2010

Essays on selection in health survey data

Maksym Obrizan
University of Iowa

ESSAYS ON SELECTION IN HEALTH SURVEY DATA

by

Maksym Obrizan

An Abstract

Of a thesis submitted in partial fulfillment of the
requirements for the Doctor of Philosophy
degree in Economics
in the Graduate College of
The University of Iowa

May 2010

Thesis Supervisor: Professor John Geweke

ABSTRACT

In this dissertation I examine the effects of sample selection on the probability of stroke among older adults. If study subjects are selected into the sample based on some non-experimental selection process, then statistical analysis may produce inconsistent estimates.

Chapter 1 develops a model of non-ignorable selection for a discrete outcome variable, such as whether stroke occurred or not. I start by noticing that in the literature there are relatively few applications of the Heckman model to the case of a discrete outcome variable and they are limited to a bivariate case. After that I extend the Bayesian multivariate probit model of Chib and Greenberg (1998) broadly following the logic of Heckman's original (1979) work. The model in the first chapter of my dissertation is set in a way general enough to handle multiple selection and discrete-continuous outcome equations.

The first extension of the multivariate probit model in Chib and Greenberg (1998) allows some of the outcomes to be missing. In particular, stroke occurrence is missing whenever the person is not selected into the sample. In terms of latent variable representation this implies that multivariate normal distribution is not truncated in the direction of missing outcome. I also use Cholesky factorization of the variance matrix to avoid the Metropolis-Hastings algorithm in the Gibbs sampler.

Chapter 2 evaluates how severe the problem of sample selection is in Assets and HEAlth Dynamics among the Oldest Old (AHEAD) data set. I start with a more restrictive assumption of ignorable selection. In particular, I apply the propensity

score method as in a recent paper by Wolinsky *et al.* (2009) and find no selection effects in the study of stroke. Then I consider the model developed in Chapter 1, which is based on a less restrictive assumption of non-ignorable selection, and also find no evidence of selection. Thus, the main substantive contribution of this chapter is the absence of selection effects based on either ignorable or non-ignorable sample selection model.

Abstract Approved: _____

Thesis Supervisor

_____

Title and Department

_____

Date

ESSAYS ON SELECTION IN HEALTH SURVEY DATA

by

Maksym Obrizan

A thesis submitted in partial fulfillment of the
requirements for the Doctor of Philosophy
degree in Economics
in the Graduate College of
The University of Iowa

May 2010

Thesis Supervisor: Professor John Geweke

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

PH.D. THESIS

This is to certify that the Ph.D. thesis of

Maksym Obrizan

has been approved by the Examining Committee for the thesis requirement for the Doctor of Philosophy degree in Economics at the May 2010 graduation.

Thesis Committee: _____
John Geweke, Thesis Supervisor

_____
Gustavo Ventura

_____
Forrest Nelson

_____
Gene Savin

_____
George Wehby

ACKNOWLEDGEMENTS

I was very lucky to be raised in a truly Ukrainian family and I will always be grateful to my Parents and my brother for the many happy years we enjoyed together. This thesis is just a small token of gratitude that I can give to my family.

I would like to thank all the people in Iowa who made this thesis happen. Kirill and Michelle Nourski supported me enormously during my final months of work on the dissertation. Our graduate coordinator Renea Jay and my roommate Latchezar Popov deserve special thanks. Most of my understanding of the AHEAD data set was acquired from fellow research assistants Suzanne Bentler, Li Liu and Beth Cook.

Financial support from Professor Fredric Wolinsky through NIH grant R01 AG-022913 in the last two years is kindly acknowledged. Professor John Geweke helped on many derivations in this thesis in addition to providing financial support in the spring semester of 2009.

ABSTRACT

In this dissertation I examine the effects of sample selection on the probability of stroke among older adults. If study subjects are selected into the sample based on some non-experimental selection process, then statistical analysis may produce inconsistent estimates.

Chapter 1 develops a model of non-ignorable selection for a discrete outcome variable, such as whether stroke occurred or not. I start by noticing that in the literature there are relatively few applications of the Heckman model to the case of a discrete outcome variable and they are limited to a bivariate case. After that I extend the Bayesian multivariate probit model of Chib and Greenberg (1998) broadly following the logic of Heckman's original (1979) work. The model in the first chapter of my dissertation is set in a way general enough to handle multiple selection and discrete-continuous outcome equations.

The first extension of the multivariate probit model in Chib and Greenberg (1998) allows some of the outcomes to be missing. In particular, stroke occurrence is missing whenever the person is not selected into the sample. In terms of latent variable representation this implies that multivariate normal distribution is not truncated in the direction of missing outcome. I also use Cholesky factorization of the variance matrix to avoid the Metropolis-Hastings algorithm in the Gibbs sampler.

Chapter 2 evaluates how severe the problem of sample selection is in Assets and HEAlth Dynamics among the Oldest Old (AHEAD) data set. I start with a more restrictive assumption of ignorable selection. In particular, I apply the propensity

score method as in a recent paper by Wolinsky *et al.* (2009) and find no selection effects in the study of stroke. Then I consider the model developed in Chapter 1, which is based on a less restrictive assumption of non-ignorable selection, and also find no evidence of selection. Thus, the main substantive contribution of this chapter is the absence of selection effects based on either ignorable or non-ignorable sample selection model.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

CHAPTER 1
MULTIVARIATE PROBIT AND SAMPLE SELECTION

1.1 Introduction

In this chapter I develop a model of *sample selection* with multiple outcome and selection equations in which dependent variables are dichotomous. As an illustration of the sample selection model, consider a sample of elderly Medicare-eligible Americans. Suppose that some of the respondents have allowed the researcher to get access to their Medicare claims data. This constitutes the selection equation which is usually modeled as univariate probit. Suppose further, that the researcher is concerned about estimating amount of Medicare spending per year, which is the outcome equation of interest, but she observes only the amounts for patients that allowed access to their Medicare claims. Economists have been aware for a long time that estimating such a model by ordinary least squares leads to inconsistent estimates.[1] Gronau (1974) seems to be among the first to recognize this problem, but Heckman (1979) offers a truly pioneering work with a simple two-step estimator that has been widely used for more than three decades.

In general, the model of sample selection, also referred to as a *model with incidental truncation*, has a dependent variable that is missing as a result of a non-experimental selection process.[2] Heckman (1979) recognizes the sample selection

[1]There is no such problem if the disturbances in two equations have zero correlation.

[2]Models of sample selection in some classifications also include models with truncation and censoring. Throughout this chapter I use the terms "sample selection" and "incidental truncation" interchangeably to refer to a Heckman-type model.

problem as specification error and offers the following two-step estimator. On the first step, the binary selection equation is estimated by probit. (Exogenous variables are assumed to be known for all respondents in the sample.) On the second step, the outcome equation is estimated only for the observed subsample with one additional variable: inverse Mills ratio obtained on the first step. Heckman (1979) shows that this procedure results in consistent coefficient estimates and also provides a corrected variance matrix for hypothesis testing.

Sample selection remains an active and ongoing research area in literature and recent textbook presentations (such as Greene 2003 and Wooldridge 2002), as well as review articles (Vella 1998 and Lee 2003) are available. Most of the research, however, is limited to the case where the endogenous variable of interest in the outcome equation is continuous. In addition, the majority of papers deal with a single selection and a single outcome equation in the sample selection model. In many applications, however, sample may be chosen based on more than one criterion, or more than one outcome equations may be considered.

