BAYESIAN APPROACH FOR SAMPLE SELECTION BIAS CORRECTION IN REGRESSION

By

Labeed Mokatrin

Submitted to the

Faculty of the College of Arts and Sciences

of American University

in Partial Fulfillment of

the Requirements for the Degree

of Doctor of Philosophy

In

Statistics

Chair:

Jun Lu, Ph.D.

Mary Gray

Mary Gray, Ph.D.

Monica C. Jackson

Monica Jackson, Ph.D.

Dean of the College of Arts and Sciences

August 17, 2011

Date

2011
American University
Washington, D.C. 20016

BAYESIAN APPROACH FOR SELECTION BIAS CORRECTION IN REGRESSION

By

Labeed Mokatrin

Submitted to the

Faculty of the College of Arts and Sciences

of American University

in Partial Fulfillment of

the Requirements for the Degree

of Doctor of Philosophy

In

Statistics

Chair:

____
Jun Lu, Ph.D.

____
Mary Gray, Ph.D.

____
Monica Jackson, Ph.D.

____
Dean of the College of Arts and Sciences

____
Date

2011
American University
Washington, D.C. 20016

BAYESIAN APPROACH FOR SAMPLE SELECTION BIAS CORRECTION IN REGRESSION

BY

Labeed Mokatrin

ABSTRACT

Selection bias occurs when samples are self-selected rather than randomly selected from the target population. This is a well-known problem and has been extensively studied in research studies in statistics and economics. In this work, I adopt a Bayesian approach to correct sample selection bias under the self-selection setup proposed in Heckman model. Bayesian methods treat the population parameters of interest as random variables instead of unknown constants. The distributions of these random parameters are called prior distributions. Statistical inference is based on the posterior distribution, which combines information from the data and the prior. Markov Chain Monte Carlo (MCMC) methods are used for Bayesian computation of the posterior distributions. The results from the proposed Bayesian method are compared to that of Heckman's two-step estimation via various simulation studies. A comprehensive simulation study is conducted where various scenarios are considered for the simulation setup and design. Furthermore, in addition to the most common self-selection setup, the new approach is extended to handle self-selection with Binary outcome model.

ACKNOWLEDGMENTS

First, I would like to thank my dissertation advisor Dr. Jun Lu. His advice, insightful criticisms, and patient encouragement aided the writing of this dissertation in innumerable ways. I would also like to thank the committee members, Dr. Mary Gray and Dr. Monica Jackson. This work would not have been possible without their valued advice and suggestions. Dr. Gray's steadfast support was greatly needed and deeply appreciated. I would like to thank the faculty and staff members of the department of Mathematics and Statistics at AU, in particular, Dr. Behzad Jalalai and Linda Greene for all their help. I am forever grateful to my family, Shireen, Adam, and Rany, for their continued support and encouragement. The numerous sacrifices they have made over the last few years allowed me to reach important milestones in my professional career. Finally, I would like to extend my gratitude to my mom and dad for all their support and enthusiasm for my education throughout my life.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF ILLUSTRATIONS

Figure

CHAPTER 1

INTRODUCTION

The Problem of Sample Selection Bias

In some sociology and economics studies, samples are self-selected rather than randomly selected from the target population. Bias can occur when using self-selected samples, because the selection criteria are often correlated with the variables of interest. Such bias is often called the Sample Selection Bias.

For example, Manski & Wise (1983) studied the relationship between SAT scores and potential college achievement. The researchers could only sample from students who were already admitted to college, but not from all students who could potentially go to college. In this case, students who scored well in the SAT were more likely to attend college, and hence are more likely to be selected into the sample. Another example is the study of women's education background and their earnings (Heckman 1979). Samples were selected from women with labor force participation. However, individuals only join the labor force if their potential earnings or occupational status meet some criteria. As a result, sampling from women in labor force ignored the women who had low potential earnings.

Heckman's Two-Step Method

Heckman (1976) raised the issue of sample selection bias when a dependent variable in the regression has values that are missing not at random. He proposed to estimate the full information maximum likelihood (FIML) by way of a two-step method. This method is called a Limited Information Maximum Likelihood (LIML).

FIML is a well-known econometric technique for estimating equation models in which the parameters of all equations are estimated simultaneously, with all the information in the model (Maddala 1977). Similar to FIML, LIML is a maximum likelihood basis for estimating one structural equation, or a proper subset of structural equations from a system of equations (Anderson and Rubin in 1949).

Heckman (1976) discussed the common structure of statistical models of limited dependent variable as well as a simple estimator for this model. He presented a unified summary of statistical model selection and limited dependent variables. Heckman (1979) proposed a solution to sample selection bias using the two-step estimator method. According to this method, in the first step, we use probit regression to model the sample selection process. A new variable called the *Inverse Mills Ratio* is calculated based on the probit regression results. In the second step, we add the Inverse Mills ratio to the regression analysis as an independent variable and simply use Ordinary Least Squares (OLS) to estimate the regression coefficients. Heckman's two-step estimation procedure is easy to implement. It has been well recognized in the applied fields, such as economics and sociology, as a correction for sample selection bias. Examples of the application of Heckman's two-step estimation can be found in, for example, (Mroz 1987), (Nawata 1994), and (Leung and Yu 1996).

Limitations of Heckman's Method and Known Alternatives

Discussions of Heckman's two-step estimation and other approaches to sample selection bias present themselves readily in the literature in the last two decades. Winship and Mare (1992) discussed the difficulties and limitation of several sample selection bias correction techniques. They show how self-selection leads to biased estimates in regression, review models that have been proposed, discuss Heckman's estimator and its limitations, and discuss other approaches to selection such as nonparametric approaches to estimating selection models. They suggest that, when selection is an issue, researchers should present estimates using a variety of methods, because the results may depend on the method used.

Nawata (1993) analyzes methods for estimating models with selection bias by comparing Maximum Likelihood Estimation (MLE) and Heckman's two-step estimator with Monte Carlo experiments. The results show that Heckman's two-step estimator can perform well when there is no multicollinearity between the Inverse Mills Ratio and the explanatory variables. However, it will perform relatively poorly when multicollineartiy exists and MLE becomes more efficient.

Stolzenberg and Relles (1997) provide mathematical tools to assist intuition about selection bias in concrete empirical analysis. They indicate that there is no general solution to the selection bias problem, but they present a new decomposition of selection bias. In this decomposition, the analyst should be able to develop intuition and make reasonable judgments about the source, severity, and direction of sample selection bias in a particular analysis. The authors also list several bias correction procedures that are available. They suggest that the safest approach to sample selection bias problem is first to understand how nonrandom selection occurs in the data. If the data seem to be selected as described by Heckman, then it is appropriate to use Heckman's two-step model.

Pahni (2000) discusses Monte Carlo studies of Heckman’s correction and illustrates a critique of Heckman's estimator. He indicates that the explanatory variables in Heckman’s two-step model may have a large set of variables in common which causes collinearity with the Inverse Mills Ratio. Pahni concludes that we should diagnose collinearity problems before deciding which estimator to use. If there is no collinearity between the regressors and the Inverse Mills Ratio, the Heckman two-part model is the most robust approach. On the other hand, if collinearity problems exist, the MLE approach is preferable to Heckman’s two-step method.

Bayesian Approach

In this study, we propose a Bayesian approach to correct sample selection bias under the self-selection setup proposed in Heckman (1979). Bayesian methods treat the population parameters of interest as random variables instead of unknown constants. The distributions of these random parameters are called prior distributions. Often both expert knowledge and mathematical convenience play a role in selecting a particular type of prior distribution. Statistical inference is based on the posterior distribution, which combines information from the data and the prior. We use Markov Chain Monte Carlo (MCMC) methods for Bayesian computation of the posterior distributions in this study. We also compare the performance of the proposed Bayesian method and that of Heckman’s two-step estimation via simulation study.

In the next chapter, we give a more detailed introduction to the sample selection problem and Heckman's two-step estimation. Using the same assumption on the data collection as used by Heckman, a Bayesian model to correct the selection bias is introduced in chapter three. A simulation study is presented in chapter four, where we demonstrate the proposed Bayesian method and compare the estimates from various approaches using the Women Wage data example and also by using a comprehensive simulation study where various scenarios will be considered for the simulation setup and design. In chapter five, we will apply the Bayesian model to a real-world data example by using AU students' placement exam data. In chapter six, we will extend the proposed Bayesian method to the Generalized Linear Model (GLM).

CHAPTER 2

SELF-SELECTED SAMPLING MODEL AND BIAS

Self-Selected Sampling Model

A famous example of sample selection bias is the estimation of the wage equation (Pahni 2000). When trying to estimate the results of schooling on the wage rate, the researcher faces the problem that some individuals who have received schooling do not work. These individuals have not received an offer that meets their reservation wage. If we assume a positive relationship between schooling and wages, people with little schooling will on average have a lower offered wage and therefore a lower employment rate than those with more years of schooling. But we only observe the wage offers which exceed an individual's reservation wage. As a consequence, we only observe the wages of those people with few years of schooling that receive comparatively high wage offers. In this case, there is self-selected sampling, and the OLS estimate is biased.

In this example, simple OLS regression of wages on years of schooling will lead to bias estimates, because the sample (working people) is unrepresentative of the population one is interested in (all people who have received schooling). The selection problem can be viewed as a problem of missing observations, except that they are not missing at random.

A linear regression model with self-selection samples can be presented using the following two-equation model:

$$Y_i = X_{1i}\beta_1 + \varepsilon_i \tag{1}$$

$$Z_i = X_{2i}\beta_2 + u_i \tag{2}$$

We call Equation 1 the observation equation and Equation 2 the selection equation. In the previous example, individuals who are only able to achieve low wage rate given their level of schooling will decide not to work. Therefore, the probability that their offered wage is below their reservation wage is highest. In other words, ε_i and u_i can be expected to be positively correlated which causes sample selections bias.

When observations are missing at random, Equation 1 can still be estimated by OLS. Typically there are three causes of non-randomly missing observations: *censoring*, *truncation* or *self-selected sampling*. A sample is *censored* when observations on Y_i are not available in some range and are reported at a cutoff value, but, the explanatory variables X_{1i} are all available. When observation on the X_{1i} are also unavailable, the sample is said to be *truncated*. When *self-selected sampling* occurs, observations on Y_i are recorded only if another variable Z_i takes on a value above or below some cutoff value. In this article, we discuss self-selected sampling.

Y_i is observed only if Z_i is greater than a cutoff value C, meaning the ith subject is selected. Without loss of generality, we can assume the cutoff C to be 0. From equation (1), the population regression function is

$$E[Y_i | X_{1i}] = X_{1i}\beta_1 \tag{3}$$

The regression function for the incomplete sample is

$$E[Y_i \mid X_{1i}, selection\ \ rule] = X_{1i}\beta_1 + E[\varepsilon_i \mid selection\,rule] = X_{1i}\beta_1 + E[\varepsilon_i \mid u_i - X_{2i}\beta_2] \qquad (4)$$

The last term in Equation 4 is equal 0 if ε_i and u_i are uncorrelated and not equal 0 otherwise.

Depending on whether Z_i is directly observed or not, we consider the following two scenarios:

Scenario A

Assuming Z_i is fully observed, we have

$$Z_i = X_{2i}\beta_2 + u_i \qquad (5)$$

$$Y_i = \begin{cases} X_{1i}\beta_1 + \varepsilon_i & if \quad Z_i > 0 \\ \text{missing} & Otherwise \end{cases} \qquad (6)$$

Scenario B

Assuming Z_i is not fully observed, then we observe a dummy D_i where

$$D_i = \begin{cases} 1 & if \quad Z_i > 0 \\ 0 & Otherwise \end{cases} \qquad (7)$$

Hence, we can write the observation equation as:

$$Y_i = \begin{cases} X_{1i}\beta_1 + \varepsilon_i & if \quad D_i = 1 \\ \text{missing} & Otherwise \end{cases} \qquad (8)$$

In practice, model B is used more often than model A. We describe the bias that arises from each model in the next section.

Self-Selected Sampling Bias

Consider scenario A. The regression function for the subsample where the data are available can be written as:

$$\begin{aligned} E\left(Y_i \mid X_{1i}, Z_i\right) &= E\left(X_{1i}\beta_1 + \varepsilon_i \mid X_{1i}, Z_i\right) \\ &= X_{1i}\beta_1 + E\left(\varepsilon_i \mid X_{2i}\beta_2 + u_i\right) \\ &= X_{1i}\beta_1 + E\left(\varepsilon_i \mid u_i = Z_i - X_{2i}\beta_2\right) \end{aligned} \tag{9}$$

Assuming (ε_i, u_i) has a bivariate normal distribution,

$$\begin{pmatrix} \varepsilon_i \\ u_i \end{pmatrix} \sim N\left(\begin{pmatrix} X_{1i}\beta_1 \\ X_{2i}\beta_2 \end{pmatrix}, \begin{pmatrix} \sigma_{11}^2 & \rho\sigma_{11}\sigma_{22} \\ \rho\sigma_{11}\sigma_{22} & \sigma_{22}^2 \end{pmatrix} \right) \tag{10}$$

hence,

$$f(\varepsilon_i, u_i) = \frac{1}{2\pi\sigma_{11}\sigma_{22}\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)} \left(\frac{\varepsilon_i^2}{\sigma_{11}^2} + \frac{z_i^2}{\sigma_{22}^2} - \frac{2\rho\varepsilon_i\, z_i}{\sigma_{11}\sigma_{22}} \right) \right). \tag{11}$$

Then the correction will have the following form

$$E(Y_i \mid X_{1i}, Z_i) = X_{1i}\beta_1 + \rho \frac{\sigma_{11}^2}{\sigma_{22}^2} (Z_i - X_{2i}\beta_2) \tag{12}$$

Now consider scenario B. The regression function for the subsample where the data are available can be written as:

$$\begin{aligned} E\left(Y_i \mid X_{1i}, D_i = 1\right) &= E(X_{1i}\beta_1 + \varepsilon_i \mid X_{1i}, Z_i > 0) \\ &= X_{1i}\beta_1 + E(\varepsilon_i \mid X_{2i}\beta_2 + u_i > 0) \\ &= X_{1i}\beta_1 + E(\varepsilon_i \mid u_i > -X_{2i}\beta_2) \end{aligned} \tag{13}$$

then the correction will have the following form

$$E(Y_i \mid X_{1i}, D_i = 1) = X_{1i}\beta_1 + \frac{\rho\sigma_{11}\sigma_{22}}{\sqrt{\sigma_{22}}} \times \frac{\phi\left(\frac{-X_{2i}\beta_2}{\sqrt{\sigma_{22}}} \right)}{1 - \Phi\left(\frac{-X_{2i}\beta_2}{\sqrt{\sigma_{22}}} \right)} \tag{14}$$

where ϕ and Φ are the standardized normal density and distribution functions respectively.

