Scientific Research Publishing

Open Journal of Statistics, 2016, 6, 741-748
http://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Sample Selection Model with Bootstrap (BPSSM) Approach: Case Study of the Malaysian Population and Family Survey

Muhamad Safiih Lola, Wan Saliha Wan Alwi, Nurul Hila Zainuddin

School of Informatics and Applied Mathematics, Universiti Malaysia Terengganu, Kuala Terengganu, Terengganu, Malaysia
Email: safiihmd@umt.edu.my, salihaalwi@gmail.com, hila.zainuddin@gmail.com

How to cite this paper: Lola, M.S., Alwi, W.S.W. and Zainuddin, N.H. (2016) Sample Selection Model with Bootstrap (BPSSM) Approach: Case Study of the Malaysian Population and Family Survey. *Open Journal of Statistics*, **6**, 741-748.
http://dx.doi.org/10.4236/ojs.2016.65060

Received: June 30, 2016
Accepted: September 19, 2016
Published: September 22, 2016

Abstract

Heckman Sampel Selection Model (PSSM) has been adopted widely in the study of labour work. This model contains exogenous, endogenous and standard error variables. However, this model is constantly exposed to high inaccuracy of estimation result. Therefore, to obtain an accurate and precise estimation, the bootstrap approach is introduced. The bootstrap approach will be hybrid with PSSM model known as BPSSM to achieve estimation result that is more precise. Then, the BPSSM is applied to Malaysian Population and Family Survey 1994 (MPFS-1994) data. The results showed that BPSSM provide a smaller standard error and shorter confidence intervals.

Keywords

Sampel Selection, Bootstrap, Standard Error, Confidence Intervals

1. Introduction

Sample selection model is part of the field of econometrics. The term "selection" or "select" is the term commonly used and it is mentioned in a number of different issues related with the econometric data. Sample selection was developed [1]. This model also has good interaction of relation and ideal to expert evaluation and quantitative information.

The earliest introduced model consists of female labour force and wage equality [2], [3]. They have used this model widely in the model affects unions, occupational choices, schooling choices, options and also a residential area of the industry. According to [4] in the female labour supply model, the binary model choice whether a person's functioning and conditions of work can be seen in working hours. Traditionally,

this model has been estimated by maximum likelihood or using computerized methods called Heckman two-step method.

Model selection consists of two parts [5]. The first part is the selection equation (known as the participation or equality of results). In this section, not random sample taken into account for the observation of the process of entering the sample and determine a probability sample of the population. This combination can fix the not random sample and the estimated relation of the population. The second part is the structural equation. This is also known as equality of outcome or equality of salaries with relationship desired study population centres.

Therefore, to get the accurate and precise estimation, bootstrap approach was introduced. Through this method, Bootstrap approach will be hybridized with PSSM model called Bootstrap PSSM to get more accurate estimator result. The best model is the model contains of consistent and efficient.

Bootstrap introduced by [6]. Bootstrap is a common technique that builds confidence interval by resampling with replacement sample from finite. The main objective of this article, we propose to develop a sample selection model (PSSM) based on bootstrap approach to achieve estimation result that is more precise. After that, the model of BPSSM was applied for the Family and Population Data 1994.

2. Methods

Bootstrapping the Base Model of BPSSM

In this study, a mean process is considered to monitor individual observations of $X_1, X_2, X_3, \cdots, X_n$ with assumption of dependent and uncorrelated distributed. Thus, the base model and residual of BPSSM can be given by:

$$y_i^* = \beta_0 + x_i'\beta_1 + \varepsilon_i$$

$$d_i = \begin{cases} 1, & d_i = z_i'\gamma + v_i \\ 0, & \text{others} \end{cases}$$