This chapter substantially extends Heckman's (1979) classic model by adding two additional features. First of all, it allows binary dependent variable in the outcome equation as well.[3] Continuing with health economics, it might be of interest to model risk factors of a certain morbid event (such as hip fracture). Secondly, adding extra selection or outcome equations with dichotomous or continuous dependent vari-

[3]I review some earlier work on a discrete outcome variable in Heckman's (1979) model and explain how my model differs further on.

ables is straightforward. This extension is crucial in the presence of multiple selection equations, as explained below. These two extensions seem to be an important contribution to the existing literature with potential applications in health, labor and related empirical economic research.

While technical issues limited the use of sample selection models with multiple binary dependent variables, their applicability is potentially very wide. To continue with the Medicare-eligible sample of elderly Americans, suppose that the researcher is interested in joint estimation of two or more binary morbid health events (for example, hip fracture and stroke) but she observes those outcomes only for respondents that gave her access to Medicare claims. Clearly, joint estimation of the two health events (outcome variables of interest) with a third equation for being in the analytic sample (selection equation) tends to be more efficient than estimating them equation-by-equation.[4] More importantly, in order to obtain consistent estimates *all* of the selection equations have to be included.

Consider another example from financial economics. Suppose that a credit card company studies the probability of default (outcome equation) for respondents who received a credit card offer. The first selection equation may be if they accepted the offer and applied for a card, and the second whether their application was approved by the bank. In this model, the agent can default only if she was approved for a credit card, which in turn is possible only if she has responded to such an offer. In

[4]This is a standard result in seemingly unrelated regression model, which does not apply if the explanatory variables are the same or if the correlation/covariance terms are zero.

labor economics it might be of interest to study employment discrimination (observed for candidates that seek a job) and wage discrimination (observed for candidates that seek a job and are hired). These two outcome equations can be estimated together with selection equation (if a candidate is seeking a job or not). All these and related models can be estimated in the framework developed in this chapter.

How is the problem of sample selection accounted for in the multivariate probit model? To continue with the health economics example, suppose that there exists some unobserved factor that affects both the probability of being selected into a sample and of having a morbid health event. If healthier individuals are more likely to allow access to their Medicare claims, then estimating the probability of a morbid health event only for the observed subsample is not representative of the entire population, as only its healthier part is considered. From the discussion above, it is apparent that in order to consistently estimate a model with incidental truncation, it is necessary to account for an omitted variable problem. In general, the sample selection problem arises if the unobserved factors determining the inclusion in the subsample are correlated with the unobservables that affect the endogenous variable of primary interest (Vella 1998). In the current chapter the specification error of omitted variable resulting from selection is dealt with by considering the unobserved omitted variable as a part of the disturbance term and then jointly estimating the system of equations accounting for the correlations in the variance-covariance matrix.

The multivariate probit model can be used to handle multiple correlated dichotomous variables along the lines of Ashford and Sowden (1970) and Amemiya

(1974). It seems, however, that the potential of this model has not been fully realized despite its connection to the normal distribution, which allows for a flexible correlation structure. As noticed in Chib and Greenberg (1998), the problem arises from the difficulties associated with evaluating the likelihood function by classical methods, except under simplifying assumptions like equicorrelated responses, as in Ochi and Prentice (1984).

Chib and Greenberg (1998) describe how the model can be reformulated in a Bayesian context using the technique of *data augmentation* (discussed in Albert and Chib [1992], among others). The discrete dependent variable in the probit model can be viewed as the outcome of an underlying linear regression with some *latent* dependent variable (i.e. unobserved by the researcher). Consider a decision to make a large purchase, as in Greene (2003, p. 669). If the benefits outweigh the costs (benefits-costs>0) then the latent dependent variable is positive and the purchase is made (the observed discrete outcome is 1), and vice versa. If the researcher makes a further assumption that the disturbance term in the model with the latent dependent variable has a standard normal distribution, then the univariate probit model results. The extension to the multivariate case is relatively straightforward.

The latent variables are clearly not observed, but their distributions are specified to be normal. Chib and Greenberg (1998) use this fact and re-introduce the latent variable back into the multivariate probit model. In a typical Bayesian model the prior distribution of the parameters and the likelihood function are used to obtain the joint posterior distribution, which combines the information from the prior

and the data. Chib and Greenberg (1998) find the joint posterior distribution of the multivariate probit model as the product of the prior distribution of the parameters and *augmented* likelihood function. The latter is obtained as the product of normal distributions for latent variables taken over all respondents in the sample. It is easy to show that, after integrating over the latent variables, the joint posterior distribution of the parameters is exactly the same as the posterior distribution obtained without introducing any latent variables (see Koop, Poirier and Tobias [2007] for related examples). The computational advantage of this method — it does not require the evaluation of the truncated multivariate normal density — is the greater the more discrete dependent variables are included into the model.

Using the full conditional posterior distributions of the coefficient vector, along with elements in the variance matrix and the latent data, it is possible to construct a Markov Chain Monte Carlo (MCMC) algorithm and simulate the parameters jointly with the latent data. In the Chib and Greenberg (1998) formulation, the conditional posterior distribution for the elements in the variance matrix has a nonstandard form and the authors use a Metropolis-Hastings algorithm to draw those elements. The current chapter modifies the Chib and Greenberg (1998) procedure by using the Cholesky factorization of the variance matrix. This allows a convenient multivariate normal representation of the parameters that are used to obtain the variance matrix, which considerably facilitates estimation.

Another complication in the sample selection model follows from the fact that some of the dependent binary variables in the outcome equation are not observed

given the selection rule into the sample. The posterior distribution of the latent data can be used to simulate those missing observations conditional on the covariance structure of the disturbance term. Consider first an individual t with complete data in $m \times 1$ vector of binary responses $y_{.t} = (y_{1t}, ..., y_{mt})'$ for all selection and outcome equations. The Chib and Greenberg (1998) procedure implies that at each MCMC simulation the latent vector $\widetilde{y}_{.t} = (\widetilde{y}_{1t}, ..., \widetilde{y}_{mt})'$ is drawn from the truncated multivariate normal distribution with a $m \times 1$ mean vector and $m \times m$ covariance matrix Σ.[5] The distribution is truncated for the ith element $\widetilde{y}_{it}$ to $(-\infty, 0]$ if the binary outcome $y_{it} = -1$ and to $(0, +\infty)$ if $y_{it} = 1$. Now suppose that individual t has missing binary outcome y_{it} for some i. The only difference with the case of an observed binary outcome y_{it} comes from the fact that the conditional multivariate normal distribution for $\widetilde{y}_{it}$ is no longer truncated in the ith dimension. That is, if y_{it} is missing for some i, then the latent variable $\widetilde{y}_{it}$ is unrestricted and can take any value in the interval $(-\infty, \infty)$.

Identification of the parameters is an important issue in models of discrete choice. It is well-known that the multivairate probit model is not likelihood-identified with unrestricted covariance matrix. Even though the formulation of the variance matrix in this chapter uses only $m(m-1)/2$ identified parameters, this turns out not to be sufficient for identification. Meng and Schmidt (1985) offer an elegant treatment of the problem of identification in the censored bivariate probit model

[5]The mean vector for individual t is a product of $m \times k$ matrix of covariates and a $k \times 1$ vector of coefficients to be defined later.

using the general principle in Rothenberg (1971) that the parameters in the model are (locally) identified if and only if the information matrix is nonsingular. The conclusion in Meng and Schmidt (1985), that the bivariate probit model with sample selection is in general identified, applies also with my parameterization of the model.