We can write Equation 14 as:

$$E(Y_i \mid X_{1i}, D_i = 1) = X_{1i}\beta_1 + w\lambda_i \tag{15}$$

equation 15 highlights the omitted variable λ_i that causes OLS estimation of Equation 1 to be biased. The variable λ_i is the hazard ratio or the Inverse Mills Ratio. For both scenarios (A and B), Equations 12 and 14 show that the estimated β_1 will be unbiased when ε_i is uncorrelated with u_i $(\rho = 0)$, so that the data are missing randomly, or the selection process is "ignorable".

In general, assume that ε_i and u_i follow a joint distribution function $f(\varepsilon_i, u_i)$ where θ is a finite set of parameters. Applying the Bayes rule, we can write:

$$E[\varepsilon_i \mid u_i > -X_{2i}\beta_2] = \frac{\int_{-\infty}^{\infty}\int_{-X_{2i}\beta_2}^{\infty} \varepsilon_i f(\varepsilon_i, u_i, \beta_2)\, d\varepsilon_i du_i}{\int_{-X_{2i}\beta}^{\infty}\int_{-\infty}^{\infty} f(\varepsilon_i, u_i, \beta_2)\, d\varepsilon_i du_i} = \lambda(X_{2i}, \beta_2) \tag{16}$$

Here $\lambda(X_{2i}, \beta_2)$ could be a nonlinear function of $X_{2i}\beta_2$ and the parameters θ. This means that the conditional expectation of Y_i given X_{1i} and the probability that Y_i is observed will be equal to the usual regression function X_{1i}, β_1 plus a nonlinear function of the selection equation regressors X_{2i} that has non-zero mean as we showed in Equation 4.

Therefore, when estimating β_1, we can conclude that the estimated intercept will be biased because the mean of the residuals is not zero. Also, if X_{1i} and X_{2i} are not completely uncorrelated (i.e. they have variables in common or they are correlated), the estimated slope coefficient will be biased because there is an omitted variable in the regression, namely $\lambda(X_{2i}, \beta_2)$, that is correlated with the included variable X_{1i}. We can see that even if X_{1i} and X_{2i} are independent, the fact that the data is nonrandomly missing will introduce heteroskedasticity to the error term, so OLS is not fully efficient.

MLE and Heckman Correction

There are two major existing approaches for estimating the self-selected sample model under the assumption of bivariate normal. The first method is FIML and the second is Heckman's well-known two-step procedure. We discuss each of these methods and adopt Heckman’s method as a benchmark for simulation comparisons because of its popularity. We will consider scenario B in the selection stage for both methods. In practice, it is more common to assume that Z_i is not fully observed.

In the maximum likelihood approach, we specify a complete model setup as in Equations 1 and 2, and we assume the following joint distribution for (ε_i, u_i)

$$\begin{pmatrix} \varepsilon_i \\ u_i \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1 \\ \rho\sigma_1 & 1 \end{pmatrix}\right) \tag{17}$$

We typically assume a bivariate normal distribution with zero and means and correlation ρ. There is no generally accepted name for this model. The restriction $\sigma_2^2 = 1$ is used to simplify the calculations of the likelihood function.

We divide the observations into groups according to the type of data observed. Each group of observations will have a different form for the likelihood. For example, for the sample selection model, there are two types of observations: (1) those where Z>0 and, (2) those where Y_i is not observed and we know that $Z_i \leq 0$.

For those where $Z > 0$. For these observations, the likelihood function is the probability of the joint event Y and $Z > 0$. We can write this probability for the ith observation as the following:

$$
\begin{aligned}
P(Y_i, Z_i > 0 \mid X_{1i}, X_{2i}) &= f(Y_i)P(Z_i > 0 \mid Y_i, X_{1i}, X_{2i}) \\
&= f(\varepsilon_i)P(u_i > -X_{2i}\beta_2 \mid \varepsilon_i, X_{1i}, X_{2i}) \\
&= \frac{1}{\sigma_1}\phi\left(\frac{Y_i - X_{1i}\beta_1}{\sigma_1}\right)\int_{-X_{2i}\beta_2}^{\infty} f(u_i \mid \varepsilon_i)du_i \\
&= \frac{1}{\sigma_1}\phi\left(\frac{Y_i - X_{1i}\beta_1}{\sigma_1}\right)\int_{-X_{2i}\beta_2}^{\infty} \phi\left(\frac{u_i - \frac{\rho}{\sigma_1}(Y_i - X_{1i}\beta_1)}{\sqrt{1-\rho^2}}\right)du_i \\
&= \frac{1}{\sigma_1}\phi\left(\frac{Y_i - X_{1i}\beta_1}{\sigma_1}\right)\left(1-\Phi\left(\frac{-X_{2i}\beta_2 - \frac{\rho}{\sigma_1}(Y_i - X_{1i}\beta_1)}{\sqrt{1-\rho^2}}\right)\right) \\
&= \frac{1}{\sigma_1}\phi\left(\frac{Y_i - X_{1i}\beta_1}{\sigma_1}\right)\times\Phi\left(\frac{X_{2i}\beta_2 - \frac{\rho}{\sigma_1}(Y_i - X_{1i}\beta_1)}{\sqrt{1-\rho^2}}\right) \qquad (18)
\end{aligned}
$$

Thus, the probability of an observation for which we see the data is the density function at the point Y_i multiplied by the conditional probability distribution for Z_i given the value of Y_i that was observed.

For those where Y_i is not observed and we know that $Z_i \le 0$. For these observations, the likelihood function is just the marginal probability that $Z_i \le 0$. We have no independent information on Y_i. This probability is written as

$$P(Z_i \le 0) = P(u_i \le -X_{2i}\beta_2) = \Phi(-X_{2i}\beta_2) = 1 - \Phi(X_{2i}\beta_2) \qquad (19)$$

therefore, if we assume the first N_1 observations have $Z_i > 0$ and the rest have $Z_i \le 0$, then the log likelihood for the complete sample of observations is the following:

$$\log L(\beta_1, \beta_2, \rho, \sigma; data) = \sum_{i=N_0+1}^{N} [-\log\sigma_1 + \log\phi\left(\frac{Y_i - X_{1i}\beta_1}{\sigma_1}\right) + \log\Phi\left(\frac{X_{2i}\beta_2 - \frac{\rho}{\sigma_1}(Y_i - X_{1i}\beta_1)}{\sqrt{1-\rho^2}}\right) + \sum_{i=1}^{N_0} \log(1-\Phi(X_{2i}\beta_2))] \qquad (20)$$

In the above log likelihood, there are N_0 observations where we do not observe Y, and there are N_1 observations where we do observe Y. Then $N_0 + N_1 = N$. The parameter estimates for the sample selection model can be obtained by maximizing this likelihood function with respect to its arguments.

As an alternative to MLE, Heckman (1979) developed a two-step model that is widely used for sample selection bias. Heckman’s model is based on two latent dependent variables. The steps of Heckman's estimation are:

(a) Estimate β_2 in Equation 2 using a probit model;

(b) Use the estimated $X_{2i}\beta_2$ to calculate

$$\lambda(X_{2i}\beta_2) = E(\varepsilon_i \mid u_{2i} > -X_{2i}\beta_2) = \frac{\phi(-X_{2i}\beta_2)}{1-\Phi(-X_{2i}\beta_2)}; \qquad (21)$$

(c) Estimate β_1 in equation (5) by replacing $E(\varepsilon_i \mid u_{2i} > -X_{2i}\beta_2)$ with $\lambda(X_{2i}\beta_2)$. Estimation of Equation 15 by OLS gives consistent parameter estimates, but special formulas are needed to get correct standard errors because the errors V_i are correlated.

If $\rho = 0$, the usual formula provides a consistent estimate of the covariance matrix of the parameters in the second-stage regression. Heckman suggests that we use the t-test of the coefficient on the λ variable as a test of sample selection bias. Melino (1982) shows that this represents the optimal test of selectivity bias, under the maintained distributional assumptions, as it is based on the same moment as the Lagrange multiplier test. That is, both the Lagrange multiplier test and the t-test for the coefficient on λ_i are based on the correlation between the errors in the primary equation and the errors from the selection equation. Note that the Inverse Mills Ratio is the error from the probit equation explaining selection.

In other words, Heckman's proposal is to estimate the Inverse Mills Ratio

$$\lambda\left(\frac{X_{2i}\beta_2}{\sigma_{22}}\right) = \frac{\phi\left(\frac{-X_{2i}\beta_2}{\sqrt{\sigma_{22}}}\right)}{1-\Phi\left(\frac{-X_{2i}\beta_2}{\sqrt{\sigma_{22}}}\right)} \tag{22}$$

in a probit model and then estimate Equation 15 by OLS to obtain consistent estimates of β_1 and λ_i,

Although Heckman's two-step procedure gives a consistent estimator, various papers criticize its small sample properties. Many claims were that the predictive power of subsample OLS or the two-step model is at least as good as the one of Heckman's procedure or MLE. Here the two-step model gives the conditional expectation of wages. Daun (1984) contends that the conditional expectation is of interest to us. In addition, we interpret the coefficient of the two part model with the same way that we estimate the wage equation by subsample OLS. Stolzenberg and Relles (1990) provide evidence that the higher the correlation between the error terms, the greater the superiority of the maximum likelihood (and maybe OLS) estimator over Heckman procedure in terms of efficiency.

The most important line of criticism of Heckman's procedure is based on practical rather than theoretical grounds. If the set of X_{1i} variables that affect the wage in the wage equation are almost identical with the set of X_{2i} variables that affect labor force participation in selection equation, then the second step of Heckman's method is only identified through the nonlinearity of the Inverse Mills Ratio. In many practical cases, we only observe values within the quasi-linear (not completely linear) range of the inverse mills ratio. Then we need X_{2i} variables that are good predictors of labor force participation and do not appear in X_{1i} which are difficult to find in practice.

Most studies find that the two-step approach can be unreliable in the absence of exclusion restrictions. Generally, an exclusion restriction is required to generate credible estimates: there must be at least one variable which appears with a non-zero coefficient in the selection equation but does not appear in the equation of interest. If no such variable is available, it may be difficult to correct for sampling selectivity. Leung and Yu (1996) conclude that this result is due to experimental design. They find that Heckman's two-step estimator is effective, provided that at least one X_i displays sufficient variation to induce tail behavior in the Inverse Mills Ratio. Under certain circumstances, even when its assumptions and formal requirements are satisfied, the two-step selection bias correction is known to produce estimates that are farther from true parameter values than estimates obtained by uncorrected OLS. Puhani (2000) strongly recommends exploratory work to check for collinearity problems before deciding on which estimator to apply. If there is no collinearity between the X_{1i} regressors and the Inverse Mills Ratio, the Heckman two-part model is the most robust approach. On the other hand, if collinearity problems exist, the MLE approach is preferable to Heckman’s two-step method.

In the next chapter, we propose a Bayesian method to estimate sample selection bias. We study its behavior by comparing our estimates to Heckman's estimates.

CHAPTER 3

MODEL DEVELOPMENT

Brief Introduction to Bayesian Methods and MCMC

The Bayesian approach is fundamentally different from the conventional approach. In the conventional approach a sample $X_1,........X_n$ is drawn from a population with an unknown but fixed parameter θ. Knowledge about θ is obtained from the observed random sample. In the Bayesian approach, θ is considered to be a random variable and its variation can be described by a probability distribution called the *prior distribution*. This is a subjective distribution, based on the researcher's belief, and is formulated before the data are seen (hence "prior"). When a sample from a population indexed by θ is observed, the prior distribution is updated with the information in the sample. The updated prior is called the *posterior distribution*. The Bayesian approach is concerned with generating the posterior distribution of the parameters and provides a more complete picture of the uncertainty in the estimation of unknown parameters, especially after the confounding effects of nuisance parameters are removed. A complete introduction to Bayesian analysis can be found in Lee (1997) and Draper (2000).

The foundation of Bayesian statistics is Bayes' Theorem which is used to update the posterior distribution. Bayes' Theorem is named after Thomas Bayes (1702-1761). He calculated the probability of a new event on the basis of earlier probability estimates that have been derived from empirical data. Bayes' work became the basis of a statistical technique, which is now called Bayesian statistics or Bayes' method.

The basic principle of Bayes theorem is as follows. If event A occurred, the probability that event E_i also occurred is

$$P(E_i \mid A) = \frac{p(A \mid E_i)p(E_i)}{p(A)} = \frac{p(A \mid E_i)p(E_i)}{\Sigma p(A \mid E_i)p(E_i)}$$

In structured modeling and analysis, Bayes' method can be written in the following equation. Assume we observe data y from distribution with parameter of θ and we wish to make inference about another random variable θ, where θ is drawn from some distribution π(θ). Then

$$p(\theta \mid y) = \frac{\pi(\theta)p(y \mid \theta)}{p(y)} = \frac{\pi(\theta)p(y \mid \theta)}{\int p(y)p(y \mid \theta)d\theta}$$

where y is a vector of the observed data and θ is the unknown parameters. The posterior probability conditional on y is p(θ|y). The prior distribution is π(θ) and it can be informative or non-informative. (An informative prior expresses specific, definite information about a variable). The likelihood function is p(y|θ) when it is regarded as a function of θ for a fixed y^*. The prior predictive distribution, also called the marginal distribution of y, is p(y).

With Bayes' model, we can estimate the posterior distribution, p(θ|y), by integrating the full Bayes equation (the likelihood and prior probability functions). For example, if y represents a random sample from $N(\theta, \sigma^2)$, then we have:

$$p(y \mid \theta, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(y-\theta)^2}{2\sigma^2}\right)$$

under the non-informative prior:

$$\pi(y \mid \theta, \sigma^2) \propto \frac{1}{\sigma^2}$$

the posterior distribution is:

$$\pi(\theta, \sigma^2 \mid y) \propto \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2+1}} \times \left(-\frac{\sum (y_i - \theta)^2}{2\sigma^2}\right)$$

Another approach is to use MCMC simulation to obtain the posterior distribution. Metropolis (1953) showed how this method helps in constructing a Markov Chain with stationary distribution. The method was generalized by Hastings (1970) and is now widely used to sample from analytically intractable probability distributions arising in statistics (Gilks 1996; Robert and Casella, 1999). The efficiency of MCMC methods is of significant practical importance, and loosely speaking, is determined by the convergence rate of the chain. In contrast to the maximum likelihood method, the MCMC Bayesian method is useful and reliable even for finite sample sizes, since convergence results depend only on the number of iterations.

The main advantage of Bayesian methodology is that in the absence of much data, the prior distribution carries a lot of weight; but the more data that are observed, the less influence the prior distribution has on the posterior distribution. The most common criticism of Bayesian methodology is that since there is no single correct prior distribution, then all conclusions drawn from the posterior distribution are suspect.

We develop a Bayesian method for estimating the parameters of the self-selected sampling model. We implement MCMC methods and Gibbs sampling to facilitate computation for the posterior estimates. We also conduct a simulation study to determine the performance of the MCMC algorithm for various prior distributions.