$$y_i = y_i^* d_i, \quad i = 1, 2, \cdots, N \tag{1}$$

where

y_i^* and d_i = dependent variables

x and z = vectors of exogenous remaining variables

β_0, β_1 and γ = unknown parameter vectors

ε_i and v_i = zero mean error terms

The standard approach is to assume that follow a bivarite normal distribution and then applied to the maximum likelihood estimation or a two-stage estimation procedure purposed by Heckman (1979). Firstly, how to estimate β and γ consistently from the data. In general, both the error terms are correlated, since that the regression of y on x for the selected sample will not give consistent estimates of γ. It is well known that the consistency of those estimators depends on the assumption of bivariate normality. For a random sample from the population it is observed that d_i x and z. If and only if, observation of $d_i = 1$ then, we observed y_i. This sample selection models

in (1) consist of two equations or parts; the first structural part, embodying the desired population relationship or is the equation of primary interest and second, the selection part or is the reduced form takes account of the non-representative nature of the present non-random sample. From the literature (Martin, 2001), the identification purpose, the variable z contains at least one variable which does not appear the relation between an outcome in interest y and a vector of covariates x and the selection equation describing the relation between a binary participation decision d_i and another vector of covariates z.

In this study, the hybridization of bootstrap approach in base model (1). This hybridization produced a hybrid control charts where the basic model named by Bootstrap PSSM (BPSSM). The algorithm for this hybrid process is as follows:

Step 1: A sample data, $z_t = z_1, \cdots, z_{t-1}$ generated by time $i = 1, \cdots, m$ from a dependent and colerated distribution. The data then will be used to estimate parameter of γ, base model of PSSM $d_i^* = z_i'\gamma + v_i$

Step 2: Find bootstrap replication, $B(c)$ by using:

$$BiasB = \frac{\sum_{i=1}^{N}\left[E\left(e^{B(c)}\right) - e_i^{B(c)}\right]}{N}$$

where $E\left(e^{B(c)}\right)$ and $e_i^{B(c)}$ defines as average of expected of error and true value of error respectively. This estimation will be repeated several times to get a constant value of bias and this new value will be used to analyze data in Step 1.

Step 3: By continue from Step 2, the residual value will be used in sampling with replacement method to get a matrix of residual bootstrap, $e_i^{B(c)}$

$$v_i^{B(t)} = \begin{pmatrix} v_1^{B(1)} & \cdots & v_1^{B(c)} \\ \vdots & & \vdots \\ v_m^{B(1)} & \cdots & v_m^{B(c)} \end{pmatrix}$$

Step 4: For each residual in Step 3, compute new data $d_i^{*B(c)}$

Step 5: Compute average of column of $d_i^{*B(c)}$ by using $z_i'^{B(c)} = \frac{d_i^* - v_i^{B(c)}}{\gamma}$, with c defines as summation of bootstrap replication, for example $c = 1000$

Step 6: By using bootstrap data, $d_i^{*B(c)}$ compute parameter of γ

Step 7: For complete PSSM model, $y_i^* = \beta_0 + x_i'\beta_1 + \varepsilon_i$ must be hybrid with bootstrap approach. Step 1 until Step 6 rapidly to compute bootstrap data of $x_i'^B$.

Step 8: After the both equation have bootstrap data, Heckman Two Step estimation was used base on model BPSSM;

$$\begin{aligned} &y_i^{*B} = \beta_0 + x_i'^B\beta_1 + \varepsilon_i^B \\ &d_i^B = 1 \quad \text{if } d_i^B = z_i'^B\gamma + v_i^B \\ &d_i^B = 0 \quad \text{others} \\ &y_i^B = y_i^{*B}d_i^B, \end{aligned} \tag{2}$$