This chapter is organized as follows. Section 1.2 reviews the literature on sample selection especially on extensions to models with discrete outcome equation and Bayesian treatment. Section 1.3 sets up the model and derives the details of the multivariate probit estimator. Section 1.4 develops the Gibbs sampler. Section 1.5 considers the problem of identification in greater detail. Finally, section 1.6 provides an illustrative example and the last section concludes the discussion.

1.2 Heckman Model: Relevant Literature

Before developing a Bayesian model of sample selection with binary outcome variables, it is worth reviewing the relevant previous studies. After formulating a textbook variant of a Heckman (1979) model, I consider its extensions to models with discrete outcome equation prevailing in classical econometrics. The second feature of my research, namely Bayesian modeling of sample selection, is addressed in the third subsection of this literature review.

1.2.1 Introducing sample selection model

The model of incidental truncation, which is another name for sample selection model, has been widely used in economic applications when the variable of interest is observed only for people who are selected into a sample based on some threshold

rule.[6] Consider a simple model in health economics where it is of interest to assess the amount of Medicare spending for elderly Americans in a given year (outcome equation), which is observed only for people who allowed access to their Medicare claims (selection equation). Define the Medicare spending equation for respondent t as

$$y_{ot} = x_t'\delta + \epsilon_t, \tag{1.1}$$

and the selection equation of being linked to Medicare claims as

$$\widetilde{y}_{st} = z_t'\gamma + u_t, \tag{1.2}$$

where $\widetilde{y}_{st}$ is unobserved. What is observed is a binary variable y_{st} which equals 1 if $\widetilde{y}_{st} > 0$ (the agent allows the linking of her Medicare claims to the survey data) and 0 otherwise. The selection rule is that Medicare spending y_{ot} is observed only when $\widetilde{y}_{st} > 0$. If ϵ_t and u_t have a bivariate normal distribution with zero means and correlation ρ, then

$$E[y_{ot}|y_{ot}\ observed] = E[y_{ot}|\widetilde{y}_{st} > 0] = x_t'\delta + \rho\sigma_\epsilon\lambda(-z_t'\gamma/\sigma_u), \tag{1.3}$$

where $\lambda(-z_t'\gamma/\sigma_u) = \phi(z_t'\gamma/\sigma_u)/\Phi(z_t'\gamma/\sigma_u)$ as in Greene (2003). OLS estimator of Medicare spending using equation (1.1) and only the data on respondents who allowed access to Medicare claims, gives inconsistent estimates of δ as long as $\rho \neq 0$. This

[6]There exists extensive research in classical statistics (as opposed to classical econometrics) on a related topic of nonignorable nonresponse. A few relevant papers are Conaway (1993), Baker (1995), Baker and Laird (1988), Diggle and Kenward (1994) and Park (1998). This literature typically uses some form of ML estimator or EM algorithm, and it is only remotely related to my current research.

model can be estimated via maximum likelihood (ML), but Heckman's (1979) two-step estimator is typically used instead (Greene 2003). To obtain estimates of γ, the probit equation for y_{st} is estimated and for each observation in the selected sample $\lambda(-z_t'\widehat{\gamma})$ is computed. In the second stage, δ and $\delta_\lambda = \rho\sigma_\epsilon$ are estimated by the OLS regression of y_{ot} on x and $\widehat{\lambda}$. A t-test of the null hypothesis that the coefficient on $\widehat{\lambda}$ is equal to zero represents a test of no sample selectivity bias (Vella 1998).

Heckman's (1979) sample selection model is a standard topic in most modern econometric textbooks (such as Greene 2003 and Wooldridge 2002). Thorough review of the literature on sample selection is beyond the scope of this chapter, given the considerable attention that the model has acquired. A few recent review articles (Vella 1998, Lee 2003 and Greene 2006) seem to be a good starting point for an interested reader. In the next subsection I consider extensions of the Heckman model to the case of discrete outcome variable, relevant for my current research, developed in classical econometrics.

1.2.2 Discrete outcome equation in classical econometrics

There are relatively few applications of Heckman's (1979) model to discrete (and count) data and Greene (2008) reviews a handful of such models, starting with Wynand and van Praag (1981). In a recent application to teen employment, Mohanty (2002) uses the formulation in Meng and Schmidt (1985), which is very similar to the bivariate probit model with sample selection in Wynand and Praag (1981). In Mohanty (2002) the applicant i for a job can be selected ($SEL_i = 1$) or not ($SEL_i =$

0) only if she has applied for a job ($SEEK_i = 1$). Both discrete variables are modeled as the latent variables y_{1i} ($SEEK_i = 1$ if $y_{1i} > 0$ and $SEEK_i = 0$ otherwise) and y_{2i} ($SEL_i = 1$ if $y_{2i} > 0$ and $SEL_i = 0$ otherwise) that have bivariate normal distribution with correlation coefficient ρ.

Estimating the hiring equation (SEL_i) only for the subsample of teens who applied for a job ($SEEK_i = 1$) produces inconsistent estimates as long as $\rho \neq 0$. Indeed, univariate probit shows misleading evidence of employment discrimination against Black teens, which disappears when participation and hiring equations are estimated jointly (Mohanty 2002).

Another relevant example in classical econometrics is Greene (1992), who refers to an earlier paper by Boyes, Hoffman and Low (1989). The (part of the) model in Greene (1992) is bivariate probit where the decision to default or not on a credit card is observed only for cardholders (and not the applicants that were rejected by a credit card company).

Terza (1998) is another important reference in this literature. He develops a model for count data that includes an endogenous treatment variable. For example, the number of trips by a family (count variable of interest) may depend on the dummy for car ownership (potentially endogenous). In this case the dependent variable for car ownership in the first equation appears as explanatory variable in the equation for the number of trips and the two equations are estimated jointly. Terza (1998) compares three estimators for this model: full information ML, non-linear weighted least squares (NWLS) and a two-stage method of moments (TSM) similar to Heckman's (1979)

estimator.[7]

The setup in Terza (1998) can be potentially used in models of discrete choice with sample selection, as in a recent paper by Kenkel and Terza (2001). Kenkel and Terza (2001) use a two-step estimator in the model of alcohol consumption (number of drinks) with an endogenous dummy for advice (from a physician to reduce alcohol consumption). The first stage is univariate probit for receiving advice and the second stage applies non-linear least squares to the demand for alcohol (number of drinks). Kenkel and Terza (2001) find that advice reduces alcohol consumption in the sample of males with hypertension, and the failure to account for the endogeneity of advice would mask this result.

Munkin and Trivedi (2003) discuss the problems with different estimators of selection models with discrete outcome equation in classical econometrics. The first class of models, which uses moment-based procedures, results in inefficient estimates and does not allow the estimation of the full set of parameters in the presence of correlated multiple outcomes. A second possibility is a weighted nonlinear instrumental variable approach that has not been very successful because of difficulties in consistent estimation of weights (Munkin and Trivedi 2003). Finally, simulated maximum likelihood method requires a sufficient number of simulations for consistency where it is not clear what is "...the operational meaning of sufficient" (Munkin and Trivedi 2003, p. 198).

---

[7]The estimators are listed in the order of decreasing efficiency and computational difficulty. NWLS estimator may result in correlation coefficient being greater than one in absolute value.

It seems that none of the models discussed so far allows multiple correlated discrete dependent variables in the presence of sample selection (except for the bivariate case). Continuing with health economics, it might be of interest to estimate a model with a single selection equation (in sample or not) and two or more morbid health events (such as hip fracture and stroke). More importantly, if selection takes place along multiple dimensions, then each one should be accounted for to avoid the problems discussed in Heckman (1979). For example, if the sample is limited to participants who (i) allowed access to their Medicare claims and (ii) are self-respondents, then two selection equations can be easily introduced in my model. To the best of my knowledge, a model capable of estimating this kind of relationships has not been developed in classical econometrics yet.