Missing Values and Latent Variables

Recall scenario B in section 2.1, and latent variable $Z_i = X_{2i}\beta_2 + u_i$, where

$$Z_i \sim N(x_{2i}\beta_2, 1) \text{ and } \varepsilon \sim N(0,1)$$

One can show that $P(D_i = 1) = P(Z_i > 0) = \Phi(x_{2i}\beta)$, where Φ is the cumulative density function (cdf) of $N(0,1)$.

The priors for other parameters remain the same:

$$\begin{pmatrix} Y_i \\ Z_i \end{pmatrix} \sim BVN\left(\begin{pmatrix} x_{1i}\beta_1 \\ x_{2i}\beta_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1 \\ \rho\sigma_1 & 1 \end{pmatrix}\right)$$

note that Z_i is a latent variable and it needs to be sampled. Also, Y_i $(i = m+1, ..., n)$ are missing and they will be sampled as well. The below steps show how to sample Z_i

i) If $D_i = 1$, then y_i is observed for $i = 1, ...m$ given initial or sampled values (MCMC) of $\beta_1, \beta_2, \rho, and\ \sigma^2$, we can show that Z_i is normally distributed with following conditional mean and conditional variance:

$$(Z_i \mid \beta_1, \beta_2, \rho, \sigma_1^2; y_i, D_i = 1) \sim N(x_{2i}\beta_2 + \frac{1}{\sigma_1} \times \rho(Y_i - x_{1i}\beta_1), (1-\rho^2))$$

and $Z_i > 0$

We can sample Z_i from truncated Normal by the above equation and $Z_i > 0$.

ii) If $D_i = 0$, then y_i is observed for $i = m+1,........n$ hence, sample

$$\begin{pmatrix} Y_i \\ Z_i \end{pmatrix} \sim BVN\left(\begin{pmatrix} x_{1i}\beta_1 \\ x_{2i}\beta_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1 \\ \rho\sigma_1 & 1 \end{pmatrix}\right)$$

with the restriction $Z_i < 0$ (since $D_i = 0$). One way to do this is to generate (Y_i, Z_i) jointly until we get a sample with $Z_i < 0$.

Now recall scenario B, Z_i is fully observed and Y_i is not. The joint distribution is:

$$\begin{pmatrix} Y_i \\ Z_i \end{pmatrix} \sim BVN\left(\begin{pmatrix} x_{1i}\beta_1 \\ x_{2i}\beta_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right),$$

for missing Y_i $(i = m+1,.......n)$, one can sample

$$Y_i \mid \beta_1, \beta_2, \rho, \sigma_1^2, \sigma_2^2; Z_i \sim N(x_{1i}\beta_1 + \frac{\sigma_1}{\sigma_2} \times \rho(Z_i - x_{2i}\beta_2), (1-\rho^2) \times \sigma_1^2)$$

Prior Distribution of Linear Component

Consider the multivariate regression model $Y_i = X_i\beta + \varepsilon_i \quad for \quad i = 1,.....n$ where Y_i is an m-vector of dependent variables for unit i; X_i is an $m \times p$ matrix of independent variables for unit i, β is a p-vector of regression coefficients; and ε_i is the error term. The error terms are mutually independent random variables from a multivariate normal distribution with mean zero and covariance matrix Σ , $\varepsilon \sim N_m(0, \Sigma)$.

A well-accepted Bayesian approach is to consider the normal distribution as the prior of β is because it is quite flexible. We will consider the prior distributions:

$$\beta \sim N_p(\mu_\beta, \Sigma_\beta)$$

$$\Sigma \sim IW_m(df, H)$$

Where $IW_m(df, H)$ is the m*-dimensional*, inverted Wishart distribution with df prior degrees of freedom and scale parameters H with density function:

$$p(H) = \frac{|H|^{\frac{1}{2}(m-p-2)} \exp(-\frac{1}{2} tr(\beta\Sigma^{-1}))}{2^{-\frac{1}{2}p(m-1)} \pi^{\frac{1}{4}p(p-1)} |\Sigma|^{\frac{1}{2}(m-1)} \Pi_1^p \Gamma\left(\frac{m-1}{2}\right)}$$

Now, recall Heckman's sample selection model Equations 1 and 2:

$$Y_i = X_{1i}\beta_1 + \varepsilon_i$$

$$Z_i = X_{2i}\beta_2 + u_i$$

for the Bayesian analysis of the above model, we assume the joint distribution of ε_i and u_i is bivariate normal:

$$\begin{pmatrix} \varepsilon_i \\ u_i \end{pmatrix} \sim N(0, \Sigma), \qquad \Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

where ρ is the correlation coefficient. To compute this model we need to provide the prior distribution using β

$$\beta_1 \sim N_p(\mu_\beta, \Sigma_{\beta_1})$$

$$\beta_2 \sim N_p(\mu_q, \Sigma_{\beta_2})$$

since ρ is unknown, we use the Inverse Wishart distribution:

$$\Sigma \sim IW_m(df, H)$$

Sampling Scheme of Linear Component

Considering the multivariate normal distribution $Y \sim N(X\beta, \Sigma)$, where

$$\beta \sim N(\mu_\beta, \Sigma_\beta) \text{ And } \Sigma \sim \pi(\Sigma)$$

the conditional posterior of β can be calculated using

$$f(\beta \mid Y, \Sigma) = f(y \mid \beta) \times f(\beta)$$

the conditional posterior is proportional to the exponent part, so we can write:

$$\begin{aligned} f(\beta \mid Y, \Sigma) &\propto \exp[-\frac{1}{2}(Y - X\beta)^T \Sigma^{-1}(Y - X\beta)] \times \exp[-\frac{1}{2}(\beta - \mu_\beta)^T \Sigma^{-1}(\beta - \mu_\beta)] \\ &\propto \exp[-\frac{1}{2}(Y - X\beta)^T \Sigma^{-1}(Y - X\beta)] \times \exp[-\frac{1}{2}(\beta - \mu_\beta)^T \Sigma^{-1}(\beta - \mu_\beta)] \end{aligned}$$

taking the first and second derivatives with respect to β, we can find the conditional precision matrix Ω. Thus the first derivative:

$$\frac{\partial f(\beta)}{\partial f\beta} = -2X^T\Sigma^{-1}(Y - X\beta) + 2\Sigma_\beta^{-1}(\beta - \mu_\beta)$$

and the second derivative:

$$\frac{\partial^2 f(\beta)}{\partial f\beta^2} = -2X^T\Sigma^{-1}X + 2\Sigma_\beta^{-1} = 2(X^T\Sigma^{-1}X + \Sigma_\beta^{-1})$$

therefore, the variance or the conditional precision matrix is:

$$\Omega^{-1} = X^T\Sigma^{-1}X + \Sigma_\beta^{-1}.$$

setting the first derivative equal to zero, we can calculate the mean

$$\begin{aligned} 0 &= -2X^T\Sigma^{-1}(Y - X\beta) + 2\Sigma_\beta^{-1}(\beta - \mu_\beta) \\ &= -Y^T\Sigma^{-1}X + \beta^T X^T\Sigma^{-1}X + \beta^T\Sigma_\beta^{-1} - \mu_\beta^T\Sigma_\beta^{-1} \end{aligned}$$

we can write:

$$\beta^T X^T\Sigma^{-1}X + \beta^T\Sigma_\beta^{-1} = Y^T\Sigma^{-1}X + \mu_\beta^T\Sigma_\beta^{-1}$$

thus,

$$\begin{aligned} \beta^T &= (Y^T\Sigma^{-1}X + \mu_\beta^T\Sigma^{-1}) \times (X^T\Sigma^{-1}X + \Sigma_\beta^{-1})^{-1} \\ &= (Y^T\Sigma^{-1}X + \mu_\beta^T\Sigma_\beta^{-1})\Omega^{-1} \end{aligned}$$

and

$$\beta = \Omega^{-1}(X^T \Sigma^{-1} Y + \Sigma_\beta^{-1} \mu_\beta)$$

therefore,

$$f(\beta \mid Y, \Sigma, \mu_\beta, \Sigma_\beta) \sim MVN[\Omega^{-1}(X^T \Sigma^{-1} Y + \Sigma_\beta^{-1} \mu_\beta), \Omega^{-1}]$$

the above posterior mean is a weighted average of the data Y and the prior mean (μ_β), with weights given by the data prior precision matrices $(\Sigma_\beta^{-1}\, and\, \Omega^{-1})$.

The conditional posterior of $\Sigma = (\Sigma^{-1})$ can be calculated using

$$f(\Sigma^{-1} \mid Y, \beta) = f(y \mid \Sigma) \times \pi(\Sigma^{-1})$$

thus, the conditional posterior can be written as:

$$f(\Sigma^{-1} \mid Y, \beta) \propto \prod_{i=1}^{m} f(y_i \mid \beta, \Sigma) \times \pi(\Sigma^{-1})$$

$$\propto \left|\Sigma\right|^{-\frac{m}{2}} \exp(-\frac{1}{2} \sum_{i=1}^{m} (y_i - x_i \beta)^T \Sigma^{-1} (y_i - x_i \beta))$$

the above can be expressed as:

$$\propto \left|\Sigma^{-1}\right|^{-\frac{1}{2}(n+m-df-1)} \exp[-\frac{1}{2} tr(\sum_{i=1}^{m} (y_i - x_i \beta)^T \Sigma^{-1} (y_i - x_i \beta) + H^{-1}) \Sigma^{-1})]$$

also, we can write:

$$\propto \left|\Sigma^{-1}\right|^{-\frac{1}{2}(df+m-2-1)} \exp[-\frac{1}{2}(S + H^{-1}) \Sigma^{-1}]$$

where,

$$S = \sum_{i=1}^{m} (y_i - x_i \beta)(y_i - x_i \beta)^T$$

this leads to

$$\Sigma^{-1} \mid Y, \beta \sim Wishart(df + m, (S + H^{-1})^{-1})$$

The model developed in this section will be used to analyze data examples and simulated data in the next chapter. We show the results of MCMC simulations carried out and evaluate sampling properties of the estimators discussed in this section.

CHAPTER 4

SIMULATION STUDY

In this chapter we apply the proposed Bayesian approach by conducting two simulations studies using MCMC methods and Gibbs sampling to facilitate computation of posterior estimates. The first simulation study uses an artificial data example to determine the performance of MCMC using various priors. The second simulation is a comprehensive simulation study using a generated data with the ability to test various data scenarios such as: sample missing rate, residuals correlation, multicollinearity, and sample size.

Women Wage Example

Discussions in the context of labor economics concerning labor force population, wages, and earnings highlight the importance of sample selection. One representative example is the estimation of women's wages. Since we only observe the wages of women who enter the workforce, our sample represents only one part of the wage offer distribution. Other secondary wage groups, such as married women and teenagers, are not represented. Therefore, estimation procedures may involve certain bias when applied to the secondary wage groups. This is an example of self-selected sampling bias.

We propose a Bayesian MCMC algorithm to estimate parameters of a self-selected sampling model. We consider the labor force example in the STATA user manual. This data is used to illustrate Heckman's approach by predicting women's wages from their education and age. To evaluate the performance of the proposed Bayesian approach and compare with other methods, we simulate sample selection as it was specified in the example, and perform the MCMC algorithm using different Wishart prior specifications.

MCMC methods use simulation of Markov chains in the parameter space. The Markov chains are defined in such a way that the posterior distribution in the gives statistical inference problem is asymptotic distribution. This allows using averages to approximate the desired posteriors expectations. Several standard algorithms to define such Markov chains exist, including Gibbs sampling and Metropolis-Hasting. Using these algorithms it is possible to implement posterior simulation in essentially any problem which allows pointwise evaluation of prior distribution and likelihood functions.

The data contain a sample of 2,000 observations of 15 variables. A brief description of the variables that are relevant for our analysis is shown in Table 1 From among the 2,000 observations; we observe wage data for only 1,343. The remaining 657 women were not in the paid work force and so did not receive wages. We are interested in modeling two things: (1) the decision of the women to enter the labor force and (2) predicting women's hourly wage. We will consider a reasonable assumption that the women's decision to enter the labor force is a function of age, marital status, the number of children, and her level of education. Also, the wage rate a woman earns is a function of her age and education.

Table 1

Women Wage Data Variable Description

Variable Name	Definition
Age	Age of the woman
Education	Number of years of education of the woman
Married	Dummy variable equal to 1 if the woman is married 0 otherwise
Children	Number of children that the woman has in her household
Wage	Hourly wage of the woman

We begin with OLS estimation of the regression model using only the observations that have wage data. The estimates can be found in Table 2 (see `OLS-Selected Wage' row). This analysis would be fine if, in fact, the missing wage data were missing completely at random. However, the decision to work or not work was made by the individual woman. Thus, those who were not working constitute a self-selected sample and not a random sample. It is likely that some of the women who would have earned low wages chose not to work. If so, this would account for much of the missing wage data. Thus, it is likely that we will over-estimate the wages of the women in the population. So, somehow, we need to account for information that we have on the non-working women. We attempt to do this by replacing the missing values with zeros for wage variable. The estimates can be found in Table 2 (see ‘OLS- Non-missing Wage’ row).

Table 2

Estimation Using OLS and Heckman Models

Method	Parameter	Parameter Estimate	Standard Error	Bias
OLS – Full Wage	Intercept	1.381	0.743	NA
OLS – Selected Wage	Intercept	6.085	0.890	-4.704
OLS – Non-Missing Wage	Intercept	-12.168	1.398	13.550
Heckman	Intercept	0.734	1.166	0.647
OLS – Full Wage	Education	1.004	0.045	NA
OLS – Selected Wage	Education	0.897	0.050	0.108
OLS – Non-Missing Wage	Education	1.065	0.084	-0.060
Heckman	Education	0.983	0.051	0.022
OLS – Full Wage	Age	0.187	0.016	NA
OLS – Selected Wage	Age	0.147	0.019	0.041
OLS – Non-Missing Wage	Age	0.391	0.031	-0.203
Heckman	Age	0.212	0.021	-0.024
OLS – Full Wage	Inverse Mills	NA	NA	NA
OLS – Selected Wage	Inverse Mills	NA	NA	NA
OLS – Non-Missing Wage	Inverse Mills	NA	NA	NA
Heckman	Inverse Mills	4.002	0.577	NA

This analysis is also troubling. It is true that we are using data from all 2,000 women but zero is not a fair estimate of what the women would have earned if they had chosen to work. It is likely that this model will under-estimate the wages of women in the population. The solution to our quandary is to use the Heckman selection model (Heckman 1979).

The Heckman selection model allows us to use information from non-working women to improve the estimates of the parameters in the regression model. The Heckman selection model provides consistent, asymptotically efficient estimates for all parameters in the model. In our example, we have one model predicting wages and one model predicting whether a woman will be working. We will use marital status, children, education and age to predict selection. In addition to the two equations, Heckman estimates ρ (actually the inverse hyperbolic tangent of ρ), the correlation of the residuals in the two equations and Σ (actually the log of Σ), the standard error of the residuals of the wage equation. Then λ = ρΣ. The estimates can be found in Table 2 (see 'Heckman' row).