In this study, we intend to examine the performance of BPSSM in terms of effective-

ness and efficiency control. Numerical estimation was selected to be used in this study. For numerical estimation, basically it is used to examine effectiveness of base model where is evident in two kind of methods, *i.e.* confidence interval, bootstrap percentile (PB) and Biased Corrected and Accelerate (BCa). BP and BCa selection is motivated by the advantages of these two methods in which BP is the basic method for estimating bootstrap intervals while BCa is a method that can improve BP interval estimation [7]. The main reason for selecting different methods is to finds the differences in the effectiveness of the hybrid model when using those interval methods. In theory, a model that gives the shortest interval estimation is said to be more effective model. This is because; short interval giving the idea that model estimation is closer to real interval estimation. Therefore, the lower and upper limit $\left[\hat{\theta}_{\mathbf{B}}, \hat{\theta}_{A}\right]$ can be given by:

Student's-*t*:

$$\left[\hat{\theta}_{\mathbf{B}}, \hat{\theta}_{A}\right] = \left[\hat{\theta} - t_{n-1}^{\alpha} \cdot \hat{RP}, \hat{\theta} + t_{n-1}^{\alpha} \cdot \hat{RP}\right] \tag{3}$$

Bootstrap Persentile (BP):

$$\left[\hat{\theta}_{B}, \hat{\theta}_{A}\right] = \left[\hat{\theta}^{(\alpha B)}, \hat{\theta}^{(1-\alpha)B}\right] \tag{4}$$

Bias Corrected and accelerated (Bca):

$$\left[\hat{\theta}_{\mathbf{B}}, \hat{\theta}_{A}\right] = \left[\hat{\theta}^{\alpha_1}, \hat{\theta}^{\alpha_2}\right] \tag{5}$$

where t_{n-1}^{α} in Equation (3) represent value of percentage α for student's-t distribution with $n-1$ degree of freedom.

In this study, α is valued as $\alpha = 0.05$ and standard error estimation, $\hat{RP}$ can be calculated using the discussion on [7].

For Equation (4), length of this interval based on percentile of mean estimation of bootstrap replication, B of 100α-th where $\alpha = 0.05$ and $1-\alpha = 0.95$ [7]. In other word, upper and lower limit of BP refers to the interval length from 50-th through 950-th replication. While α_1 and α_2 in Equation (5) refers to normal confidence interval with $\alpha_1 = \alpha$ and $\alpha_2 = 1-\alpha$ respectively.

Mean Square Error (MSE) and Root Mean Square Error (RMSE) was used for error estimation in this study. In theory, a model that gives the smallest estimation value is said to be more efficient and automatically show the effectiveness of the model itself [8]. By taking the idea of small error in base model, it clearly shows that model is giving more accurate estimation. Thus, error estimation used in this study is given:

$$\text{MSE} = \frac{\sum_{i=1}^{N}\left[e_i^B - E\left(e^B\right)\right]^2}{N} \tag{6}$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}\left[e_i^B - E\left(e^B\right)\right]^2}{N}} \tag{7}$$

where for both MSE and RMSE refers to differences of real error, e_i^B with expected of error estimation, $E\left(e^B\right)$.

3. Results and Discussion

Application of Hybrid Model: The Malaysian Population and Family Survey 1994 Data

A comparison of performance of the real model, PSSM and hybrid model, BPSSM in terms of effectiveness or efficiency base on model estimation. The data set used for this study is from the Malaysian population and family survey 1994 (MPFS-1994). This survey was conducted by National Population and Family Development Board of Malaysia under Ministry of Women, Family and Community Development Malaysia. This survey was specifically for married women, providing relevant and significant information for the problem of married women status regarding wages, educational attainment, household composition and other socioeconomic characteristics. The original MPFS-94 sample data comprises 4444 married women.

The whole data sets used in this study consisted of 2792 married women. The selection rules [9] were applied to create the sample criteria of choosing for participant and non-participant married women on the basis of the MPFS-94 data set, which are as follows:

- Married and aged below 60
- Not in school or retired
- Husband present in 1994
- Husband reported positive earnings for 1994

The empirical results of the basic specification one are presented for the Heckman two-step approach. These approaches consider the probit estimates for the participation equation as a first step and OLS estimates for the wage equation as the second step. We discuss both the participation and wage equation on the estimated coefficient for interval method, the significant effect, and consistency and for PSSM, as well as BPSSM for comparison purposes.