The approach that I adopt in this chapter is to apply the multivariate probit model in Bayesian framework, allowing for some missing responses. Chib and Greenberg (1998) discuss the problems with estimation of multivariate probit model by methods of classical econometrics and offer a Markov Chain Monte Carlo algorithm which constitutes the starting point of my investigation. I review existing Bayesian treatments of sample selection in the next subsection and then provide further details on Chib and Greenberg (1998).

1.2.3 Bayesian treatment of Heckman model

Recent Bayesian treatments of sample selection model are almost exclusively based on *Markov Chain Monte Carlo* (MCMC) methods with *data augmentation*.[8] The idea of data augmentation was introduced by Tanner and Wong (1987), and used in Bayesian discrete choice models starting (at least) from Albert and Chib (1993).[9] Latent variables in these models are treated as additional parameters and are sampled from the joint posterior distribution. In these models, however, the joint posterior distribution for parameters and latent variables typically does not have a recognizable form. *Gibbs sampler* is an MCMC method used when the joint posterior distribution can be represented as a full set of (simpler) conditional distributions. It is possible then to obtain the sample from the joint posterior distribution by iteratively drawing from each conditional distribution, *given* the values obtained from the remaining distributions. The model developed herein shares the two aforementioned features (data augmentation and Gibbs sampling) and simultaneous equation structure with previous studies by Li (1998), Huang (2001) and van Hasselt (2008).

Li (1998) develops Bayesian inference in the following simultaneous equation

---

[8]Earlier developments in Bayesian statistics model selection by means of various weight functions. For example, Bayarri and DeGroot (1987 and four other papers, as cited in Lee and Berger 2001) mostly concentrate on indicator weight function: potential observation is selected into a sample if it exceeds a certain threshold. Bayarri and Berger (1998) develop nonparametric classes of weight functions that are bounded above and below by two weight functions. Lee and Berger (2001) use the Dirichlet process as a prior on the weight function.

[9]Notice that the selection equation in a Heckman-type model is univariate probit.

model with limited dependent variables (SLDV):

$$\begin{aligned} y_1^* &= y_2\gamma_1 + X_1\delta_1 + u_1 \\ y_2^* &= X_2\delta_2 + u_2, \end{aligned} \tag{1.4}$$

where y_1^* is of Tobit type (a researcher observes $y_1 = y_1^*$ if $y_1^* > 0$ and $y_1 = 0$ otherwise) and y_2^* is of probit type (the researcher observes $y_2 = 1$ if $y_2^* > 0$ and $y_2 = 0$ otherwise).[10] The vector of disturbances $(u_1, u_2)'$ is assumed to follow bivariate normal distribution with the variance of u_2 set to 1 for model identification:

$$\Sigma = \begin{pmatrix} \sigma_{11}^2 & \sigma_{12} \\ \sigma_{12} & 1 \end{pmatrix},$$

where σ_{11}^2 is the variance of u_1 and σ_{12} is the covariance between u_1 and u_2. Decomposing the joint bivariate distribution of $(u_1, u_2)'$ into the product of the marginal distribution of u_2 and the conditional distribution of $u_1|u_2$ allows convenient blocking in the Gibbs sampler. This decomposition in Li (1998), together with the more convenient reparametrization of the variance matrix

$$\Sigma = \begin{pmatrix} \sigma^2 + \sigma_{12}^2 & \sigma_{12} \\ \sigma_{12} & 1 \end{pmatrix},$$

appear repeatedly in later studies. With these changes the model is now re-defined as

$$\begin{aligned} y_1^* &= y_2\gamma_1 + X_1\delta_1 + u_2\sigma_{12} + v_1 \\ y_2^* &= X_2\delta_2 + u_2, \end{aligned} \tag{1.5}$$

[10]To avoid the confusion with my parameters later on, I use different Greek letters from those used in the original papers throughout the literature review.

with $u_2 = y_2^* - X_2\delta_2$, $\sigma^2 = \sigma_{11}^2 - \sigma_{12}^2$, and $v_1 \sim N(0, \sigma^2)$. In the resulting Gibbs sampler with data augmentation, all conditional distributions have recognizable forms that are easy to draw from (multivariate normal, univariate truncated normal and gamma).[11]

Huang (2001) develops Bayesian seemingly unrelated regression (SUR) model, where dependent variables are of the Tobit type (researcher observes $y_{ij} = y_{ij}^*$ if $y_{ij}^* > 0$ and $y_{ij} = 0$ otherwise). The Gibbs sampler with data augmentation in Huang (2001) consists of multivariate normal, Wishart and truncated multivariate normal distributions.

In the paper by van Hasselt (2008), two sample selection models — with unidentified parameters and with identified parameters — are compared.[12] The idea behind the first model is borrowed from McCulloch and Rossi (1994), who used a similar approach in multinomial probit context. The output from the Gibbs sampler is used to approximate the posterior distribution of the identified parameters. The model with identified parameters in van Hasselt (2008) uses marginal-conditional decomposition of the disturbance terms together with more convenient parameteriza-

[11] Chakravarti and Li (2003) apply this model to estimate dual trade informativeness in futures markets. Probit equation estimates a trader's decision to trade on her own account and tobit equation measures her (abnormal) profit from her own account trading. Chakravarti and Li (2003) did not find significant correlation between a dual trader's private information and her abnormal profit.

[12] Another interesting paper by van Hasselt (2005) compares the performance of sample selection and two-part models (when two equations are estimated independently) in a Bayesian setup. In classical econometrics Leung and Yu (1996) provide conclusive evidence against negative results in Manning, Duan and Rogers (1987) who claim that two-part model performs better than sample selection model even when the latter is the true model. Leung and Yu (1996) show that problems with sample selection model are caused by a critical problem in the design of experiments in Manning, Duan and Rogers (1987).

tion of the variance matrix, as in Li (1998).[13] The major contribution of van Hasselt (2008) is relaxing the normal distribution assumption in the sample selection model via mixture of normal distributions. I do not follow that route and my model remains fully parametric.

In all the papers cited above the outcome variable is continuous and not discrete. There are two Bayesian papers with discrete outcome variable (and multiple outcome equations) that are worth mentioning: Munkin and Trivedi (2003) and Preget and Waelbroeck (2006).

Munkin and Trivedi (2003) develop a three-equation model with the first equation for count data (the number of doctor visits), the second equation for a continuous variable (the associated health expenditures) and the third equation for a dummy variable (the type of health insurance plan). The selection problem — demand for health care that potentially depends on the type of health insurance — is modeled by using an (endogenous) dummy variable for private health plan. There is no problem of missing dependent variable for respondents that are not in the sample (i.e. who did not purchase private insurance). Neither of the correlation coefficients for private health plan with two variables of interest is statistically different from zero and the type of insurance does not affect the level of health care use (Munkin and Trivedi 2003).[14]

---

[13]McCulloch, Polson and Rossi (2000) show that fully identified multinomial probit model comes at a cost: higher autocorrelation in the Markov Chain.

[14]In a later work, Deb, Munkin and Trivedi (2006), perhaps dissatisfied with a sample selection model, use a two-part model with endogeneity in a similar context.