Recall that we do have full wage information on all 2,000 women. We can therefore run a regression using the full wage information to use as a comparison. The estimates can be found in Table 2 (see 'OLS- Full Wage' row). The 'Selected Wage' model tends to over-estimate wages; the 'Non-Missing Wage' model tends to severely under-estimate wages; and the Heckman model does the best job in predicting wages.

Finally, we consider the Bayesian approach to predicting women's wage from their education and age. In this approach, the posterior distributions are too complicated to evaluate analytically. However, by using MCMC methods and Gibbs sampling, this posterior distribution can be sampled indirectly by generating a sample of parameter values from the conditional distribution of interest. Posterior Bayes estimates are then obtained from the generated samples. We estimate the parameters using the MCMC algorithm as the following:

1. Use Bayesian approach using many loops for Gibbs sampling / MCMC to repeatedly sample from the conditional distribution
   - Sample latent variable in selection stage
   - Update missing y1's and mean prediction of y1's
   - Update β_1 and β_2 jointly
   - Update Σ using Wishart prior which leads to Wishart posterior

2. Repeat the above steps 1,500 times

The goal is to see how this method performs when we use different priors and different correlations of the error terms in the two-model equation. We run the algorithm for 1,500 iterations after convergence, discarding the first 500 iterations. The estimates are in Table 3. Comparing the results in Table 2 and Table 3 we find that the Bayesian approach is providing estimates that are at least as effective as the Heckman's estimates and are better than the OLS using selected wage.

The Wishart distribution is an objective prior because the posterior mean will be affected by the prior choice. We perform our analysis under two instances of a Wishart prior: Wish(3,H) & Wish(4,H). Also, we change the inverse scale matrix in the Wishart prior in using different values of σ: (0.01,0.1,1,5,10,100). The results in Table 3 show that the parameters did not change much compared with the results in Table 2. This indicates that the Bayesian approach is performing as well as the Heckman approach or even better in some scenarios. We can see good estimates for β_1 under Wishart(3,H) with Sigma(10,0,0,10). The standard error and bias are low compared to other estimates.

Table 3

Estimation Using Bayesian Methods

		Intercept			Education		
Wishart	Inverse Scale	Parameter	Standard		Parameter	Standard	
Prior	Matrix	Estimate	Error	Bias	Estimate	Error	Bias
(3, H)	(.01,0,0,.01)	0.5157	1.1578	0.8654	0.9920	0.0544	0.0124
(3, H)	(.1,0,0,.1)	0.4286	1.0115	0.9525	0.9914	0.0526	0.0130
(3, H)	(1,0,0,1)	0.7894	1.0293	0.5917	0.9867	0.0531	0.0177
(3, H)	(5,0,0,5)	0.9148	1.0784	0.4663	0.9788	0.0513	0.0256
(3, H)	(10,0,0,10)	1.0153	1.0484	0.3658	0.9816	0.0542	0.0228
(3, H)	(100,0,0,100)	3.3957	0.9819	-2.0146			

		Age		
Wishart	Inverse Scale	Parameter	Standard	
Prior	Matrix	Estimate	Error	Bias
(3, H)	(.01,0,0,.01)	0.2116	0.0212	-0.0242
(3, H)	(.1,0,0,.1)	0.2135	0.0200	-0.0261
(3, H)	(1,0,0,1)	0.2095	0.0202	-0.0221
(3, H)	(5,0,0,5)	0.2089	0.0210	-0.0215
(3, H)	(10,0,0,10)	0.2067	0.0203	-0.0193
(3, H)	(100,0,0,100)	0.1786	0.0196	0.0088

		Intercept			Education		
Wishart	Inverse Scale	Parameter	Standar		Parameter	Standard	
Prior	Matrix	Estimate	d Error	Bias	Estimate	Error	Bias
(4, H)	(.01,0,0,.01)	0.3787	1.0783	1.0024	0.9901	0.0547	0.0143
(4, H)	(.1,0,0,.1)	0.2861	1.0329	1.0950	0.9952	0.0530	0.0092
(4, H)	(1,0,0,1)	0.1410	1.0890	1.2401	1.0003	0.0544	0.0041
(4, H)	(5,0,0,5)	0.7795	1.0366	0.6016	0.9865	0.0556	0.0179
(4, H)	(10,0,0,10)	0.9294	1.1339	0.4517	0.9831	0.0538	0.0213
(4, H)	(100,0,0,100)	3.5064	1.0478	-2.1253	0.9375	0.0493	00669

		Age		
Wishart	Inverse Scale	Parameter	Standar	
Prior	Matrix	Estimate	d Error	Bias
(3, H)	(.01,0,0,.01)	02151	0.0201	-0.0277
(3, H)	(.1,0,0,.1)	0.2149	0.0211	-0.0275
(3, H)	(1,0,0,1)	0.2160	0.0202	-0.0286
(3, H)	(5,0,0,5)	0.2090	0.0196	-0.0216
(3, H)	(10,0,0,10)	0.2079	0.0216	-0.0205
(3, H)	(100,0,0,100)	0.1780	0.0204	0.0094

Comprehensive Simulation Study

Here we conduct a more comprehensive simulation study to further investigate the effect of prior distributions and the robustness of the Bayesian approach. Unlike the work in the previous section, the data sets here will be generated from several specific conditions. More specifically, I consider the effects of the fraction of selection, correlation between selection of regression models, sample sizes, and the fraction of the independent variables that appear in both selection and regression model. The simulated data will be analyzed by both Heckman's two-stage estimator and the Bayesian methods proposed above. The estimates from both methods will be evaluated and compared in terms of Bias and RMSE (Root Mean Square Error).

The Bias of an estimator is the difference between the estimator's expected value and the true parameter value of the parameter being estimated. An estimator or decision rule with zero bias is called unbiased. Otherwise the estimator is said to be biased. RMSE is based on two sums of squares: Sum of Squares Total (SST) and Sum of Squares Error (SSE). SST measures how far the data are from the mean and SSE measures how far the data are from the model's predicted vales.

The data set will be generated using the two model equation with the below self-selected samples.

Observation stage: $Y_i = X_{1i}\beta_1 + \varepsilon_i$

Selection stage: $Z_i = X_{2i}\beta_2 + u_i$

Where $\begin{pmatrix} \varepsilon_i \\ u_i \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{matrix} 1 & \rho\sigma_1 \\ \rho\sigma_1 & 1 \end{matrix}\right)$

Note that Y_i is the observation stage and Z_i is the selection stage. Without loss of generality, we will observe $y_1, y_2, ...y_n \ (n < m)$ where $z_1, z_2, ...z_n$ are greater than the selection cutoff value c. Hence, $y_{(n+1)}, y_{(n+2)}, ...y_m$ will be considered missing at the analysis step. We consider the following scenarios at the data generation step:

- Change the level or percentage of missing values
- Change the value of ρ to control the level of correlation
- Change the two scalars x_{1i} and x_{2i} to control the level of multicollinearity
- Change the sample size

The data are simulated as follows: we generate design matrices X_{1i} and X_{2i} in observation and screening stage with 3 × 1 vector for each with their first rows fixed as one to make constant terms for each equation. The two other rows in X_{1i} and X_{2i} are independently generated from a uniform distribution on [0,1]. I set the true parameter value for $\beta_0, \beta_1, \beta_2$ as 1,2,3 respectively and I generate random error ε_i and u_i from a bivariate normal density with zero mean and variance-covariance matrix (1 ρ, ρ 1).

In each case 1,500 Gibbs samples were drawn, the first 500 were discarded, and the remaining 1,000 were used for posterior inference. I tried multiple runs to ensure convergence of the results. The main simulation didn't include 'Thinning' (strategy of reducing autocorrelation by storing only every mth point after the burn-in period), however, we are confident with the results because we use different settings with different starting points and the results are obtained from 100 replications to avoid systematic mistakes. Since 1,500 might not be long enough for MCMC to converge, I selected one simulation scenario from one chain and ran 11,000 MCMC samples using two different starting values. The first 1,000 were discarded as burn-in, and we applied thinning of 5, leaving 2,000 effective posterior samples to show plot behavior and how the distribution converges. Convergence refers to the idea that eventually the MCMC and Gibbs Sampler that we choose did eventually reach a stationary distribution, which is also our target distribution.

To test the results, we generated the following diagnosis plots: trace plots to show the sampling path, kernel density to show the posterior density function, and moving averages of posterior samples to show that samples are converging to similar values. One way to see if our chain has converged is to see how well our chain is mixing, or moving around the parameter space. If our chain is taking a long time to move around the parameter space, then it will take longer to converge. We can see how well our chain is mixing through visual inspection. We will discuss these inspections for every parameter.

Figure 1 shows the sampling paths for $\beta_0, \beta_1, \beta_2$ from two different starting chains. This figure contains plots known as trace plots of the iteration number against the value of the draw of the parameters at each iteration. These plots are useful to show whether our chain is converging to the same value or gets stuck in certain areas of the parameter space, which indicate bad mixing. Our results show that all samples converge to the same distribution and there seems to be large spread for β_2 estimates.

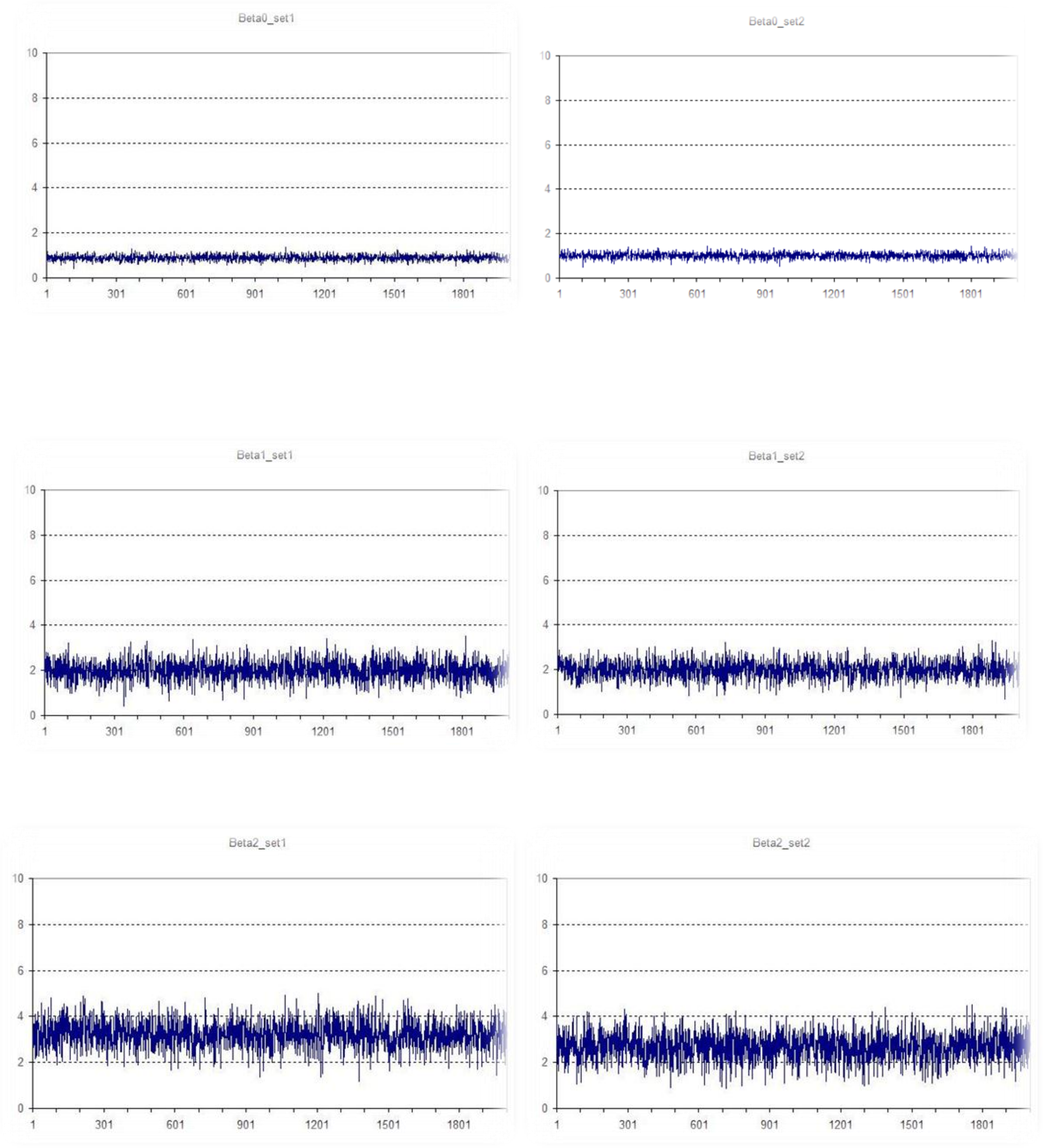


Figure 1. *Sample Path for Comprehensive Simulation*

Figure 2 shows the posterior density plots for the estimates from the two MCMC chains with the normal density curve. The plots show strong evidence for convergence for β_0 and β_1 which is reflected in the distributions. Usually, non-convergence is reflected in multimodal distribution and this is especially true if the kernel density is not just multi-modal, but lumpy.

Figure 3 shows the plots of the moving averages of $\beta_0, \beta_1, \beta_2$ from the two MCMC chains. The x-axis represents the number of iterations and the y-axis shows the posterior mean from these iterations. As a result, all the paths are believed to be stationary in an acceptable rang. When comparing the two settings, β_0 and β_1 seem to converge to the same value fairly quickly, however, β_2 convergence does not seem to be apparent which is consistent with the results in previous studies.

The following sections summarize the simulation results for the various set-ups we mentioned earlier. For each sample of data generated, we obtain MCMC estimates by calculating the mean of the conditional posterior densities for the simulated samples, RMSE, and Bias. We also include the estimates from Heckman's method and from OLS using 'all' data and 'subset' data.

