow the steps: (1) generating a stratified population, (2) from the generated stratified population repeatedly draw 10,000 stratified unequal probability samples, (3) obtaining the regression coefficients for each sample by four proposed estimators, and (4) comparing the bias and variance of the estimators.

2.1 Population and Unequal probability sampling

In stratified sampling the population of N units is divided into k strata (subpopulation) of N_1 , N_2…, N_k units, respectively. For each stratum, it is assumed that the finite subpopulation of N_h units is a simple random sample of size N_h from an infinite subpopulation. The population value obtained for the j^{th} unit within i^{th} stratum is denoted by (x_{ij}, y_{ij}), $i=1,2,\ldots,k$; $j=1,2,\ldots,N_i$, where x is an independent variable, and y is an dependent variable. The population data are characterized by a regression model of the form

$$\mathbf{Y}_N = \mathbf{X}_N\boldsymbol{\beta} + \boldsymbol{\varepsilon}_N, \text{ where } E(\boldsymbol{\varepsilon}_N \mid \mathbf{X}_N) = \mathbf{0} \text{ for all } \mathbf{X}, \tag{1}$$

where $\mathbf{Y}_N = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_k)^T$ with $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{iN_i})^T$;

$\mathbf{X}_N = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k)^T$ with $\mathbf{x}_i = (\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{iN_i})^T$, $\mathbf{x}_{ij} = (1, x_{ij1}, x_{ij2}, \ldots, x_{ijq})$;

$\boldsymbol{\varepsilon}_N = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N)^T$ with $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \varepsilon_{i2}, \ldots, \varepsilon_{iN_i})^T$; $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_q)^T$.

The sample is drawn by an unequal probability sampling. The j^{th} unit within i^{th} stratum (x_{ij}, y_{ij}) is assigned independently a probability π_{ij} of entering the sample. A sample consisting of K unequal probability samples of $n_1, n_2, \ldots, n_k$ units are sampled from k strata (subpopulation) of N_1 , N_2…, N_k units, respectively. The stratified unequal probability sampling is repeated for 10,000 times to obtain 10,000 samples.

2.2 Estimators of regression coefficients

For each sample, regression coefficients are estimated by the ordinary least squares,

weighted least squares, probability weighted least squares, and Quasi-Aitken probability weighted least squares estimator.

2.2.1 Ordinary least squares estimator

The estimator $\hat{\boldsymbol{\beta}}_{LSE}$ and its variance-covariance matrix $\mathbf{V}(\hat{\boldsymbol{\beta}}_{LSE})$ of the ordinary least squares (OLS) estimator are as follows:

$$\hat{\boldsymbol{\beta}}_{LSE} = (\mathbf{x}_n^T \mathbf{x}_n)^{-1} \mathbf{x}_n^T \mathbf{y}_n, \quad \mathbf{V}(\hat{\boldsymbol{\beta}}_{LSE}) = (\mathbf{x}_n^T \mathbf{x}_n)^{-1} \mathbf{x}_n^T \boldsymbol{\Sigma} \mathbf{x}_n (\mathbf{x}_n^T \mathbf{x}_n)^{-1}, \tag{2}$$

where $\boldsymbol{\Sigma} = diag(\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2)$. For error term with homogeneity, the $\mathbf{V}(\hat{\boldsymbol{\beta}}_{LSE})$ can be consistently estimated by following equation:

$$\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}}_{LSE}) = \hat{\boldsymbol{\sigma}}(\mathbf{x}_n^T \mathbf{x}_n)^{-1}, \text{ where } \hat{\boldsymbol{\sigma}} = \frac{\sum_{i=1}^{n} (y_i - \mathbf{x}_i \hat{\boldsymbol{\beta}}_{LSE})^2}{n - (q+1)}, \tag{3}$$

For error term with heteroscedasticity, the $\mathbf{V}(\hat{\boldsymbol{\beta}}_{LSE})$ can be consistently estimated by Eicker-White variance-covariance matrix $\hat{\mathbf{V}}_{EW}(\hat{\boldsymbol{\beta}}_{LSE})$ as follows:

$$\hat{\mathbf{V}}_{EW}(\hat{\boldsymbol{\beta}}_{LSE}) = (\mathbf{x}_n^T \mathbf{x}_n)^{-1} \mathbf{x}_n^T \hat{\boldsymbol{\Sigma}}_{EW} \mathbf{x}_n (\mathbf{x}_n^T \mathbf{x}_n)^{-1}, \text{ where} \tag{4}$$
$$\hat{\boldsymbol{\Sigma}}_{EW} = diag(\hat{\varepsilon}_1^2, \hat{\varepsilon}_2^2, \ldots, \hat{\varepsilon}_n^2), \ \hat{\varepsilon}_i^2 = (y_i - \mathbf{x}_i \hat{\boldsymbol{\beta}}_{LSE})^2,$$