Preget and Waelbroeck (2006) develop a three-equation model with application to timber auctions. There are two binary dependent variables (if a lot received any bids and, conditional on receiving at least one bid, if a lot received two or more bids) and one continuous variable (highest bid for a lot) with an endogenous dummy variable for the number of bids. Preget and Waelbroeck (2006) comment that in such models the likelihood function is not always well behaved, especially in the direction of the correlation coefficients.[15] While in Preget and Waelbroeck (2006) the correlation coefficients are never statistically different from zero, they find that their Bayesian algorithm "...yields a remarkably stable coefficient for the binary endogenous variable and was able to deal with irregularities in the likelihood function."

Two conclusions seem to follow from my review of relevant studies. First of all, there exist serious computational difficulties when the sample selection model with multiple dichotomous dependent variables is estimated by methods of classical econometrics. For example, Munkin and Trivedi (2003) comment on difficulties associated with estimating their model in a simulated maximum likelihood framework. This provides strong motivation for a Bayesian econometric methodology and also explains why models similar to mine are typically estimated in a Bayesian and not classical tradition. Second, even in the Bayesian literature, there seem to be no

[15]Consider the following sequential probit model: the second binary outcome is missing for all respondents whose first outcome is "No." The third binary outcome, if present, is missing for all respondents who answered "No" in the second equation and so on. Waelbroeck (2005) argues that in this model the likelihood function is not globally concave and flat in some directions, which limits practical applicability of the model. Notice that in a two-equation case, sequential probit is the same model as censored probit except that the two models may have different interpretation. Keane (1992) discusses similar computational issues in multinomial probit model.

published papers that can be used directly to estimate a model with three or more dichotomous dependent variables. This constitutes an important contribution of the current chapter.

While my work shares the methods with previous studies (data augmentation, Gibbs sampling and simultaneous equation structure) it comes from a different area — multivariate probit model developed in Chib and Greenberg (1998). The next section introduces the multivariate probit in Chib and Greenberg (1998) and provides the extensions that make it applicable in the sample selection model.

1.3 Multivariate Probit and Sample Selection

Suppose that a researcher observes a set of potentially correlated binary events $i = 1, ..., m$ over an independent sample of $t = 1, ..., T$ respondents. Consider the multivariate probit model reformulated in terms of latent variables as in Chib and Greenberg (1998). For each of the events $i = 1, ..., m$ define a $T \times 1$ vector of latent variables $\widetilde{y}_{i\cdot} = (\widetilde{y}_{i1}, ..., \widetilde{y}_{iT})'$ and a $T \times k_i$ matrix of explanatory variables Z_i where each row t represents a $1 \times k_i$ vector Z_{it}. Then each latent variable can be modeled as

$$\widetilde{y}_{i\cdot} = Z_i \beta_i + \varepsilon_{i\cdot}, \tag{1.6}$$

where $\varepsilon_{i\cdot}$ is a vector of disturbance terms that have normal distribution. There is potential correlation in the disturbance terms for respondent t across events $i = 1, ..., m$ coming from some unobserved factor that simultaneously affects selection and outcome variables. Let $\widetilde{y}_{\cdot t} = (\widetilde{y}_{1t}, ..., \widetilde{y}_{mt})'$ be the vector of latent variables for

respondent t such that

$$\widetilde{y}_{.t} \sim N_m(Z_t\beta, \Sigma), \tag{1.7}$$

where $Z_t = diag(Z_{1t}, ..., Z_{mt})$ is an $m \times k$ covariate matrix, $\beta_i \, \epsilon \, R^{k_i}$ is an unknown parameter vector in equation $i = 1, ..., m$ with $\beta = (\beta_1', ..., \beta_m')' \, \epsilon \, R^k$ and $k = \sum_{i=1}^{m} k_i$, and Σ is the variance matrix.

The sign of $\widetilde{y}_{it}$ for each dependent variable $i = 1, ..., m$ uniquely determines the observed binary outcome y_{it}:

$$y_{it} = I(\widetilde{y}_{it} > 0) - I(\widetilde{y}_{it} <= 0) \quad (i = 1, ..., m), \tag{1.8}$$

where $I(A)$ is the indicator function of an event A. Suppose it is of interest to evaluate the probability of observing a vector of binary responses $Y_{.} = (Y_1, ..., Y_m)'$ for indivial t. Chib and Greenberg (1998) show that the probability $y_{.t} = Y_{.t}$ can be expressed as

$$\int_{B_{mt}} ... \int_{B_{1t}} \phi_m(\widetilde{y}_{.t}|Z_t\beta, \Sigma) d\widetilde{y}_{.t}, \tag{1.9}$$

where $B_{it} \, \epsilon \, (0, \infty)$ if $y_{it} = 1$ and $B_{it} \, \epsilon \, (-\infty, 0]$ if $y_{it} = -1$. Define $B_t = B_{1t} \times ... \times B_{mt}$.

Alternatively, the probability $y_{.t} = Y_{.t}$ can be expressed without introducing latent variables as

$$pr(y_{.t} = Y_{.t}|\beta, \Sigma) = \int_{A_{mt}} ... \int_{A_{1t}} \phi_m(w|0, \Sigma) dw, \tag{1.10}$$

where $\phi_m(w|0, \Sigma)$ is the density of a m-variate normal distribution and A_{it} is the interval defined as

$$A_{it} = \begin{cases} (-\infty, Z_{it}\beta_i) & \text{if } y_{it} = 1, \\ (Z_{it}\beta_i, \infty) & \text{if } y_{it} = -1. \end{cases}$$

The multidimensional integral over the normal distribution in (1.10) is hard to evaluate by conventional methods.[16]

Instead of evaluating this integral, Chib and Greenberg (1998) use the formulation in (1.9) and simulate the latent variable $\widetilde{y}_{.t}$ from the conditional posterior distribution with mean $Z_t\beta$ and variance matrix Σ. This distribution is truncated for the ith element to $(-\infty, 0]$ if the observed outcome is $y_{it} = -1$ and to $(0, +\infty)$ if $y_{it} = 1$. The current model also assumes that $y_{it} = 0$ when the response for event i is missing for t.

It is important to understand what missing binary response means in terms of the latent data representation. If respondent t has missing binary response y_{it} for some i then no restriction can be imposed on the latent normal distribution in the ith dimension. Then the vector $\widetilde{y}_{.t}$ is simulated from the m-variate normal distribution with the same mean and variance as in the complete data case but the distribution is not truncated for the ith element. For the case of missing outcome i the latent variable $\widetilde{y}_{it}$ can take any value in the interval $(-\infty, \infty)$.[17]

The multivariate model of incidental truncation can not be estimated using only the observed data because the endogenous selection variables are constant and

[16]Quadrature method is an example of nonsimulation procedure that can be used to approximate the integral. Quadrature operates effectively only when the dimension of integral is small, typically not more than four or five (Train 2003). The GHK simulator is the most widely used simulation method after Geweke (1989), Hajivassiliou (as reported in Hajivassiliou and McFadden 1998) and Keane (1994).

[17]This methodology allows for continuous endogenous variables as well. In this case $\widetilde{y}_{jt}$ is trivially set to the observed y_{jt} for a continuous variable j in each iteration of the MCMC algorithm introduced below.

equal to 1. Now, due to simulated missing data one can estimate the variance matrix Σ, which is the focus of the procedure to account for sample selection. The covariances in Σ effectively adjust for sample selectivity in the outcome equations by controlling for unobserved heterogeneity.