Based on **Table 1**, the estimate for the standard intervals (Student's-t confidence interval) using hybrid model gives a shorter interval. See for example, the length of BPSSM-t valued 0.0512791 compare to real model, PSSM-t which is more length 1.0898210. This result can be seen clearly in **Figure 1**. The differences shown in these two models proved that the bootstrap approach fixed the interval estimation and gives a good performance for hybrid model.

In **Figure 1** and **Figure 2**, a plot of interval estimation for Student's-t method in **Figure 1** and also a plot of Bootstrap Percentile (BP) and Bias Corrected and accelerated (BCa) showed in **Figure 2**. Based on these two plots, we found that bootstrap interval method, *i.e.* BP and BCa, gave short length compare to the standard interval which show a length interval either for real or hybrid model. Significant difference on the length value shows that the bootstrap percentile interval method gave a good performance compare to Student's-t estimation. By considering this result, we can say that BPSSM model gives a better performance when using bootstrap interval method. With this result, it's clearly proved that hybrid model provides an effective estimation compare to the real model.

Table 1. Interval estimation for the real and hybrid model.

Model	Lower	Upper	Length
PSSM-t	22.8134200	23.9032400	1.0898210
BPSSM-t	23.2661200	23.3284100	0.0622905
BPSSM-PB	2.1398290	2.1136020	0.0262270
BPSSM-BCa	23.3015500	23.2929800	0.0085664

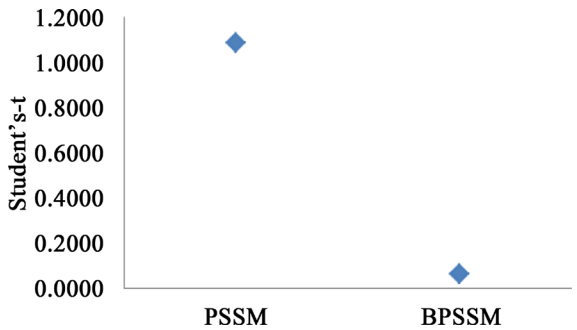

Figure 1. Interval estimation of standard interval: Student's-t.

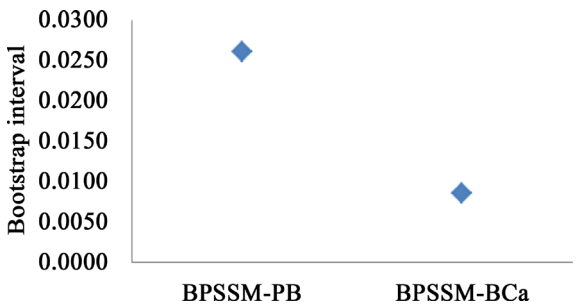

Figure 2. Interval estimation of bootstrap interval method.

Table 2. Error value for real and hybrid model.

Model	MSE	RMSE
PSSM	0.000574969	0.023978510
BPSSM	0.000394221	0.019855010

Next, estimation of the effectiveness of the real and hybrid models is seen in the results of the MSE and RMSE and the good performance of the model is based on the theory of effectiveness estimation model, as discussed in the previous section. Therefore, the results of this error can be referred to Table 2.

Based on the results of Table 2, a model with the bootstrapping approach provides the smallest error value compared to the real model, PSSM. For example, the estimation result for MSE of hybrid model is 0.019855010 while real model gives large error estimation, *i.e.* 0.023978510. These results are plotted in Figure 3 and Figure 4 for an illustration comparing the results for both models of MSE and RMSE estimation value.