2.2.2 Weighted least squares estimator

Stratified sampling with sampling fraction P_i and sample weight W_i of ith stratum expressed as follows:

$$P_i = \frac{n_i}{N_i}, \quad W_i = \frac{1}{P_i} = \frac{N_i}{n_i}, \ i = 1, 2, \ldots, k \ , \tag{5}$$

the regression model is expressed as following:

$$\mathbf{W}_{str}^{1/2} \mathbf{y}_n = \mathbf{W}_{str}^{1/2} \mathbf{x}_n \boldsymbol{\beta} + \mathbf{W}_{str}^{1/2} \boldsymbol{\varepsilon}_n, \text{ where} \tag{6}$$
$$\mathbf{W}_{str} = diag(\mathbf{W}_{str\_1}, \mathbf{W}_{str\_2}, \ldots, \mathbf{W}_{str\_k}), \quad \mathbf{W}_{str\_i} = diag(W_i, W_i, \ldots, W_i)_{n_i \times n_i}$$

The weighted least squares estimator $\hat{\boldsymbol{\beta}}_{str}$ and its estimated variance-covariance matrix $\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}}_{str})$ are expressed as

$$\hat{\boldsymbol{\beta}}_{str} = (\mathbf{x}_n{}^{\mathbf{T}} \mathbf{W}_{str} \mathbf{x}_n)^{-1} \mathbf{x}_n{}^{\mathbf{T}} \mathbf{W}_{str} \mathbf{y}_n, \tag{7}$$
$$\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}}_{str}) = (\mathbf{x}_n{}^{\mathbf{T}} \mathbf{W}_{str} \mathbf{x}_n)^{-1} \mathbf{x}_n{}^{\mathbf{T}} \mathbf{W}_{str} \hat{\mathbf{D}}_{e,str} \mathbf{W}_{str} \mathbf{x}_n (\mathbf{x}_n{}^{\mathbf{T}} \mathbf{W}_{str} \mathbf{x}_n)^{-1},$$

where $\hat{\mathbf{D}}_{e,str} = diag(\hat{\mathbf{e}}_{1,str}, \hat{\mathbf{e}}_{2,str}, \ldots, \hat{\mathbf{e}}_{k,str})$, $\hat{\mathbf{e}}_{i,str} = diag(\hat{e}^2_{i1,str}, \hat{e}^2_{i2,str}, \ldots, \hat{e}^2_{in_i,str})$,
$\hat{e}_{ij,str} = y_{ij} - x_{ij}\hat{\boldsymbol{\beta}}_{str}$, $i = 1,2,\ldots,k;\ j = 1,2,\ldots,n_i$.

2.2.3 Probability weighted least squares estimator

Assume that the j^{th} unit within i^{th} stratum (x_{ij}, y_{ij}) is assigned independently a probability π_{ij} of entering the sample. The regression model is expressed as following:

$$\mathbf{W}^{1/2}\mathbf{y}_n = \mathbf{W}^{1/2}\mathbf{x}_n\boldsymbol{\beta} + \mathbf{W}^{1/2}\boldsymbol{\varepsilon}, \tag{8}$$

where $\mathbf{W} = diag(\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_k)$, $\mathbf{w}_i = diag(w_{i1}, w_{i2}, \ldots, w_{in_i})$, $w_{ij} = \dfrac{1}{\pi_{ij}}$,
$i = 1,2,\ldots,k;\ j = 1,2,\ldots,n_i$. The probability weighted least squares estimator $\hat{\boldsymbol{\beta}}_{PW}$ and its estimated variance-covariance matrix $\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}}_{PW})$ are expressed as

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{PW} &= (\mathbf{x}_n{}^{\mathbf{T}}\mathbf{W}\mathbf{x}_n)^{-1}\mathbf{x}_n{}^{\mathbf{T}}\mathbf{W}\mathbf{y}_n, \\ \hat{\mathbf{V}}(\hat{\boldsymbol{\beta}}_{PW}) &= (\mathbf{x}_n{}^{\mathbf{T}}\mathbf{W}_{PW}\mathbf{x}_n)^{-1}\mathbf{x}_n{}^{\mathbf{T}}\mathbf{W}_{PW}\hat{\mathbf{D}}_{e,PW}\mathbf{W}_{PW}\mathbf{x}_n(\mathbf{x}_n{}^{\mathbf{T}}\mathbf{W}_{PW}\mathbf{x}_n)^{-1}, \end{aligned} \tag{9}$$

where $\hat{\mathbf{D}}_{e,PW} = diag(\hat{\mathbf{e}}_{1,PW}, \hat{\mathbf{e}}_{2,PW}, \ldots, \hat{\mathbf{e}}_{k,PW})$, $\hat{\mathbf{e}}_{i,PW} = diag(\hat{e}^2_{i1,PW}, \hat{e}^2_{i2,PW}, \ldots, \hat{e}^2_{in_i,PW})$,
$\hat{e}_{ij,PW} = y_{ij} - x_{ij}\hat{\boldsymbol{\beta}}_{PW}$, $i = 1,2,\ldots,k;\ j = 1,2,\ldots,n_i$.