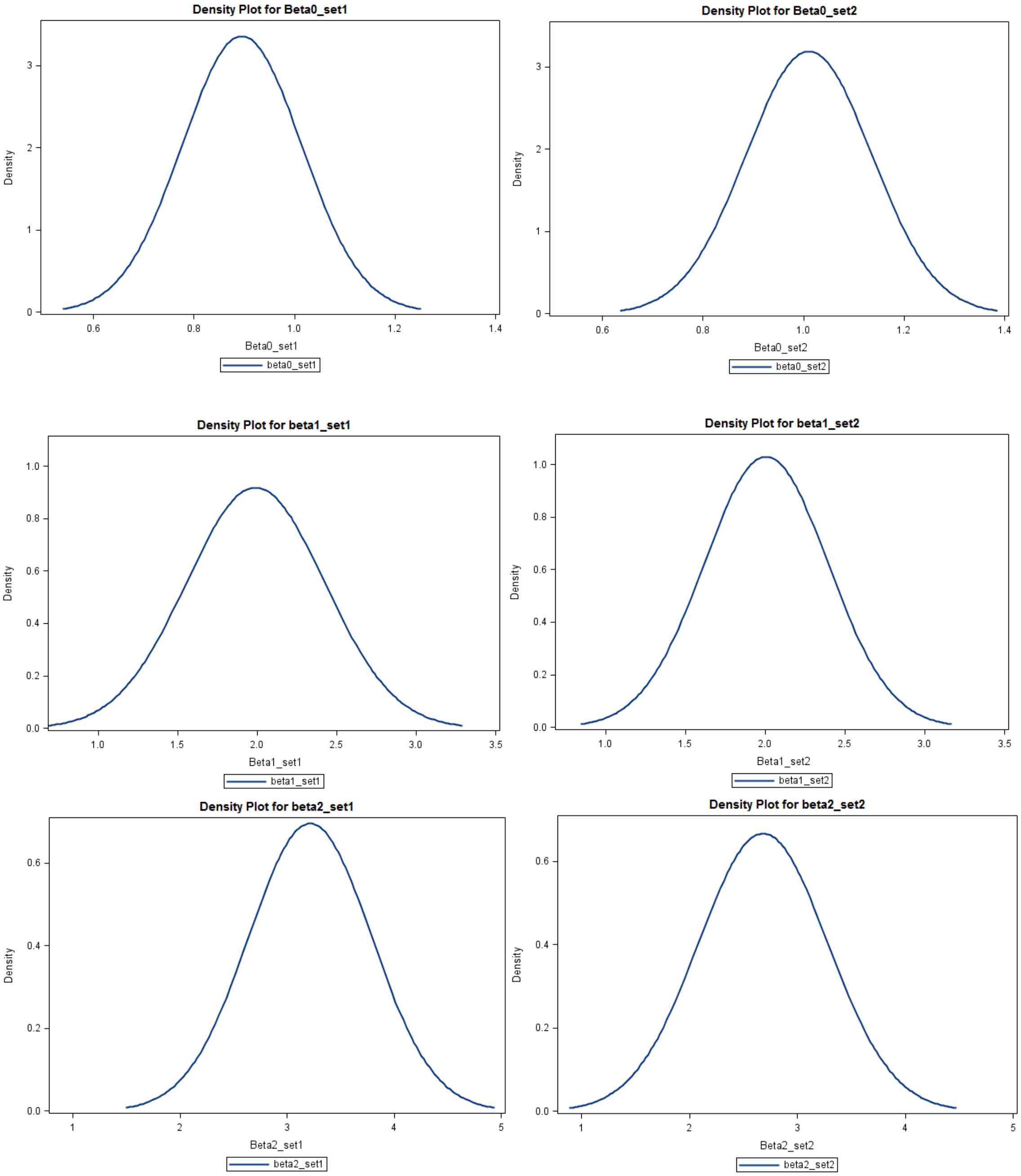


Figure 2. *Density Plots for Comprehensive Simulation*

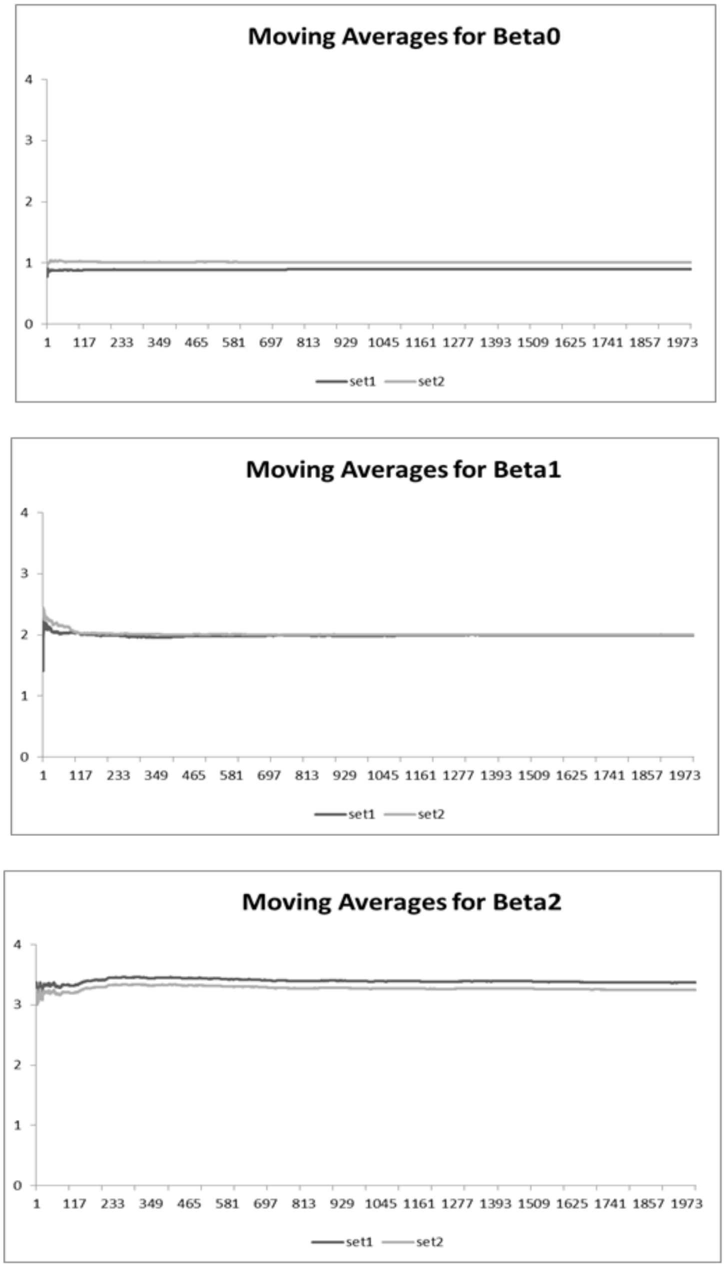


Figure 3. *Moving Averages for Comprehensive Simulation*

Selection Rate

Table 4 shows the simulation results for the regression coefficients using different 'selection rate' scenarios. To test the effect of the missing rate level, we use the following two levels: 50% and 20%. In both scenarios we notice reduction in RMSE and in bias in most coefficients when using a Bayesian approach. By comparing the Bayesian method to Heckman's method we can see 64% (0.14 versus 0.05) reduction in bias for β_0 in the first scenario and 93% reduction in Bias in the second scenario. This shows significant improvement in β_0 estimates using Bayesian methods comparing to Heckman's method when the missing rate is low. We also notice 20% reduction in RMSE in the second scenario for both β_1 and β_2. The RMSE for β_1 in Bayesian is 0.72 versus 0.83 in Heckman. There is also a slight reduction in RMSE for β_2 in the second scenario - RMSE of 0.79 in Bayesian versus 0.82 in Heckman. The results of both scenarios indicate that the Bayesian approach is performing as well as the Heckman approach or even better.

Table 4

Simulation Results for Selection Rate

Selection Rate	Method	Parameter	Mean	RMSE	Bias
50%	OLS – all data	β0	1.009	0.305	0.009
50%	OLS – subset	β0	1.291	0.532	0.291
50%	Heckman	β0	1.142	0.482	0.142
50%	Bayesian	β0	1.051	0.418	0.051
50%	OLS – all data	β1	2.016	0.405	0.016
50%	OLS – subset	β1	2.083	0.569	0.083
50%	Heckman	β1	2.009	0.584	0.009
50%	Bayesian	β1	1.977	0.526	-0.023
50%	OLS – all data	β2	2.969	0.358	-0.031
50%	OLS – subset	β2	2.922	0.542	-0.078
50%	Heckman	β2	2.940	0.546	-0.060
50%	Bayesian	β2	2.939	0.511	-0.061
20%	OLS – all data	β0	1.027	0.289	0.027
20%	OLS – subset	β0	1.519	0.705	0.519
20%	Heckman	β0	1.293	0.603	0.293
20%	Bayesian	β0	1.019	0.606	0.019
20%	OLS – all data	β1	1.990	0.365	-0.010
20%	OLS – subset	β1	1.992	0.778	-0.008
20%	Heckman	β1	1.981	0.834	-0.019
20%	Bayesian	β1	2.031	0.717	0.031
20%	OLS – all data	β2	2.939	0.363	-0.061
20%	OLS – subset	β2	3.038	0.820	0.038
20%	Heckman	β2	3.005	0.820	0.005
20%	Bayesian	β2	3.043	0.787	0.043

Note. Sample size = 80, correlation level = 0.50, and multicollinearity COR(X12, X21) < 0.1.

Correlation Level

Table 5 shows the simulation results for the regression coefficients using different correlation levels. The correlation level is controlled by the value of correlation between the error terms $cor(\varepsilon_i, u_i)$. To check the Bayesian approach's performance, we assign the following three correlation values: 0.3, 0.5, and 0.75. The overall results show that the Bayesian approach provides a significant reduction in RMSE and in bias when correlation level is high. We see reduction in bias in the first two scenarios (0.3 and 0.5) and less reduction for β_1 and β_2 bias when the correlation level is high ($cor(\varepsilon_i, u_i)$ =0.75). The results for the lowest correlation level ($cor(\varepsilon_i, u_i)$ =0.3) show slight reduction in Bias for all coefficients with no significant improvement in RMSE. However, significant reduction in RMSE seems to exist in the high correlation scenarios.

Table 5

Simulation Results for Correlation Level

Correlation	Method	Parameter	Mean	RMSE	Bias
0.30	OLS – all data	β0	1.007	0.339	0.007
0.30	OLS – subset	β0	1.433	0.888	0.433
0.30	Heckman	β0	1.144	0.906	0.144
0.30	Bayesian	β0	1.129	0.935	0.129
0.30	OLS – all data	β1	1.991	0.441	-0.009
0.30	OLS – subset	β1	1.864	0.963	-0.136
0.30	Heckman	β1	1.860	0.998	-0.140
0.30	Bayesian	β1	1.882	0.970	-0.118
0.30	OLS – all data	β2	2.984	0.403	-0.016
0.30	OLS – subset	β2	2.922	0.965	-0.078
0.30	Heckman	β2	3.119	1.038	0.119
0.30	Bayesian	β2	3.013	1.002	0.013
0.50	OLS – all data	β0	1.027	0.289	0.027
0.50	OLS – subset	β0	1.519	0.705	0.519
0.50	Heckman	β0	1.293	0.603	0.293
0.50	Bayesian	β0	1.019	0.606	0.019
0.50	OLS – all data	β1	1.990	0.365	-0.010
0.50	OLS – subset	β1	1.992	0.778	-0.008
0.50	Heckman	β1	1.981	0.834	-0.019
0.50	Bayesian	β1	2.031	0.717	0.031
0.50	OLS – all data	β2	2.939	0.363	-0.061
0.50	OLS – subset	β2	3.038	0.820	0.038
0.50	Heckman	β2	3.005	0.820	0.005
0.50	Bayesian	β2	3.043	0.787	0.043
0.75	OLS – all data	β0	1.027	0.289	0.027
0.75	OLS – subset	β0	1.519	0.705	0.519
0.75	Heckman	β0	1.293	0.603	0.293
0.75	Bayesian	β0	1.019	0.606	0.019
0.75	OLS – all data	β1	1.990	0.365	-0.010
0.75	OLS – subset	β1	1.992	0.778	-0.008
0.75	Heckman	β1	1.981	0.834	-0.019
0.75	Bayesian	β1	2.031	0.717	0.031
0.75	OLS – all data	β2	2.939	0.363	-0.061
0.75	OLS – subset	β2	3.038	0.820	0.038
0.75	Heckman	β2	3.005	0.820	0.005
0.75	Bayesian	β2	3.043	0.787	0.043

Note. Sample size = 80, selection rate = 20%, and multicollinearity COR(X12, X21) < 0.1.

Multicollinearity

The results in Table 6 show the simulation results for the regression coefficients using three different multicollinearity levels. This level is controlled by the level of correlation between two explanatory variables in observation and screening stages [$cor(x_{12}, x_{21})$]. To test the effect of multicollinearity, we use the following levels: 0.06, 0.67, and 1.00. The third scenario [$cor(x_{12}, x_{21})$ =1] refers to the case where the screening stage and the observation stage contain one common explanatory variable. The Bayesian method seems to provide best results for all coefficients when the multicollinearity level is high. The third case show large decrease in RMSE and bias with a Bayesian approach comparing to Heckman. The Bias for β_0 dropped 63% and the RMSE for β_1 dropped 25% in Bayesian estimation when comparing to Heckman method. There seems to be no significant improvement with the Bayesian approach in the first scenario where the multicollinearity level is very low. In this case, Heckman's approach seems to be a good choice for estimation but definitely not when a high level of multicollinearity exists.

Table 6

Simulation Results for Multicollinearity

Multicollinearity	Method	Parameter	Mean	RMSE	Bias
0.06	OLS – all data	β0	1.027	0.289	0.027
0.06	OLS – subset	β0	1.519	0.705	0.519
0.06	Heckman	β0	1.293	0.603	0.293
0.06	Bayesian	β0	1.019	0.606	0.019
0.06	OLS – all data	β1	1.990	0.365	-0.010
0.06	OLS – subset	β1	1.992	0.778	-0.008
0.06	Heckman	β1	1.981	0.834	-0.019
0.06	Bayesian	β1	2.031	0.717	0.031
0.06	OLS – all data	β2	2.939	0.363	-0.061
0.06	OLS – subset	β2	3.038	0.820	0.038
0.06	Heckman	β2	3.005	0.820	0.005
0.06	Bayesian	β2	3.043	0.787	0.043
0.67	OLS – all data	β0	1.011	0.288	0.011
0.67	OLS – subset	β0	1.728	0.982	0.728
0.67	Heckman	β0	1.452	0.837	0.452
0.67	Bayesian	β0	1.119	0.777	0.119
0.67	OLS – all data	β1	2.004	0.353	0.004
0.67	OLS – subset	β1	1.708	0.914	-0.292
0.67	Heckman	β1	1.993	0.931	-0.007
0.67	Bayesian	β1	2.021	0.902	0.021
0.67	OLS – all data	β2	2.928	0.364	-0.072
0.67	OLS – subset	β2	2.870	0.866	-0.130
0.67	Heckman	β2	2.789	0.918	-0.211
0.67	Bayesian	β2	2.825	0.889	-0.175
1	OLS – all data	β0	1.014	0.343	0.014
1	OLS – subset	β0	1.952	1.317	0.952
1	Heckman	β0	1.566	1.592	0.566
1	Bayesian	β0	1.213	1.283	0.213
1	OLS – all data	β1	1.984	0.429	-0.016
1	OLS – subset	β1	1.470	1.179	-0.530
1	Heckman	β1	1.862	1.610	-0.138
1	Bayesian	β1	1.916	1.205	-0.084
1	OLS – all data	β2	3.006	0.385	0.006
1	OLS – subset	β2	2.967	0.790	-0.033
1	Heckman	β2	2.958	0.815	-0.042
1	Bayesian	β2	3.008	0.769	0.008

Note. Sample size = 80, selection rate = 20%, and correlation level = 0.50.