The issue of sample selection arises whenever the unobserved factors determining the inclusion in the sample are correlated with the unobservables that affect the outcome variable(s) of primary interest (Vella 1998). The critical idea in the current work is to account for selection in binary outcome equation(s) by jointly estimating selection and outcome equations while controlling for possible unobserved effect through multivariate probit with correlated responses. If the covariance terms belong to the highest posterior density region, this indicates the presence of unobserved effect and, hence, sample selection bias.

The elements in the variance matrix in the Chib and Greenberg (1998) formulation do not have the conditional posterior distribution of a recognizable form, which forces them to employ a Metropolis-Hastings algorithm. This chapter makes the technical advance that allows convenient multivariate normal representation of the parameters used to obtain the variance matrix. Consider the Cholesky factorization of the inverse of the variance matrix $\Sigma^{-1} = \breve{F} \cdot \breve{F}'$ where $\breve{F}$ is the lower triangular matrix. If the diagonal elements of $\breve{F}$ are arrayed in a diagonal matrix Q then $\Sigma^{-1} = \breve{F}Q^{-1}Q^2Q^{-1}\breve{F}' = FQ^2F$ (Greene 2003). In the current work the variance matrix is defined by F which is a lower triangular matrix that has ones on the main

diagonal and $D^{-1} = Q^2$ which is a diagonal matrix. Then

$$\Sigma = (F')^{-1} D F^{-1}, \tag{1.11}$$

with $D = diag\{d_{11}, ..., d_{mm}\}$ and F is lower triangular

$$F = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ f_{21} & 1 & 0 & \cdots & 0 \\ f_{31} & f_{32} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ f_{m1} & f_{m2} & f_{m3} & \cdots & 1 \end{pmatrix}.$$

Finally, consider the system of m equations

$$\underbrace{\widetilde{y}}_{Tm\times 1} = \begin{pmatrix} \widetilde{y}_{1.} \\ \widetilde{y}_{2.} \\ \vdots \\ \widetilde{y}_{m.} \end{pmatrix}, \quad \underbrace{Z}_{Tm\times k} = \begin{bmatrix} Z_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & Z_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & Z_m \end{bmatrix}, \quad \underbrace{\beta}_{k\times 1} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{pmatrix},$$

so that the model can be represented as

$$\widetilde{y} = Z\beta + \varepsilon, \tag{1.12}$$

where $k = \sum_{i=1}^{m} k_i$ and

$$\underbrace{\varepsilon}_{Tm\times 1} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_m \end{pmatrix}.$$

Under the maintained assumption of the normally distributed vector ε it follows that

$$\varepsilon|(\beta, F, D, Z) \sim N(0, (F')^{-1} D F^{-1} \otimes I_T). \tag{1.13}$$

1.4 Deriving the Gibbs Sampler

Consider a sample of $m \times T$ observations $y = (y_{.1}, ..., y_{.T})$ that are independent over $t = 1, ..., T$ respondents but are potentially correlated over $i = 1, ..., m$ events. Given a prior density $p(\beta, F, D)$ on the parameters β, F and D the posterior density is equal to

$$p(\beta, F, D|y) \propto p(\beta, F, D) p(y|\beta, \Sigma), \tag{1.14}$$

where $p(y|\beta, \Sigma) = \prod_{t=1}^{T} p(y_{.t}|\beta, \Sigma)$ is the likelihood function. Define $y_{.t} = (y_{st}, y_{ot})$, where y_{st} and y_{ot} are selection and outcome variables with some of the $y_{.t}$'s missing. In this representation the evaluation of the likelihood function is computationally intensive from a classical perspective. Albert and Chib (1993) developed an alternative Bayesian framework that focuses on the joint posterior distribution of the parameters and the latent data $p(\beta, F, D, \widetilde{y}_1, ..., \widetilde{y}_T|y)$. It follows then that

$$\begin{aligned} p(\beta, F, D, \widetilde{y}|y) &\propto p(\beta, F, D) p(\widetilde{y}|\beta, \Sigma) p(y|\widetilde{y}, \beta, \Sigma) \\ &= p(\beta, F, D) p(\widetilde{y}|\beta, \Sigma) p(y|\widetilde{y}). \end{aligned} \tag{1.15}$$

It is possible now to implement a sampling approach and construct a Markov chain from the distributions $[\widetilde{y}_{.t}|y_{.t}, \beta, \Sigma]$ $(t \leq T)$, $[\beta|y, \widetilde{y}, \Sigma]$ and $[F, D|y, \widetilde{y}, \beta]$.

With unrestricted F or D matrix the multivariate probit model is not identified. The observed outcomes $y_{.t}$ for respondent t depend only on signs but not magnitudes of the latent data $\widetilde{y}_{.t}$. In a multivariate probit model with m equations only

$m(m-1)/2$ parameters in the variance matrix are identified. Consider the following transformation of the model $F'\widetilde{y}_{.t} \sim N(F'Z_t\beta, D)$, where D is some unrestricted diagonal matrix. The latent regression has the form $F'\widetilde{y}_{.t} = F'Z_t\beta + D^{1/2}\varepsilon_{.t}$, where $\varepsilon_{.t}$ is m-variate normal with a zero mean vector and an $m \times m$ identity variance matrix. However, pre-multiplying this equation by $\alpha > 0$ results in $\alpha F'\widetilde{y}_{.t} = F'Z_t(\alpha\beta) + \alpha D^{1/2}\varepsilon_{.t}$ which is the same model corresponding to the same observed data $y_{.t}$. Since the parameters in $D^{1/2}$ cannot be identified, D is set to identity matrix extending the logic from the univariate probit model in Greene (2003).[18]

The posterior density kernel is the product of the priors and the augmented likelihood in equation (1.15).[19] The parameters in β and F are specified to be independent in the prior. Let the prior distribution for β be normal $\phi_k(\beta|\underline{\beta}, \underline{B}^{-1})$ with the location vector $\underline{\beta}$ and the precision matrix $\underline{B}$.

It is convenient to concatenate the vectors below the main diagonal in F matrix as

$$F_{vector} = \begin{pmatrix} F_{2:m,1} \\ F_{3:m,2} \\ \vdots \\ F_{m,m-1} \end{pmatrix},$$

where $F_{i+1:m,i}$ for $i = 1, ..., m-1$ represents elements from $i+1$ to m in column i.

[18]Observe that this is not sufficient for identification and later I give an example from Meng and Schmidt (1985) when the model is not identified with two equations.

[19]The term "augmented likelihood" emphasizes the fact that the likelihood includes latent variables.

The prior distribution of F_{vector} is assumed to be $\left(\frac{m(m-1)}{2}\right)$-variate normal

$$F_{vector} \sim N(\underline{F}_{vector}, \underline{H}^{-1}). \tag{1.16}$$

In this expression $\underline{F}_{vector}$ is the prior mean of the normal distribution, and the prior variance matrix $\underline{H}^{-1}$ is block-diagonal with

$$\underline{H} = \begin{pmatrix} \underline{H}_{2:m,2:m} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \underline{H}_{3:m,3:m} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \underline{H}_{1,1} \end{pmatrix}.$$

This precision matrix has $(m-1) \times (m-1)$ matrix $\underline{H}_{2:m,2:m}$ in the upper left corner and the matrix dimension is decreasing by one in each consequent block on the main diagonal. The lower right matrix $\underline{H}_{1,1}$ is a scalar. The posterior density kernel is now

$$\begin{aligned} & |\underline{B}|^{1/2} \exp\Big\{ -\frac{1}{2}(\beta - \underline{\beta})' \underline{B} (\beta - \underline{\beta}) \Big\} \\ \cdot \quad & |\underline{H}|^{1/2} \exp\Big\{ -\frac{1}{2}(F_{vector} - \underline{F}_{vector})' \underline{H} (F_{vector} - \underline{F}_{vector}) \Big\} \\ \cdot \quad & |\Sigma|^{-T/2} \prod_{t=1}^{T} \exp\Big\{ -\frac{1}{2}(\widetilde{y}_{.t} - Z_t\beta)' \Sigma^{-1} (\widetilde{y}_{.t} - Z_t\beta) \Big\} I(\widetilde{y}_{.t} \in B_t). \end{aligned} \tag{1.17}$$