Based on such a significant error reduction in Figure 3 and Figure 4, show the boot-

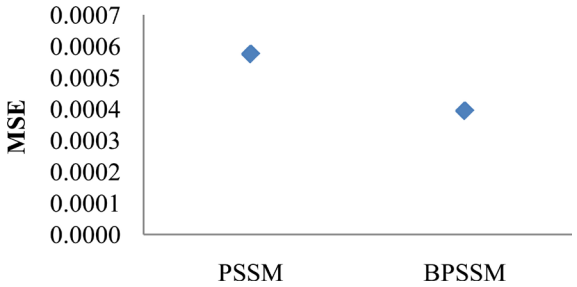

Figure 3. Plot of error estimation, MSE.

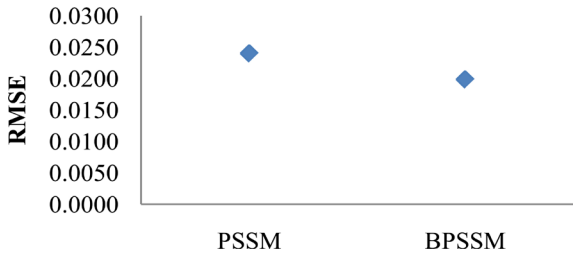

Figure 4. Plot of error estimation, RMSE.

strap approach in real base model of control charts fixing the estimation of real model and provide a more accurate estimation for the model. This small error values also indicate that the hybrid model is more effective and gives good performance compared to the real model, PSSM.

4. Conclusion

In this study, a PSSM model was hybrid with bootstrap method using an alternative algorithm. Using an alternative algorithm, the hybrid process was involved the construction of a standard error of PSSM confidence interval and proposed a new hybrid model of BPSSM. The data set Malaysian population and Family survey 1994 (MPFS-1994) was used. Participation and wage equation on the estimated coefficient for interval method, the significant effect, and consistency and for PSSM, as well as BPSSM for comparison purposes was discussed. The differences shown in these two models proved that the bootstrap approach fixed the interval estimation and gives a good performance for hybrid model. Estimation of the effectiveness of the real and hybrid models is seen of the MSE and RMSE. This small error value also indicates that the hybrid model is more effective.

Acknowledgements

A special gratitude for School of Informatics and Applied Mathematic (SIAM) and Research Management Centre (RMC), Universiti Malaysia Terengganu for supported this research paper.

References

[1] Heckman, J.J. (1974) Shadow Wages, Market Wages and Labor Supply. *Econometrica*, **42**, 679-693. http://dx.doi.org/10.2307/1913937

[2] Gronau, R. (1974) The Effect of Children on the Housewife's Value of Time. Economics of the Family: Marriage, Children, and Human Capital, University of Chicago Press, 457-490.

[3] Heckman, J.J. (1976) The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models. *Annals of Economic and Social Measurement*, **5**, 475-492.

[4] Zhou, C. (2010) The Extent of the Maximum Likelihood Estimator for the Extreme Value Index. *Journal of Multivariate Analysis*, **101**, 971-983. http://dx.doi.org/10.1016/j.jmva.2009.09.013

[5] Markus, F. (1998) Semi-Parametric Estimation of Selectivity Models. Unbublished Ph.D. Thesis, Konstanz University, Konstanz.

[6] Efron, B. (1979) Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, **7**, 1-26. http://dx.doi.org/10.1214/aos/1176344552

[7] Efron, B. (2003) Second Thoughts on the Bootstrap. *Statistical Science*, **18**, 135-140. http://dx.doi.org/10.1214/ss/1063994968

[8] Chou, S.-R., Chien, A., Changchien, C.-C. and Wu, C.-H. (2010) A Comparison of the Forecasting Volatility Performance on EWMA Family Models. *International Research Journal of Finance and Economics*, **54**, 19-28.

[9] Martins, F.M. (2001) Parametric and Semi-Parametric Estimation of Sample Selection Models: An Empirical Application to the Female Labour Force in Portugal. *Journal of Applied Econometrics*, **16**, 23-39. http://dx.doi.org/10.1002/jae.572