Sample Size

Finally, Table 7, shows the simulation results for the regression coefficients using three different sample sizes (N=80, 120, 160). This is another situation where the results of both scenarios indicate that the Bayesian approach is performing as well as the Heckman approach or even better. The results for various sample size show similar slight reduction of less than 10% in RMSE average for Bayesian method comparing to Heckman for all coefficients. However, we see significant improvement in Bias for Bayesian method for β_0. The Bias reduction in Bayesian method comparing to Heckman for β_0 exceeds 50% in all scenarios.

This comprehensive simulation study used various scenarios that a researcher could face when dealing with a real data. The results proved the effectiveness of the Bayesian approach and showed the limitations of Heckman's approach, particularly when faced with a high level of multicollinearity. A detailed investigation of the multicollinearity issue can be found in Leung and Yu (1996). They show that the degree of multicollinearity is the main decision driver to judge the appropriateness of the LIML and FIML estimates in relation to the two-part model. In empirical analysis, in wage equations for example, the standard procedure to solve the multicollinearity problem, is to find variables that determine the probability to work (selection equation), but not the wage rate (observation equation) directly. Practical examples for these variables could be the income of the spouse, household income, etc. However, these variables are not always available in practical situations.

Table 7

Simulation Results for Sample Size

Sample Size	Method	Parameter	Mean	RMSE	Bias
N=80	OLS – all data	β0	1.027	0.289	0.027
N=80	OLS – subset	β0	1.519	0.705	0.519
N=80	Heckman	β0	1.293	0.603	0.293
N=80	Bayesian	β0	1.019	0.606	0.019
N=80	OLS – all data	β1	1.990	0.365	-0.010
N=80	OLS – subset	β1	1.992	0.778	-0.008
N=80	Heckman	β1	1.981	0.834	-0.019
N=80	Bayesian	β1	2.031	0.717	0.031
N=80	OLS – all data	β2	2.939	0.363	-0.061
N=80	OLS – subset	β2	3.038	0.820	0.038
N=80	Heckman	β2	3.005	0.820	0.005
N=80	Bayesian	β2	3.043	0.787	0.043
N=120	OLS – all data	β0	0.996	0.257	-0.004
N=120	OLS – subset	β0	1.672	0.886	0.672
N=120	Heckman	β0	1.396	0.745	0.396
N=120	Bayesian	β0	1.113	0.629	0.113
N=120	OLS – all data	β1	1.988	0.288	-0.012
N=120	OLS – subset	β1	1.884	0.681	-0.116
N=120	Heckman	β1	1.912	0.663	-0.088
N=120	Bayesian	β1	1.895	0.630	-0.105
N=120	OLS – all data	β2	3.037	0.354	0.037
N=120	OLS – subset	β2	2.883	0.723	-0.117
N=120	Heckman	β2	2.965	0.762	-0.035
N=120	Bayesian	β2	2.958	0.687	-0.042
N=160	OLS – all data	β0	0.977	0.197	-0.023
N=160	OLS – subset	β0	1.550	0.741	0.550
N=160	Heckman	β0	1.325	0.602	0.325
N=160	Bayesian	β0	0.981	0.530	-0.019
N=160	OLS – all data	β1	2.043	0.281	0.043
N=160	OLS – subset	β1	1.950	0.628	-0.050
N=160	Heckman	β1	2.000	0.636	0.000
N=160	Bayesian	β1	2.056	0.604	0.056
N=160	OLS – all data	β2	3.015	0.277	0.015
N=160	OLS – subset	β2	3.026	0.0543	0.026
N=160	Heckman	β2	3.041	0.550	0.041
N=160	Bayesian	β2	3.004	0.581	0.004

Note. Selection rate = 20%, selection level = 0.50, and multicollinearity COR(X12,X21) < 0.1.

CHAPTER 5

CASE STUDY: PLACEMENT EXAM AND MATH ACHIEVEMENT

In this chapter we apply the Bayesian model to a real-world data example by using AU students' placement exam data. The goal is to investigate how the students' placement exam scores are associated with their first year math achievement. All AU students are supposed to take the math placement exam and register for appropriate math/stat courses accordingly. The problem is that information about the students first year math achievements (grades) is only available for those who actually take and complete their first year math courses. We wish to forecast outcomes in the whole pool of freshmen but are forced to rely on a subset chosen non-randomly.

The data contain 1,012 freshmen students with 4 variables. A brief description of the variables that are relevant for our analysis is shown in Table 8. From among the 1,012 students, we observe students grades for only 752. The remaining 260 students did not register or complete a math class in fall 2010 and so did not receive a grade. We are interested to see how placement exam score are related to the student's math grade. Due to the self-selection, we need to perform accurate estimation by correcting for sample selection bias. The student grade is a function of his placement score and the recommended class. A dummy variable called 'Basic Level' was created to determine the type of recommended math class (e.g. Basic Algebra, Applied Calculus, etc.). If the recommended class type is classified as basic, the new dummy variable value is equal to 1; otherwise, the value is equal to zero. This new dummy variable was included in the observation stage as an explanatory variable.

Table 8

Variable Descriptions for Placement Exam Data

Variable Name	Definition
Placement	Recommended class based on placement exam score
Score	Student’s placement exam score
Basic Level	Dummy variable equal to 1 if the recommended class type is classified as a basic level; otherwise, the value is equal to zero
Grade	Student’s grade for the class

We begin with OLS estimation of the regression model using only the observations that have grade data. The estimates can be found in Table 9 in `OLS- subset' row. The estimated coefficient shows a very small effect of placement exam score on the student’s math achievement (β_1 =0.0326). This analysis would be fine if, in fact, the missing grade data were missing completely at random. However, the decision to register for a math class or not was made by the individual student. Thus, those who were not registered constitute a self-selected sample and not a random sample. It is likely that some of the students who had low placement scores chose not to register for any math class. If so, this would account for much of the missing grade data. Thus, it is likely that we will over-estimate the grade of the student in the population.

On the other hand, the Heckman’s method is supposed to allow us to use information from non-registered students to improve the estimates of the parameters in the regression model. However, Heckman’s method shows negative estimates and low negative effect of placement exam score on the student's math achievement (β_1=-0.0312).

Table 9

Case Study Results for Placement Exam Data

	Intercept (β0)		Score (β1)		Basic Level (β2)	
	Mean	Std. Error	Mean	Std. Error	Mean	Std. Error
OLS – subset	2.6557	0.1887	0.0326	0.0008	-0.1025	0.1109
Heckman	-0.6127	11.2704	-0.0312	0.2206	-0.1196	0.1256
Bayesian	1.0768	4.9429	0.0130	0.2432	-0.1038	0.4510

Finally, we consider the Bayesian approach using MCMC methods and Gibbs sampling, where the posterior distribution can be sampled indirectly by generating a sample of parameter values from the conditional distribution of interest. Posterior Bayes estimates are then obtained from the generated samples. In the 20,000 samples the first 1,000 are discarded, and we use thinning of 2, leaving 9,000 effective posterior samples.

Figure 4 shows the sample path for $\beta_0, \beta_1, \beta_2$ using two different starting points. The plots show that all values converge to the same distribution. Figure 5 shows the posterior density plot for $\beta_0, \beta_1, \beta_2$ and the two settings show normal distributions. Figure 6 shows the moving averages for $\beta_0, \beta_1, \beta_2$ from the two MCMC chains. The plots indicate that the two MCMC chains are converging to the same value for β_0 and β_1. The Moving Averages for β_2 (Basic Level) from the two settings seem not to converge to the same value. The Bayesian estimate for β_2 in Table 9 seem to support this result. This indicates low effect of Basic Level on student's math achievement.

Table 9 shows that the estimated effect of placement score on student's grade is positive when applying the Bayesian model. Such relationship of placement score is not identified in estimates using Heckman's method. However, Basic Level seems to have less effect in the Bayesian method compared with Heckman's method. There seems to be no improvement in the standard errors using the Bayesian approach for all coefficients. The Bayesian standard errors seem to be larger than those of the other two estimation methods (Heckman and OLS).