A Gibbs sampler is constructed by drawing from the following conditional posterior distributions: the vector of coefficients β, the F_{vector} from the variance matrix decomposition and the latent vector $\widetilde{y}_{.t}$ for each respondent $t \leq T$.[20]

In a typical iteration the Gibbs sampler initiates by drawing the vector of the coefficients β conditional on F_{vector} and $\widetilde{y}_{.t}$ obtained from the previous draw. The

[20]Appendix A.1 provides complete details of the Gibbs sampler derivation.

posterior distribution of β comes from the posterior density kernel and is normal

$$\beta|(\widetilde{y},\Sigma) \sim N_k(\beta|\overline{\beta},\overline{B}^{-1}), \tag{1.18}$$

where $\overline{B} = \underline{B} + \sum_{t=1}^{T} Z_t'\Sigma^{-1}Z_t$ and $\overline{\beta} = \overline{B}^{-1}(\underline{B}\underline{\beta} + \sum_{t=1}^{T} Z_t'\Sigma^{-1}\widetilde{y}_{.t})$. In this last expression it is understood that for each t , $\widetilde{y}_{.t} \epsilon B_t$.

To obtain the conditional posterior distribution of F, an alternative expression for the density of $\widetilde{y}$ is useful:

$$\begin{aligned} p(\widetilde{y}|y,\beta,F,D) &\propto |\Sigma|^{-T/2}\prod_{t=1}^{T}\exp\Big\{-\frac{1}{2}(\widetilde{y}_{.t}-Z_t\beta)'\Sigma^{-1}(\widetilde{y}_{.t}-Z_t\beta)\Big\}I(\widetilde{y}_{.t}\,\epsilon\,B_t) \\ &= |FD^{-1}F'|^{T/2}\prod_{t=1}^{T}\exp\Big\{-\frac{1}{2}\varepsilon_t'FD^{-1}F'\varepsilon_t\Big\}I(\widetilde{y}_{.t}\,\epsilon\,B_t) \\ &= \prod_{t=1}^{T}\prod_{i=1}^{m}\exp\Big\{-\frac{1}{2}(\varepsilon_{t,i}+F'_{i+1:m,i}\varepsilon_{t,i+1:m})^2\Big\} \\ &= \prod_{i=1}^{m}\exp\Big\{-\frac{1}{2}\sum_{t=1}^{T}(\varepsilon_{t,i}+F'_{i+1:m,i}\varepsilon_{t,i+1:m})^2\Big\}, \end{aligned} \tag{1.19}$$

where for each t , $\widetilde{y}_{.t} \epsilon B_t$. In this derivation the restriction $D = I_m$ is already imposed. Then the posterior conditional distribution of F_{vector} is also normal

$$F_{vector}|(y,\widetilde{y},\beta) \sim N_{\left(\frac{m(m-1)}{2}\right)}(\overline{F}_{vector},\overline{H}^{-1}). \tag{1.20}$$

The conditional posterior normal distribution has the posterior precision matrix

$$\overline{H} = \underline{H} + \begin{pmatrix} \sum_{t=1}^{T}\varepsilon_{t,2:m}\varepsilon_{t,2:m}' & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \sum_{t=1}^{T}\varepsilon_{t,3:m}\varepsilon_{t,3:m}' & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \sum_{t=1}^{T}\varepsilon_{t,m}\varepsilon_{t,m}' \end{pmatrix}.$$

The posterior mean of the normal distribution is equal to

$$\overline{F}_{vector} = \overline{H}^{-1}\underline{H}\ \underline{F}_{vector} - \overline{H}^{-1}\begin{pmatrix} \sum_{t=1}^{T} \varepsilon_{t,2:m}\varepsilon_{t,1} \\ \sum_{t=1}^{T} \varepsilon_{t,3:m}\varepsilon_{t,2} \\ \vdots \\ \sum_{t=1}^{T} \varepsilon_{t,m}\varepsilon_{t,m-1} \end{pmatrix}.$$

Finally, the latent data $\widetilde{y}_{.t}$ are drawn independently for each respondent $t \leq T$ from the truncated multivariate normal distribution as described in Geweke (1991). The algorithm makes draws conditional on Z_t, β and F as well as $\widetilde{y}_{.t}$ obtained in the previous draw. The multivariate normal distribution is truncated to the region defined by the $m \times 2$ matrix $[a, b]$ with a typical row i equal to $(0, \infty)$, if $y_{it} = 1$ and $(-\infty, 0)$ if $y_{it} = -1$. If y_{it} is not observed, then row i is $(-\infty, \infty)$.

Thus, this work extends Chib and Greenberg (1998) in the following two ways: (i) it permits missing outcome variables $\widetilde{y}_{.t}$, and (ii) it re-parameterizes the variance matrix in terms of more convenient multivariate normal F_{vector} that is used to obtain Σ.

1.5 The Problem of Identification

Identification is an important issue in models of discrete choice. Meng and Schmidt (1985) in their elegant article offer an excellent treatment of identification in a bivariate probit model under various levels of observability. Meng and Schmidt (1985) rely on the general principle in Rothenberg (1971) that the parameters are (locally) identified if and only if the information matrix is nonsingular. In particular, their *Case Three: Censored Probit* is very similar to the following bivariate sample

selection model: the binary variable of interest y_{2t} is observed for respondent t only if she is selected in the sample ($y_{1t} = 1$).[21]

Let $F^t = F(Z_{1t}\beta_1, Z_{2t}\beta_2; f_{21})$ specify the bivariate normal cumulative distribution function and $\Phi(Z_{ht}\beta_h)$ specify the univariate standard normal cumulative distribution function with $h = 1, 2$ for respondent t. Recall that the sign of $\widetilde{y}_{it}$ perfectly predicts y_{it} and one can write

$$
\begin{aligned}
p(y|\widetilde{y}) = &\prod_{t=1}^{T} I(\widetilde{y}_{1t} > 0)I(\widetilde{y}_{2t} > 0)I(y_{1t} = 1)I(y_{2t} = 1) \\
&+I(\widetilde{y}_{1t} > 0)I(\widetilde{y}_{2t} \leq 0)I(y_{1t} = 1)I(y_{2t} = -1) + I(\widetilde{y}_{1t} \leq 0)I(y_{1t} = -1).
\end{aligned}
$$