2.2.4 Quasi-Aitken probability weighted least squares estimator

The Quasi-Aitken probability weighted least squares estimator is proposed by Magee (1998) to reduce the variance of the probability weighted least squares estimator. The regression model is expressed as following:

$$\mathbf{A}^{1/2}\mathbf{W}^{1/2}\mathbf{y}_n = \mathbf{A}^{1/2}\mathbf{W}^{1/2}\mathbf{x}_n\boldsymbol{\beta} + \mathbf{A}^{1/2}\mathbf{W}^{1/2}\boldsymbol{\varepsilon}, \tag{10}$$

where $\mathbf{A} = diag(\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_k)$, $\mathbf{W} = diag(\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_k)$,
$\mathbf{A}_i = diag(exp(z_{i1}\theta), exp(z_{i2}\theta), \ldots, exp(z_{in_1}\theta), \ldots exp(z_{in_k}))$,
$\mathbf{w}_i = diag(w_{i1}, w_{i2}, \ldots, w_{in_i})$, $w_{ij} = \dfrac{1}{p_{ij}}$, $z_{ij} = \dfrac{x_{ij} - \bar{x}_i}{\sqrt{Var(x_i)}}$, $i = 1,2,\cdots,k;\ j = 1,2,\ldots,n_i$.

The Quasi-Aitken probability weighted least squares estimator $\hat{\boldsymbol{\beta}}_{QA}$ and its estimated variance-covariance matrix $\hat{V}(\hat{\boldsymbol{\beta}}_{QA})$ are expressed as

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{QA} &= (\mathbf{x}_n{}^{\mathbf{T}}\mathbf{A}\mathbf{W}\mathbf{x}_n)^{-1}\mathbf{x}_n{}^{\mathbf{T}}\mathbf{A}\mathbf{W}\mathbf{y}_n, \\ \hat{V}(\hat{\boldsymbol{\beta}}_{QA}) &= (\mathbf{x}_n{}^{\mathbf{T}}\mathbf{A}\mathbf{W}_{QA}\mathbf{x}_n)^{-1}\mathbf{x}_n{}^{\mathbf{T}}\mathbf{A}\mathbf{W}_{QA}\hat{\mathbf{D}}_{e,QA}\mathbf{W}_{QA}\mathbf{A}\mathbf{x}_n(\mathbf{x}_n{}^{\mathbf{T}}\mathbf{A}\mathbf{W}_{QA}\mathbf{x}_n)^{-1}, \end{aligned} \tag{11}$$

where $\hat{\mathbf{D}}_{e,QA} = diag(\hat{\mathbf{e}}_{1,QA}, \hat{\mathbf{e}}_{2,QA}, \ldots, \hat{\mathbf{e}}_{k,QA})$, $\hat{\mathbf{e}}_{i,QA} = diag(\hat{e}^2_{i1,QA}, \hat{e}^2_{i2,QA}, \ldots, \hat{e}^2_{in_i,QA})$, $\hat{e}_{ij,QA} = y_{ij} - x_{ij}\hat{\boldsymbol{\beta}}_{QA}$, $i = 1,2,\ldots,k;\ j = 1,2,\ldots,n_i$.

3. Simulation study

3.1 Sample generation

The simulation is conducted by MATLAB program in this study. In the simulation study, a population consisting of three strata with size N_1=400, N_2=500, and N_3=600 was independently generated from subpopulations of Π_1, Π_2, and Π_3, respectively. Let X_1, X_2, and X_3 in three subpopulations are distributed as following:

$$\Pi_1 : \mathbf{X}_1 = \mathbf{U}_1 + \boldsymbol{\Omega}_1;\ \Pi_2 : \mathbf{X}_2 = \mathbf{U}_2 + \boldsymbol{\Omega}_2;\ \Pi_3 : \mathbf{X}_3 = \mathbf{U}_3 + \boldsymbol{\Omega}_3 , \tag{12}$$

where $\mathbf{U}_1 \sim U(\ 70,130)$, $\boldsymbol{\Omega}_1 \sim N(0,\sigma_1^2)$, $\mathbf{U}_2 \sim U(170,230)$, $\boldsymbol{\Omega}_2 \sim N(0,\sigma_2^2)$, $\mathbf{U}_3 \sim U(270,330)$, $\boldsymbol{\Omega}_3 \sim N(0,\sigma_3^2)$. There are two cases are taken for standard deviation σ. Case I: $\sigma_i = 10, \text{for } i = 1,2,3$; Case II: $\sigma_1 = 10, \sigma_2 = 20, \sigma_3 = 30$. *Y* given *X* is generated as

$$Y_{ij} = 50 + 1.2X_{ij} + \varepsilon_{ij} \tag{13}$$

where $\varepsilon_{ij} = \psi_{ij}\nu_1 + \nu_2$, $\nu_1 \sim N(0,25^2)$, $\nu_2 \sim N(0,\delta_i^2)$.