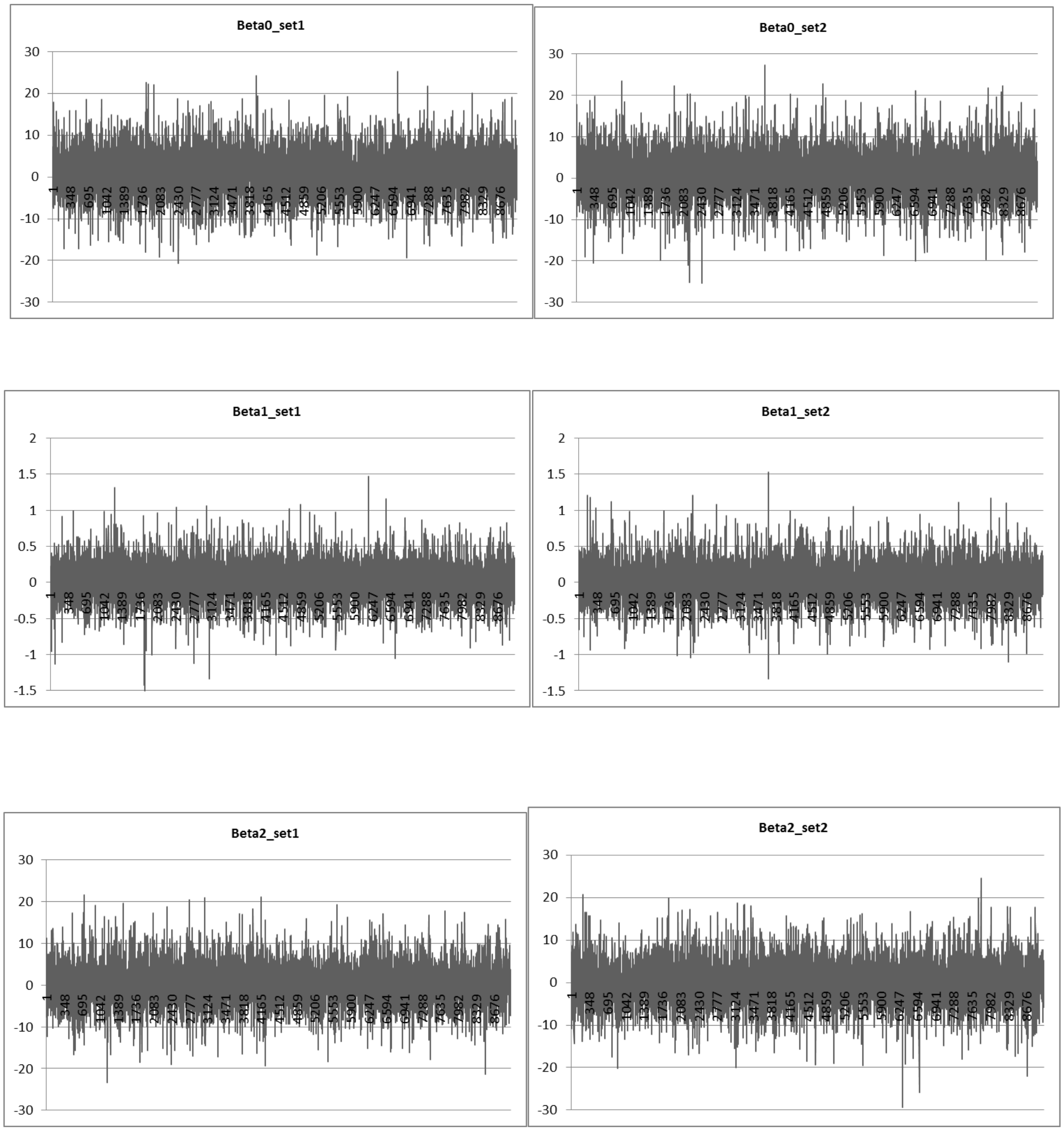


Figure 4. *Sample Paths for Placement Exam Example*

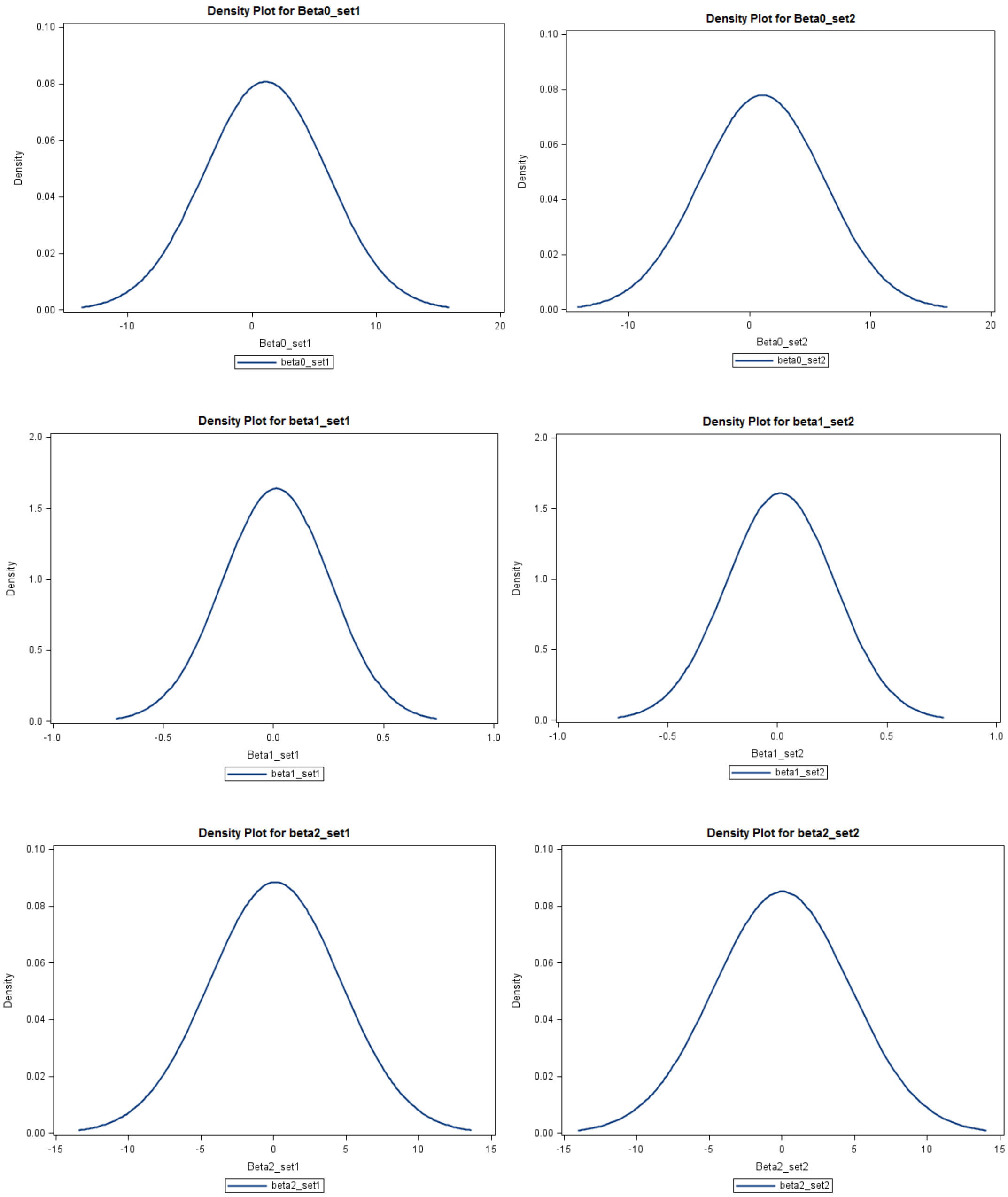


Figure 5. *Density Plots for Placement Exam Example*

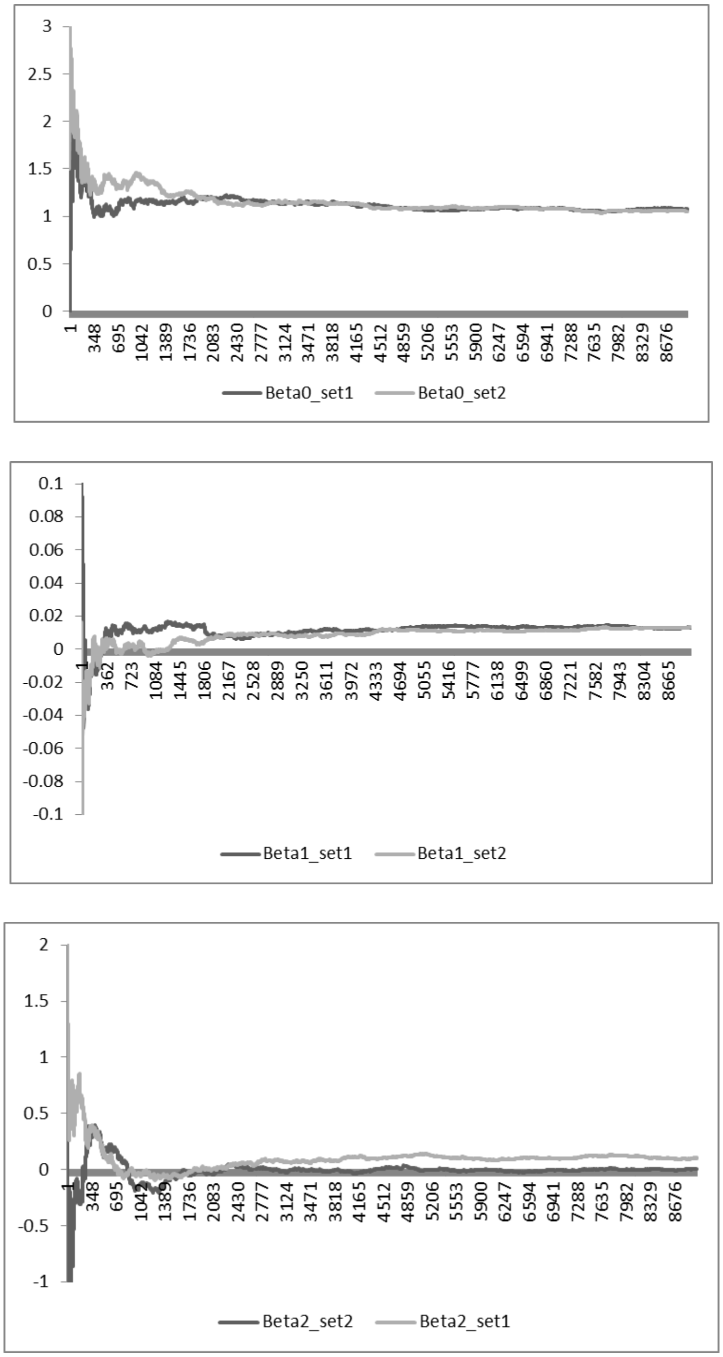


Figure 6. *Moving Averages for Placement Exam Example*

CHAPTER 6

BINARY SELECTIVITY MODEL

In this chapter, the proposed Bayesian method is extended to the Generalized Linear Model (GLM). The GLM extends the linear regression model in order to accommodate non-normal responses (e.g. binomial data, frequency data, etc.) to linear equation via a link function. Examples of GLM include well-known models such as logistic regression and log-linear models (Poisson regression) for frequency tables, etc. Conceptually the Bayesian specification is straightforward. We need to assign a prior for regression coefficients, as in the previous regression examples. There is no closed form solution available, but it is simple to obtain samples from posteriors via MCMC.

Generalized Linear Model

Let $y_1,......y_n$ denote n independent observations on a response variable and treat y_i as a realization of a random variable Y_i. In GLM, we assume that y_i that is part of the exponential family with three main components (random, systematic, and link). The random part is the distribution of the observations, the systematic component is the linear combination of explanatory variables, and the link function is the link between the random part and the systematic component. The exponential family is defined as the following:

$$f(y_i;\theta_i,\phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)\right\} \qquad (41)$$

where θ_i and ϕ are location and scale parameters respectively. The mean and variance are:

$$E(y_i) = b^{'}(\theta_i) \text{ and } V(y_i) = b^{''}(\theta_i)a(\phi)$$

and we assume that the expected value μ_i is a linear function of x_i.

$$\eta_i = g(\mu_i) = x^{'}{}_i\beta \qquad (42)$$

where η_i is the linear predictor, $g(.)$ is the link function, x_i are the predictors and β is a vector of unknown parameters (regression coefficients).

Our current model employs a linear regression in the observed stage and a probit/logistic regression in the selection stage. Let us consider $U \sim N(x^{'}\beta,1)$ which follows

$$\theta_i = P(Y = 1 | x_i) = \Phi(x_i^{'}\beta)$$

where

$$\Phi(t) = (2\pi)^{-1/2}\int_{-\infty}^{t}\exp(-\frac{1}{2}z^2)dz$$

The relation is linearized by the inverse normal transformation

$$\Phi^{-1}(\theta) = x_i^{'}\beta = \sum_{j=1}^{p} x_{ij}\beta_j$$

The cutoff value of U is fixed and the mean of U is changing with x.

The goal is to have probit (or GLM) in the observed as in the selection stage. We will obtain this by using two latent variables in each stage, and we will be able to allow correlation between these two variables. This is similar to Heckman's selection model except that now we have a binary outcomes in the observation stage.

Assume the following selection setup:

$$D_i = \begin{cases} 1 & if \quad Z_i > 0 \\ 0 & Otherwise \end{cases}$$

and $Z_i \sim N(x_{2i}\beta_2, 1)$ is a latent variable.

$$Y_i^* = \begin{cases} 1 & if \quad Y_i > 0 \quad and \quad D_i = 1\,(or\ Z_i > 0) \\ 0 & if \quad Y_i < 0 \quad and \quad D_i = 1\,(or\ Z_i > 0) \\ Mis\sin g & if \qquad\qquad D_i = 0\ (or\ Z_i < 0) \end{cases}$$

The probit regression model can expressed as: $P(Y_1^* = 1) = \Phi(x_{1i}\beta_1)$

$Y_i \sim N(x_{1i}\beta, 1)$ is also a latent variable. Therefore, the latent variables can be expressed as the following Bivariate Normal:

$$\begin{pmatrix} Y_i \\ Z_i \end{pmatrix} \sim BVN\left(\begin{pmatrix} x_{1i}\beta_1 \\ x_{2i}\beta_2 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right)$$

Bayesian Estimation

The Bayesian setup for the GLM is an extension of the framework we have used for regression models. Suppose we have $\eta_i = g(\mu_i) = X\beta$, we need to choose a prior density for the parameters $(\beta, \phi), \pi(\beta, \phi)$. The posterior density is then expressed as:

$$\pi(\beta, \phi \mid y) = \frac{f(y; \beta, \phi)\pi(\beta, \phi)}{\int f(y; \beta, \phi)\pi(\beta, \phi)d\beta d\phi} = \frac{f(y; \beta, \phi)\pi(\beta, \phi)}{\pi(y)} \propto f(y; \beta, \phi)\pi(\beta, \phi)$$

where $\pi(y)$ is the marginal likelihood of the data, obtained by integrating the likelihood conditional on the unknown regression coefficient β and dispersion parameter ϕ across the prior density.

The joint posterior density of unobserved β and Z given Y is

$$\pi(\beta, Z, y \mid D_i, y^*) \propto \prod_{i=1}^{N} \{L(D_i \mid Z_i)L(y^* \mid y)\phi(Z_i - x_{i2}\beta_2)\phi(y_i - x_{i1}\beta_1)\}\pi(\beta)$$

Let $\beta = [\beta_1^{'}\beta_2^{'}]$, from the above joint posterior, we now infer conditional posteriors and implement Gibbs sampler. We start with sampling the conditional posterior of Z_i from

$$Y_i, Z_i \mid \beta, D_i, Y^* \sim BVN\left(\begin{pmatrix} x_{1i}\beta_1 \\ x_{2i}\beta_2 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right)$$

therefore,

$$Z_i \mid \beta, Y_i, D_i, Y^* \sim N(\mu_z, 1) \text{ - - Truncated at left by 0 if } D_i = 1$$

$$Z_i \mid \beta, Y_i, D_i, Y^* \sim N(\mu_z, 1) \text{- - Truncated at right by 0 if } D_i = 0$$

where $\mu_z = x_{2i}\beta_1 + (Y_i - x_{2i}\beta_2)$

in a similar way, we calculate the conditional posterior of Y_i. We can get a result for Y_i when $D_i = 1$ and $Y^* = 1$, which also truncated normal.

$$Y_i \mid \beta, Z_i, D_i, Y^* \sim N(\mu_Y, 1) \text{- - Truncated at left by 0 if } D_i = 1 \text{ and } Y^* = 1$$

$$Y_i \mid \beta, Z_i, D_i, Y^* \sim N(\mu_Y, 1) \text{- - Truncated at left by 0 otherwise}$$

where $\mu_Y = x_{1i}\beta_2 + (Z_i - x_{1i}\beta_1)$

To sample β , we use the following prior

$$\pi(\beta \mid y, Z) \propto \pi(\beta) \prod_{i=1}^{N} \phi(Z_i, X_i\beta, 1)$$

let $X = (x_{1i}^{'}, 0, 0, x_{2i}^{'})$ and $W = (Y_i, Z_i)^{'}$ and we can get the conditional posterior function of β which is normal density,

$$\beta \mid Y, Z, \sim N(\beta, B^{-1})$$

where $\beta = B^{-1}(X^{'}\Sigma W)$ and $B = B_0 + X^{'}\Sigma X$

We can sample β and Z iteratively, by drawing β given Z and vice versa. We can sample Z_i and Y_i from the posterior marginal distribution at each iteration. This marginal distribution is conditional only on the data and not on any parameters. And then we can sample β from the same posterior full conditional distribution as the following:

- Sample Z_i and Y_i from its posterior marginal distribution.
- Sample β from the same posterior full conditional distribution as described previously.

Data Example

The previous model is applied to the same data example we used in Chapter 5 (AU Students Replacement Exam Score Data) except that the response variable is now binary. For each student I assigned a new variable called ‘Pass’ that takes the value of 1 if Grade>3.0 (Pass) and otherwise 0 (Fail).

I begin with MLE estimation for the bivariate probit model using only the observations that have grade data. The estimates can be found in this first row in Table 10. As discussed in previous examples, it is likely that we will over-estimate proportion of ‘Passed’ students in the population. Bayesian method is applied after running 10,000 of iterations MCMC using Gibbs Sampler with 1,000 as burn-in. The results of this method can be found in this first row in Table 10.

Table 10

Model Results for Binary Selectivity

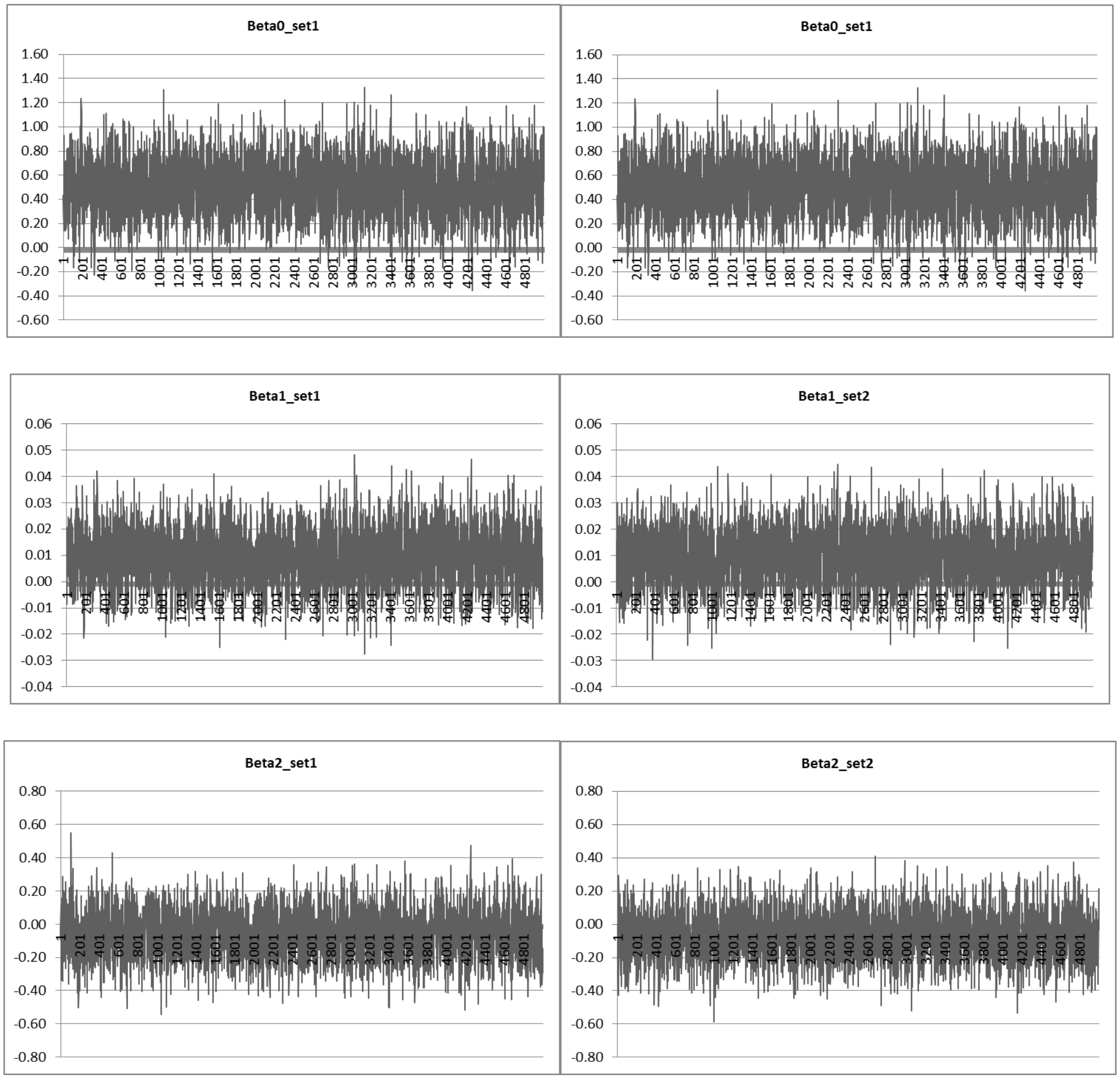
	Intercept		Score		Basic Level	
	Mean	Std. Error	Mean	Std. Error	Mean	Std. Error
MLE	0.8599	0.2555	-0.0112	0.0112	0.0309	0.1541
Bayesian	0.5020	0.2327	0.0095	0.0103	-0.0643	0.1404

Figure 7 shows the sample path for $\beta_0, \beta_1, \beta_2$ using two different starting points. The plots show that all values converge to the same distribution. Figure 8 shows the posterior density plot for $\beta_0, \beta_1, \beta_2$ and the two settings show normal distribution. Figure

9 shows the Moving Averages for $\beta_0, \beta_1, \beta_2$ from the two MCMC chains. The graph indicates that the two MCMC chains are converging to the same value.