In order to see whether the variability of variance of error inflect the performance of the estimators of regression coefficients, two cases are taken for δ_i: (1) δ_i=5 for *i*=1,2,3, (2) $\delta_1 = 5, \delta_2 = 10, \delta_3 = 15$. Moreover, three cases are taken for ψ_{ij}: (1) $\psi_{ij} = 1$, (2) $\psi_{ij} = \exp(-\frac{X_{ij} - \mu_i}{\sigma_i})$, (3) $\psi_{ij} = \exp(\frac{X_{ij} - \mu_i}{\sigma_i})$ to see how the homogeneity of variance in error term inflect the performance of the estimators of regression coefficients. Two kinds of variances in independent variable X mixed with six kinds of variance in error term results into twelve population settings for simulation study. The distributions of the twelve are displayed in Figure A.1 in Appendix A.

The sample inclusion probabilities for element (x_{ij}, y_{ij}) are generated by

$$\alpha_{ij} = [1 + \exp(\frac{|\gamma_{ij}|}{7.5})], \text{ where } \gamma_{ij} \sim N(0, 10^2) \tag{14}$$

And obtain sample inclusion probability $\pi_{ij} = \frac{n_i}{\sum_{j=1}^{N_i} \alpha_{ij}} \alpha_{ij}, i = 1,2,3, j = 1,2,\ldots,n_i$. In order to compare the effect of sample size, three cases are used for simulation: (1) with small equal size, $n_1=n_2=n_3=5$, (2) with proportional allocation, $n_1=20$, $n_2=25$, and $n_3=30$, (3) with large equal size, $n_1=n_2=n_3=35$. The expectation, variance, and mean squares of error are calculated based upon the simulation results for those four estimators as follows

$$E(\hat{\beta}) \approx \frac{\sum_{i=1}^{10000} \hat{\beta}_i}{10000}, \quad \text{var}(\hat{\beta}) \approx \frac{\sum_{i=1}^{10000} (\hat{\beta}_i - E(\hat{\beta}))^2}{10000}, \quad MSE(\hat{\beta}) \approx \frac{\sum_{i=1}^{10000} (\hat{\beta}_i - \beta)^2}{10000}. \tag{15}$$

3.2 Simulation results

Twelve population regression models carried from the twelve population settings are described in Table 1. The simulation results for different sample sizes are shown in Table 2. As we expect, for the case of homogeneity of variance in both X and in error term among three strata, model 1(i), the bias of the four estimator are all small; while the MSE of OLS and WLS estimators are smaller than that of PWLS, QA-PWLS. For the case of heteroscedasticity in the error term, model 1(ii) and 1(iii), the error term depends on *X*, the estimator of OLS and WLS have larger bias than others, the bias is significant on the case of small sample size; while the MSE for all estimators are all small. For the cases of model 2(i), 2(ii), and 2(iii), the variance of *X* among three strata are different. The biases of OLS and WLS estimators are larger than that of PWLS and QA-PWLS. The bias is significant on small sample size. The MSE of OLS and WLS are smaller than others as the error term is independent of *X*; while the MSE of OLS and WLS are larger than others as the error term depends on *X*.

For the case of model 3(i), 3(ii), and 3(iii), the variances of *X* among three strata are same, but variances of error tem are different. The biases of OLS and WLS estimators are larger than that of PWLS and QA-PWLS. The MSE of OLS is smaller than others for the case of small sample size, but the MSE are similar among the four estimators in larger sample size. That shows that the QA-PWLS can reduce the variance for large sample size. For

the case of model 4(i), 4(ii), and 4(iii), both of the variances of *X* and variance of error term are different among three strata. The biases of OLS and WLS estimators are larger than that of PWLS and QA-PWLS. The MSE of OLS is smaller than others for the case of small sample size, but the MSE are similar among the four estimators in larger sample size. The MSE of OLS and WLS are smaller than others as the error term is independent of *X*; while the MSE of OLS and WLS are larger than others as the error term depends on *X*.

Table 1: Specification of population regression models for simulation

Model No.	Population regression model	Standard deviation of *X*	Standard deviation of ε	
			δ_i	ψ_{ij}
1(i)	$\mu_{Y\|X} = 47.907 + 1.212X$	$\sigma_i = 10, \text{for } i = 1,2,3$	$\delta_i = 5, i = 1,2,3$	$\psi_{ij} = 1$
1(ii)	$\mu_{Y\|X} = 50.831 + 1.196X$			$\psi_{ij} = \exp(-\frac{X_{ij} - \mu_i}{\sigma_i})$
1(iii)	$\mu_{Y\|X} = 50.605 + 1.197X$			$\psi_{ij} = \exp(\frac{X_{ij} - \mu_i}{\sigma_i})$
2(i)	$\mu_{Y\|X} = 49.871 + 1.196X$	$\sigma_1 = 10, \sigma_2 = 20, \sigma_3 = 30$	$\delta_i = 5, i = 1,2,3$	$\psi_{ij} = 1$
2(ii)	$\mu_{Y\|X} = 52.842 + 1.189X$			$\psi_{ij} = \exp(-\frac{X_{ij} - \mu_i}{\sigma_i})$
2(iii)	$\mu_{Y\|X} = 49.273 + 1.202X$			$\psi_{ij} = \exp(\frac{X_{ij} - \mu_i}{\sigma_i})$
3(i)	$\mu_{Y\|X} = 48.492 + 1.207X$	$\sigma_i = 10, \text{for } i = 1,2,3$	$\delta_1 = 5, \delta_2 = 10, \delta_3 = 15$	$\psi_{ij} = 1$
3(ii)	$\mu_{Y\|X} = 48.783 + 1.206X$			$\psi_{ij} = \exp(-\frac{X_{ij} - \mu_i}{\sigma_i})$
3(iii)	$\mu_{Y\|X} = 51.722 + 1.194X$			$\psi_{ij} = \exp(\frac{X_{ij} - \mu_i}{\sigma_i})$
4(i)	$\mu_{Y\|X} = 48.521 + 1.203X$	$\sigma_1 = 10, \sigma_2 = 20, \sigma_3 = 30$	$\delta_1 = 5, \delta_2 = 10, \delta_3 = 15$	$\psi_{ij} = 1$
4(ii)	$\mu_{Y\|X} = 49.384 + 1.201X$			$\psi_{ij} = \exp(-\frac{X_{ij} - \mu_i}{\sigma_i})$
4(iii)	$\mu_{Y\|X} = 50.825 + 1.194X$			$\psi_{ij} = \exp(\frac{X_{ij} - \mu_i}{\sigma_i})$

In summary, the estimators of OLS and WLS are biased under stratified unequal probability sampling as the variances of *X* among three strata are different or the error

term depends on *X*. The MSE of OLS is smaller than others on small sample size. For large sample size, the QA-PWLS can reduce variance and obtain smaller variance than PWLS. The simulation results show that PWLS and QA-PWLS perform better than OLS and WLS in terms of bias under stratified unequal probability sampling; but PWLS has larger variance.

Table 2: Simulation results of $\hat{\beta}_1$

Model No.		1(i)			1(ii)			1(iii)		
(n_1, n_2, n_3)		Case 1	Case 2	Case 3	Case 1	Case 2	Case 3	Case 1	Case 2	Case 3
OLS	bias	-0.002	-0.001	-0.002	-0.012	-0.009	-0.010	0.004	0.003	0.003
	MSE	0.006	0.001	0.001	0.002	0.000	0.000	0.002	0.000	0.000
WLS	bias	-0.001	-0.001	-0.001	-0.010	-0.009	-0.008	0.004	0.003	0.003
	MSE	0.006	0.001	0.001	0.002	0.000	0.000	0.002	0.000	0.000
PWLS	bias	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.000	0.001
	MSE	0.010	0.002	0.001	0.002	0.001	0.000	0.003	0.001	0.000
QA-PWLSE (trace)[1]	bias	0.001	0.002	0.001	-0.001	0.001	0.001	0.002	0.003	0.003
	MSE	0.010	0.002	0.001	0.002	0.000	0.000	0.002	0.000	0.000
QA-PWLSE (det)[2]	bias	0.000	0.001	0.000	-0.001	0.001	0.001	0.002	0.003	0.003
	MSE	0.011	0.002	0.001	0.002	0.000	0.000	0.002	0.000	0.000

Note: Case 1: (n_1, n_2, n_3)=(5,5,5); Case 2: (n_1, n_2, n_3)=(20,25,30); Case 3: (n_1, n_2, n_3)=(35,35,35).

Table 2: Simulation results of $\hat{\beta}_1$ (Continue a)

Model No.		2(i)			2(ii)			2(iii)		
(n_1, n_2, n_3)		Case 1	Case 2	Case 3	Case 1	Case 2	Case 3	Case 1	Case 2	Case 3
OLS	bias	0.011	0.007	0.006	-0.007	-0.006	-0.004	0.018	0.021	0.018
	MSE	0.005	0.001	0.001	0.002	0.001	0.000	0.018	0.006	0.004
WLS	bias	0.010	0.007	0.005	-0.008	-0.006	-0.005	0.019	0.021	0.020
	MSE	0.005	0.001	0.001	0.003	0.001	0.000	0.022	0.006	0.005
PWLS	bias	0.001	0.000	0.000	0.000	0.001	0.001	0.007	0.001	0.001
	MSE	0.009	0.002	0.001	0.004	0.001	0.000	0.016	0.004	0.003
QA-PWLS (trace)[1]	bias	0.003	0.001	0.001	0.005	0.005	0.005	0.001	-0.001	-0.001
	MSE	0.010	0.002	0.001	0.003	0.000	0.000	0.006	0.000	0.000
QA-PWLS (det)[2]	bias	0.002	0.001	0.000	0.004	0.005	0.005	0.002	-0.001	-0.001
	MSE	0.010	0.002	0.001	0.002	0.000	0.000	0.007	0.000	0.000