The results in Table 10 indicates that the Bayesian approach is performing at least as well as the MLE approach. The Score coefficient seems to indicate positive relationship with Grade. This relationship was reversed with negative coefficient for Score in MLE method. Another advantage in the Bayesian approach seems to be in the slight reduction of the standard error for all coefficients.

In general, we would expect strong positive relationship between the placement exam score and the final outcome whether the student passed or failed math class in the first semester. Similarly, we would expect larger correlation between Basic Level which is based on the student's placement exam score and whether the student passed or failed math class. However, these strong correlations are not present in the data. This is mainly due to the limitation of the available variables that we have to use in our model. The variable 'Score' is used in both stages (selection and observation) which leads to high multicollinearity. These results confirm that the parameters estimates are not very effective when multicollinearity exists in the model.



Beta0_set1

Beta0_set1

Beta1_set1

Beta1_set2

Beta2_set1

Beta2_set2

Figure 7. *Sample Paths for Binary Selectivity*

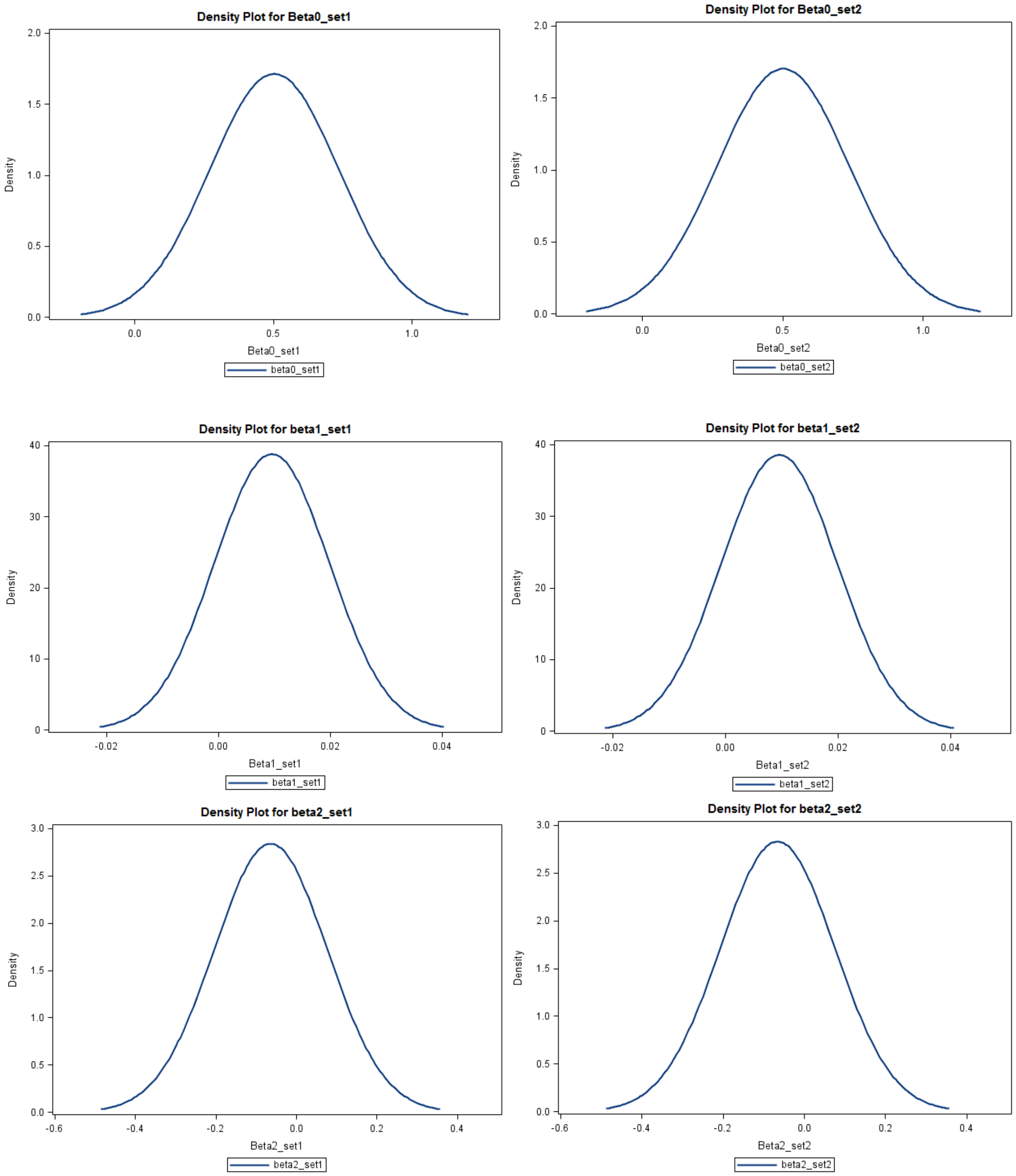


Figure 8. *Density Plots for Binary Selectivity*

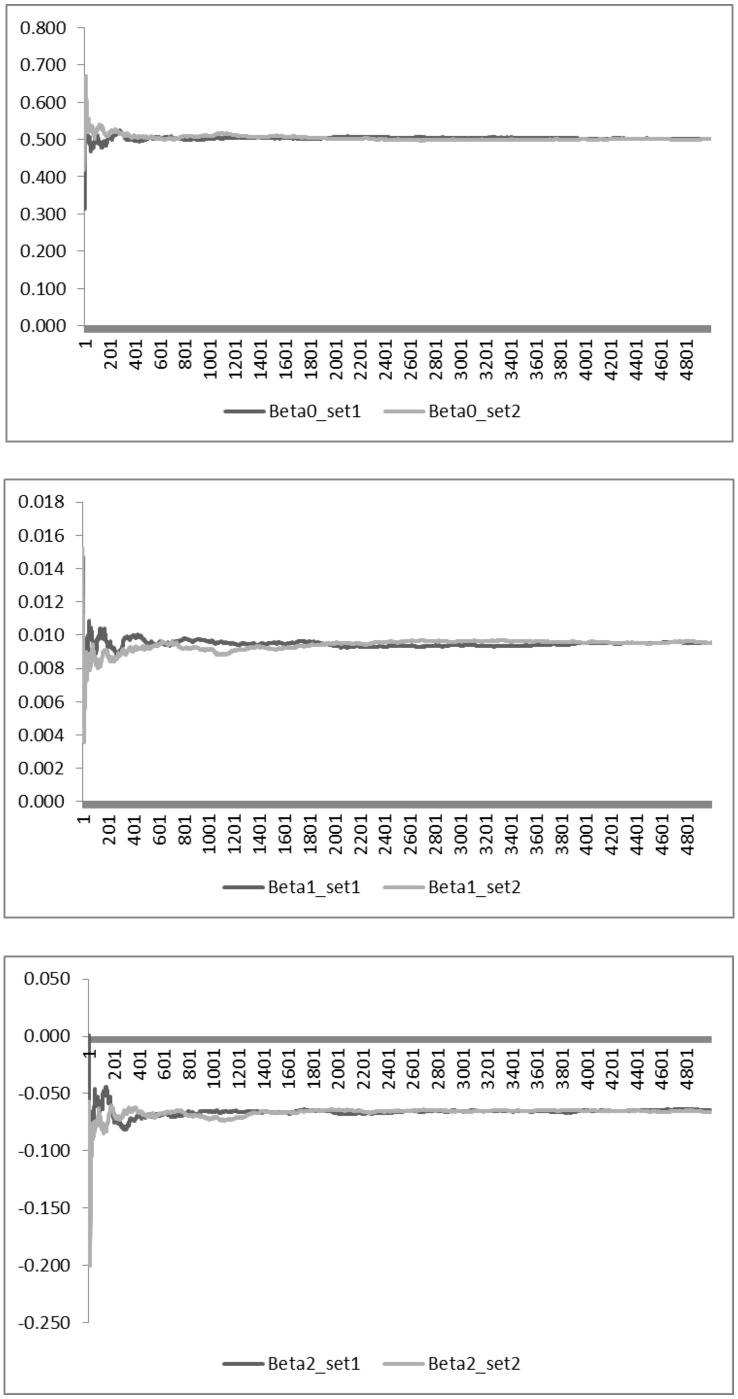


Figure 9. *Moving Averages for Binary Selectivity*

CHAPTER 7

DISCUSSION

This study has provided a Bayesian approach for the sample selection model. We investigate the effect of prior distributions and the robustness of Bayesian approach. A simulation study using generated data and a real data example are conducted using MCMC methods and Gibbs sampling. The results of the MCMC method are compared to OLS estimation and Heckman estimation. The results indicate that the Bayesian approach is performing at least as well as the Heckman approach, and outperforms the Heckman's approach in some scenarios such as when multicollinearity exists.

A comprehensive simulation study is conducted to investigate the Bayesian approach by applying several specific conditions. These conditions reflect various scenarios which the researcher might face when dealing with the problem of self-selection bias. The results proved the effectiveness of the Bayesian estimates and showed the limitation of the Heckman method. The results show that Heckman estimator can perform well when there is no multicollinearity between the Inverse Mills Ratio and the explanatory variables.

In the real data example, a Bayesian approach is applied to measure the effect of placement exam score on students' math achievement. With the spirit of posterior distributions, it is not surprising that Bayesian methods provided improvement in the parameters estimates. OLS and Heckman methods showed very small effect of placement exam score on student’s math achievement with slight effect of Basic Level. However, the Bayesian method is showing larger estimates for the same variable with less effect of Basic Level.

Finally, a Bayesian approach with Binary selection model is provided. This approach is applied to our real data example using MCMC and Gibbs sampling. The results of both MCMC method and MLE estimation are compared. It shows that both outcomes are so close that the Bayesian approach is as reliable as MLE.

This Bayesian approach can be extended to handle other models with sample selectivity problem such as multilevel models. Multilevel data are structures that consist of multiple units of analysis, one nested within the other. These models are used frequently in political science where clustering or multilevel such as various groups of people (e.g. gender, ethnic background) is important for data analysis purposes.

Furthermore, this approach can be extended to handle multinomial and ordinal probit models as well. In these models, the latent variables are divided to multiple intervals. The extension to a response consisting of a mixture of binary and continuous data could be interesting and useful in many applications.

REFERENCES

Albert, J. H. and Chib, S. (1993), “Bayesian Analysis of Binary and Polychotomos Response Data,” 88(422), 669-679

Anderson T. and Rubin H. (1949), Estimation of the parameters of a single equation in a complete system of stochastic equations, The Annals of Mathematical Statistics, 20, 46-63.

Bernd A. Berg (2004). "Markov Chain Monte Carlo Simulations and Their Statistical Analysis". Singapore, World Scientific

Browne W. and Draper J. (2000) ‘Implementation and performance issues in the Bayesian and likelihood fitting of multilevel models. Computational Statistics, 15, 391-420

Casella, G. and Edward I. (1992). Explaining the Gibbs sampler. The American Statistician 46 (3): 167–174

Draper, D. (2000). Bayesian Hierarchical Modeling. Draft version can be found on the web at http://www.bath.ac.uk/»masdd/

Duan N., Manning W., Morris C., and Newhouse J. (1984) Choosing between the sample-selection model and the multi-part model, Journal of Business & Economic Statistics, 2, 3, 283-289.

Geman S. (1984). "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images". IEEE Transactions on Pattern Analysis and Machine Intelligence, 6:721–741,

Gilks W.R., Richardson S. and Spiegelhalter D.J. (1996) Strategies for improving MCM, in Markov Chain Monte Carlo in Practice, Chapman & Hall/CRC, London, 89-114.

Hastings, W.K. (1970). Monte Carlo Sampling Methods Using Markov Chains and Their Applications. Biometrika 57 (1): 97–109. doi:10.1093/biomet/57.1.97

Heckkman 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 Social Measurement, 5, 4, 475-492.

Heckman J. (1979) Sample selection bias as a specification error, Econometrica, 47, 153-161.

Lee, P. (1997). Bayesian Statistics: An introduction, 2nd Ed. John WIley, New York.

Lee Y. and Nelder J. (1996) Hierarchical generalized linear models (with discussion). Journal of the Royal Statistical Society (Series B) 58: 619-678

Leung S. and Yu S. (1997) On the choice between sample selection and two-part models, Journal of Economics, 72, 197-229.

Maddala G. (1977) Limited dependent variable models using panel data, Journal of Human Resources, 22, 307–338.

Manski C. and Wise D. (1983) College Choice in America, Harvard University Press, Cambridge, MA

Melino A. (1982) Testing for sample selection bias, Review of Economic Studies, 49, 151-153.

Metropolis, N., A.W. Rosenbluth, M. N. Rosenbluth, A.Teller, and H. Teller. (1953). Equations of state calculations by fast computing machines. Journal of ChemicalPhysics 21: 1087–1091.

Mroz T. (1987) The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions, Econometrica, 55, 765-799.

N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller (1953). "Equation of State Calculations by Fast Computing Machines". Journal of Chemical Physics 21 (6): 1087–1092.

Nawata K. (1993) A note on the estimation of models with sample selection biases, Economis Letters, 42, 15-24.

Nawata K. (1994) Estimation of sample selection bias models by the maximum likelihood estimator and Heckman's two-step estimator, Economics Letters, 45, 33-40.

Pahani P. (2000) The Heckman correction for sample selection and its critique, Journal of Economic Surveys, Vol 14, No 1

Smith, A. F. M., and G. O. Roberts. 1993. Bayesian computation via the Gibbs sampler and related Markov chain Monte-Carlo methods (with discussion). J. Roy. Stat. Soc. Series B 55: 3-23.

Stolzenberg M. and Relles D. (1997) Tools for intuition about sample selection bias and its correction, American Sociology Review, 62, 494-507

Vella F. (1998) Estimating models with sample selection bias. The Journal of Human Resources. 33(1), 127-169

Winship C. and Mare R (1992) Models for sample bias, Annual Review of Sociology, 18, 327-350