Table 2: Simulation results of $\hat{\beta}_1$ (Continue b)

Model No.		3(i)			3(ii)			3(iii)		
(n_1, n_2, n_3)		Case 1	Case 2	Case 3	Case 1	Case 2	Case 3	Case 1	Case 2	Case 3
OLS	bias	0.029	0.019	0.016	0.004	0.007	0.005	-0.013	-0.013	-0.009
	MSE	0.013	0.002	0.002	0.003	0.001	0.000	0.005	0.001	0.001
WLS	bias	0.029	0.019	0.015	0.006	0.007	0.007	-0.015	-0.013	-0.011
	MSE	0.014	0.002	0.002	0.003	0.001	0.000	0.005	0.001	0.001
PWLS	bias	0.004	0.000	-0.002	0.002	0.001	0.000	-0.001	0.001	0.001
	MSE	0.019	0.003	0.002	0.005	0.001	0.001	0.006	0.001	0.001
QA-PWLS (trace)[1]	bias	0.006	0.001	-0.001	0.003	0.002	0.001	0.001	0.004	0.004
	MSE	0.020	0.003	0.003	0.005	0.001	0.001	0.005	0.001	0.001
QA-PWLS (det)[2]	bias	0.005	0.000	-0.002	0.003	0.002	0.001	0.001	0.004	0.004
	MSE	0.021	0.003	0.003	0.004	0.001	0.001	0.005	0.001	0.001

Table 2: Simulation results of $\hat{\beta}_1$ (Continue c)

Model No.		4(i)			4(ii)			4(iii)		
(n_1, n_2, n_3)		Case 1	Case 2	Case 3	Case 1	Case 2	Case 3	Case 1	Case 2	Case 3
OLS	bias	0.002	0.002	0.001	-0.005	-0.005	-0.005	0.022	0.027	0.025
	MSE	0.011	0.002	0.001	0.004	0.001	0.001	0.021	0.006	0.004
WLS	bias	0.002	0.002	0.002	-0.005	-0.005	-0.005	0.023	0.027	0.027
	MSE	0.013	0.002	0.002	0.005	0.001	0.001	0.025	0.006	0.005
PWLS	bias	0.001	-0.001	-0.001	-0.001	0.000	0.000	0.007	0.001	0.001
	MSE	0.019	0.003	0.003	0.008	0.001	0.001	0.048	0.014	0.012
QA-PWLS (trace)[1]	bias	0.000	-0.001	-0.001	-0.002	-0.004	-0.004	0.007	0.006	0.005
	MSE	0.020	0.004	0.003	0.007	0.001	0.001	0.017	0.001	0.001
QA-PWLS (det)[2]	bias	0.002	0.000	0.001	-0.003	-0.004	-0.004	0.006	0.007	0.007
	MSE	0.021	0.004	0.003	0.005	0.001	0.001	0.024	0.001	0.001

Monte Carlo approach is used in this paper to compare the efficiency of the four estimators of regression coefficients based upon bias, variance, and MSE. The simulation results show that probability weighted least squares estimator and Quasi-Aitken weighted least square estimator are unbiased estimators of regression coefficients. The simulation results also find that the Quasi-Aitken weighted least square estimator has a smaller asymptotic variance than least squares estimator. Simulation results show that the

ordinary least squares estimator is biased under the data collected under the unequal probability design; while under the equal probability design the weighted least squares estimator is better than ordinary least squares, but under the unequal probability design weighted least squares estimator may have a larger variance.

4. Empirical study

To examine the results carried out by simulation study in previous section. This study uses the real data set of "Taiwan Social Change Survey 2007, Phase 5, Wave 3," collected under a stratified unequal probability sampling by the Institute of Sociology Academia Sinica for empirical comparison of the three methods, namely, OLS, PWLS, and Q-A PWLS *via* comparing the estimates of regression coefficients, RMSE, and R^2.

4.1 Sampling design

The real data set of " Taiwan Social Change Survey" is collected by a complex survey, stratified multi-stage cluster sampling, which includes 1,989 observations. The population is stratified into six strata (region), each region *i* with people T_i. In each region *i*, N_i towns are selected with probability proportional to the town's population C_i. n_i villages are selected with probability proportional to the village's population V_i from each selected town. Then m_i people are selected from each selected village. The probability of the person *j* in the *i*th stratum included in sample π_{ij} and its weight wu_{ij} are shown as follows.

$$\pi_{ij} = (N_i \times \frac{C_i}{T_i}) \cdot (n_i \times \frac{V_i}{C_i}) \cdot (\frac{m_i}{V_i}) = \frac{N_i n_i m_i}{T_i}; \quad wu_{ij} = \frac{1}{\pi_{ij}} = \frac{T_i}{N_i n_i m_i}. \tag{16}$$

In order to increase the precision of estimation, recursive raking with sex, age, and stratum is used in this study to reach the consistency of the distributions of frequency between sample and population. The weight used for raking is wt_i

$$wt_i = \frac{n}{n_i} \times \frac{N_i}{N}, \; i = 1, \cdots, 6. \tag{17}$$

4.2 Variables used for regression analysis

Four variables are used for regression analysis to see the relationship between

respondents' total education years and his (her) parents' total education years. The variables are described as follows.

Dependent variable *Y* (*edu*): total years of education.

Independent variable *X1* (*age*): respondent's age.

Independent variable *X2* (*f-edu*): total years of education of respondent's father.

Independent variable *X3* (*m-edu*): total years of education of respondent's mother.

The sample statistics and the test for equality of mean and equality of variance over six strata are shown in Table 3. The hypothesis test for mean equality ($H_0: \mu_{1p} = \cdots = \mu_{6p}$) shows that all the variables have significant differences among six regions. The Bartlett's test for homogeneity of variance ($H_0: \sigma_{1p} = \cdots = \sigma_{6p}$) all shows that all the variables have heterogeneity of variance among six regions.

Table 3: Sample statistics of the variables

	Variables		Stratum (region)						P-value
			Core cities	General cities	New cities	Traditional counties	Rural counties	Senior counties	
X	*X1* (*age*)	mean	43.46	41.40	42.82	47.85	48.98	44.58	p<0.0001
		s.d.	17.52	15.38	16.37	18.57	16.86	18.30	p=0.0093
	X2 (*f_edu*)	mean	8.27	7.03	6.41	4.94	5.03	4.89	p<0.0001
		s.d.	5.05	5.05	4.59	4.41	4.36	4.57	p=0.0202
	X3 (*m_edu*)	mean	6.32	5.36	4.55	3.15	3.48	3.05	p<0.0001
		s.d.	4.93	4.75	4.404	3.86	3.98	4.13	p<0.0001
Y	*edu*	mean	11.42	8.93	8.81	12.27	9.11	10.71	p<0.0001
		s.d.	4.35	5.19	4.92	4.35	5.07	4.64	p=0.0061

Note: p-values in the last column are from AVOVA test for mean equality and Bartlett's test for homogeneity of variance, respectively.

4.3 Regression analysis

Three estimators, OLS, PWLS, and Q-A WPLS, are used to estimate the regression coefficients, in which the OLS estimator is taken from equation (2) and its variance estimator is from equation (4), PWLS estimator and its variance estimator is taken from equation (9) and QA-PWLS estimator and its variance estimator is taken from equation

(11). The weight w_{ij} for PWLS and QA-PWLS is calculated as following

$$w_{ij} = wt_i \times wu_{ij} = wt_i \times \frac{1}{\pi_{ij}}, i = 1, 2, ..., 6;\ j = 1, 2, ..., n_i . \quad (18)$$

The estimated coefficients are shown in Table 4. The empirical results consist with previous studies. The results show that there is no big difference among the estimated parameters of those three methods. The results also show that the education years of respondents have significant negative relationship with their ages but have positive relationship with their parents' education years.

Table 4: Estimated regression models

		β_0	β_1	β_2	β_3	R-Square	RMSE
OLS	Estimate	13.4186	0.3266	0.0937	-0.1191	0.5570	3.1650
	St. Error	(0.1071)	(0.0000)	(0.0001)	(0.0000)		
WPLS	Estimate	13.5639	0.3126	0.1020	-0.1226	0.5567	3.1662
	St. Error	(0.1112)	(0.0000)	(0.0001)	(0.0000)		
Q-A	Estimate	12.8397	0.3080	0.0981	-0.1037	0.5528	3.1799
WPLS	St. Error	(0.1050)	(0.0000)	(0.0001)	(0.0000)		

5. Conclusion

The sampling design is getting more complex to comply with a variety of social environment and to increase the precision of sampling survey as well. The traditional estimators used with complex survey may lower the accuracy of the statistical analysis. This study explores the methods of regression analysis on survey data obtained under a complex sampling. Four methods of multiple regression analysis proposed by this study, namely, ordinary least squares, weighted least squares, probability weighted least squares and Quasi-Aitken probability weighted least squares are used in this study for comparison analysis. Monte Carlo approach is used in this paper to compare the efficiency of the four estimators of regression coefficients based upon bias, variance, and MSE. The simulation results show that probability weighted least squares estimator and Quasi-Aitken probability weighted least squares estimator perform better than ordinary least squares estimator and weighted least squares estimations in terms of bias, but probability weighted least squares estimator has a larger variance for estimating regression

coefficients under complex survey. Quasi-Aitken probability weighted least squares estimator performs better than other estimator in terms of bias and MSE as the error term and independent variables have heterogeneity of variance among strata. The simulation results also find that the Quasi-Aitken weighted least square estimator has a smaller asymptotic variance than least squares estimator on the cases of larger sample size.

This study uses the data of "Taiwan Social Change Survey 2007, Phase 5, Wave 3," collected under a stratified unequal probability sampling by the Institute of Sociology Academia Sinica for empirical comparison of those three methods *via* comparing the estimates of regression coefficients, RMSE, and R^2. The empirical results consist with previous studies and the simulation results in this study. The results show that there is no big difference among the estimated parameters of those three methods. The results also show that the education year of respondents has significant negative relationship with their age but has positive relationship with their parents' education year.

References

Ajmani, V. B. 2009. *Applied Econometrics Using the SAS® System.*

Cochran, W. G. 1977. *Sampling Techniques,* 3^{rd} edition.

Cragg, J. G. 1992. Quasi-Aitken estimation of heteroskedasticity of unknown form.*J.econometr*.,**54**,179-201.

DuMouchel, W. H. and G. J. Duncan. 1983. Using sample survey weights in multiple regression analyses of stratified samples. *Journal of the American Statistical Association*, **78**, 383, 535-543.

Graubard, B. I. and E. L. Korn. 2002. Inference for superpopulation parameters using sample surveys. *Statistical Science*.**17**, 1, 73-96.

Hansen, M. H. and W. N. Hurwitz. 1943. On the theory of sampling from finite populations. *The Annals of Mathematical Statistics*, **14**, 333-362.

Holt, D., T. M. F. Smith, and P. D.Winter. 1980. Regression analysis of data from complex surveys. *Journal of the Royal Statistical Society*, **143**, 4, 474-487.

Kutner, M. H., C. J. Nachtsheim, J. Neter, and W. Li. 2005. *Applied Linear Statistical Models,* 5^{th} edition.

Magee, L. 1998. Improving Survey-Weighted Least Squares Regression. *Journal of Royal Statistical Society*, **60**, 1, 115-126.

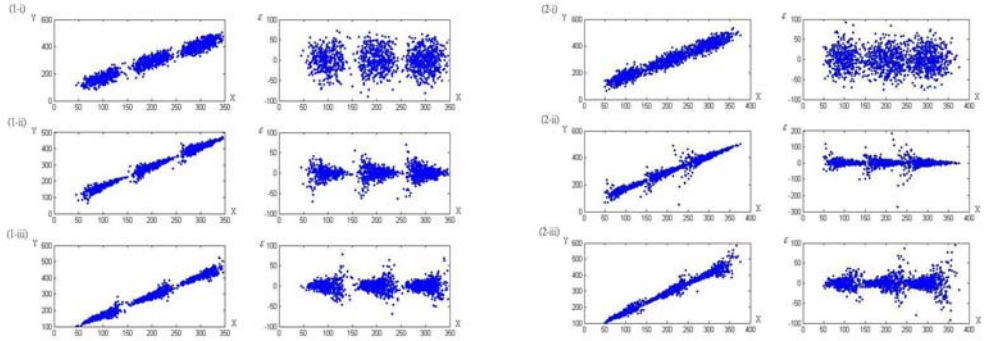
Rutemiller, H.C., and D.A. Bowers. 1968. Estimation in a Heteroscedastic Regression Model. *Journal of the American Statistical Association*, **63**, 322, 552-557.

White, H. 1980. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, **48**, 817-838.

Wu, Y.Y. and A. Fuller. 2005. Preliminary Testing Procedures for regression with survey samples. *In Proceedings of the Survey Research Method Section, American Statistical Association*, 3683-3888.

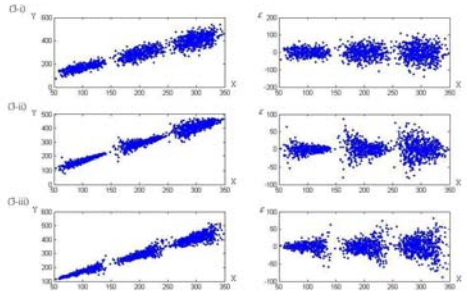
Wu, Y.Y. and A. Fuller. 2005a. Estimation of regression coefficients with unequal probability samples. *In Proceedings of the Survey Research Method Section, American Statistical Association*, 3892-3899.

Appendix A

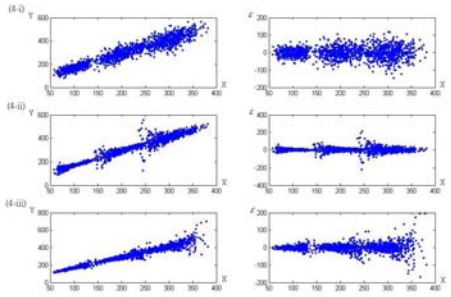


Population 1(i)~(iii)

Population 2(i)~(iii)



Population 3(i)~(iii)



Population 4(i)~(iii)

Figure A.1: Scatter diagram of populations