The likelihood function in the bivariate model can be obtained after I integrate over $\widetilde{y}$ in the following way

$$
\begin{aligned}
\int_B \quad & p(y, \widetilde{y}|\beta, \Sigma)d\widetilde{y} = \int_B p(y|\widetilde{y}, \beta, \Sigma)p(\widetilde{y}|\beta, \Sigma)d\widetilde{y} = \int_B p(y|\widetilde{y})p(\widetilde{y}|\beta, \Sigma)d\widetilde{y} \\
&= \prod_{t=1}^{T} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \Big[I(\widetilde{y}_{1t} > 0)I(\widetilde{y}_{2t} > 0)I(y_{1t} = 1)I(y_{2t} = 1) \\
&+I(\widetilde{y}_{1t} > 0)I(\widetilde{y}_{2t} \leq 0)I(y_{1t} = 1)I(y_{2t} = -1) + I(\widetilde{y}_{1t} \leq 0)I(y_{1t} = -1) \Big] \\
&\cdot f(Z_{1t}\beta_1, Z_{2t}\beta_2; f_{21})d\widetilde{y}_{1t}d\widetilde{y}_{2t} \\
&= \prod_{t=1}^{T} \int_{0}^{\infty} \int_{0}^{\infty} I(y_{1t} = 1)I(y_{2t} = 1)f_2 d\widetilde{y}_{1t}d\widetilde{y}_{2t} \\
&+ \prod_{t=1}^{T} \int_{-\infty}^{0} \left(\int_{0}^{\infty} I(y_{1t} = 1)I(y_{2t} = -1)f_2 d\widetilde{y}_{1t} \right) d\widetilde{y}_{2t} \\
&+ \prod_{t=1}^{T} \int_{-\infty}^{0} I(y_{1t} = -1)\phi(Z_{1t}\beta_1)d\widetilde{y}_{1t}
\end{aligned}
\tag{1.21}
$$

[21] I employ different parametrization of the variance matrix and, thus, the parameters have to be scaled to be comparable with Meng and Schmidt (1985).

$$= \prod_{t=1}^{T} F(Z_{1t}\beta_1, Z_{2t}\beta_2; f_{21})^{I(y_{1t}=1)I(y_{2t}=1)}$$
$$\cdot[\Phi(Z_{1t}\beta_1) - F(Z_{1t}\beta_1, Z_{2t}\beta_2; f_{21})]^{I(y_{1t}=1)I(y_{2t}=-1)}$$
$$\cdot[1 - \Phi(Z_{1t}\beta_1)]^{I(y_{1t}=-1)},$$

where $f = f(Z_{1t}\beta_1, Z_{2t}\beta_2; f_{21})$ is bivariate normal density function and $\phi(\cdot)$ is univariate normal density function. Define $q_{it} = \frac{y_{it}+1}{2}$ for $i = 1, 2$ and take the natural logarithm of this expression to obtain

$$\ln L(Z_1\beta_1, Z_2\beta_2; f_{21}) = \tag{1.22}$$
$$\sum_{t=1}^{T} \Big[q_{1t}q_{2t} \ln F^t + q_{1t}(1 - q_{2t}) \ln[\Phi(Z_{1t}\beta_1) - F^t] + (1 - q_{1t}) \ln[1 - \Phi(Z_{1t}\beta_1)]\Big],$$

which is equivalent to equation (6) in Meng and Schmidt (1985) except for the correlation coefficient ρ being replaced by the parameter f_{21} that enters F^t as defined below equation (1.11). The general conclusion reached by Meng and Schmidt (1985) is that the parameters in this model are identified except in certain "perverse" cases. First of all, peculiar configurations of the explanatory variables may cause nonidentification, but this problem can be addressed only given the data at hand. Second, nonidentification may be caused by certain combinations of parameters in the model. For example, the censored bivariate probit model with my parametrization is not identified when $Z_{1t}\beta_1 = \frac{-f_{21}}{\sqrt{1+f_{21}^2}} Z_{2t}\beta_2$ for all respondents t and I show this result in Appendix A.2.[22] The information matrix is then singular because the row for the

[22] Another example of nonidentification given in Meng and Schmidt (1985) is when there are only intercepts included in all equations. While such a model cannot be used in a meaninful way for economic analysis, it provides an interesting limiting case when all the covariate coefficients go to zero.

second intercept (i.e. for $(k_1+1)th$ term) is the last row (i.e. for the parameter f_{21}), divided by a constant. In this particular example the problem of nonidentification does not arise as long as the set of explanatory variables is not the same in two equations.

Meng and Schmidt (1985) comment that there might also be other combinations of parameters or particular configurations of explanatory variables leading to nonidentification. Since it is not possible to foresee all such problems *a priori*, it is the responsibility of the researcher to check if the parameters in the model are identified. However, it is very reassuring that the sample selection model is generally identified, except in some (not very likely) cases.

1.6 Experiments with Artificial Data

The purpose of the experiment with artificial data is to study if the model can retrieve the parameters and the correlation coefficient that are used to generate the data when some of the outcome variables are missing. It is also of interest to assess the convergence properties of the model. I construct the following bivariate probit model with sample selection. Let y_{2t} be the dichotomous dependent variable of interest that is observed only if the selection variable y_{1t} is equal to 1.

For this experiment I generate $t = 1, ..., 500$ independent latent variables $(\widetilde{y}_{1t}, \widetilde{y}_{2t})'$ from the bivariate normal distribution with mean $\mu_t = [Z_{1t}\beta_{1,.}\sqrt{1+f_{21}^2}, Z_{2t}\beta_{2,.}]'$, where a 1×3 vector Z_{it} contains intercept, one discrete and one continuous variable as described below and $\beta_{i,.} = [\beta_{i,1}\ \beta_{i,2}\ \beta_{i,3}]'$ for $i = 1, 2$. Each equation contains the

intercept denoted $\beta_{i,1}$, continuous variable $\beta_{i,2}$ and discrete variable $\beta_{i,3}$. Continuous variable in each equation is drawn from the normal distribution with $\mu = -0.5$ and $\sigma = 2$. Discrete variable takes values of -1 and 1 with equal probability. All continuous and discrete variables are independent from each other. The coefficients used to generate the artificial data are provided in the second column of Table 1.1. The correlation coefficient is set to 0.5 with the corresponding value of $f_{21} \approx -0.5774$. Finally, the 2×2 covariance matrix is the same for all respondents and is set to

$$\Sigma = \begin{bmatrix} 1 + f_{21}^2 & -f_{21} \\ -f_{21} & 1 \end{bmatrix}.$$

Observe that the true parameters of the first equation are multiplied by $\sqrt{1 + f_{21}^2}$ and in each simulation I normalize the draws of $\beta_{1,\cdot}$ by $\sqrt{1 + f_{21}^2}$ obtained in the same draw. After I obtain the 500×2 matrix of the latent variables $\widetilde{y}$, I convert it into the matrix of "observed" dichotomous dependent variables y which is used in the simulator. The coefficients that were chosen place approximately one third in each of the three bins (yes, yes), (yes, no) and $(no, missing)$.

The implementation of the Gibbs sampler is programmed in the Matlab environment with some loops written in C language. All the codes successfully passed the joint distribution tests in Geweke (2004). The results in this section are based on 24,000 draws from the posterior (the first 6,000 draws were discarded as burn-in iterations). The prior for $i = 1, 2$ vector of coefficients $\beta_{i,\cdot}$ is mutlivariate normal with the mean vector set to zeros and the variance matrix equal to the identity matrix of dimension 3. The prior for f_{21} is standard normal distribution.

The results of the experiment are shown in Figures 1.1-1.3 and Table 1.1.[23] The simulator works quite well in this experiment with low autocorrelation and stable results with histograms centered almost at the values of the parameters used to generate the data. Geweke's convergence diagnostic test (Geweke 1992) does not indicate problems with the convergence of the Markov Chain. The only slight problem is that the mean of the correlation coefficient ρ in the sample obtained from the joint posterior distribution (0.23) is somewhat lower than the value of 0.5 used to obtain the artificial data but it still belongs to the 95% highest posterior density interval.

1.7 Concluding Remarks

This chapter develops a sample selection model for discrete or mixed continuous-discrete outcomes with multiple outcome and selection equations. To facilitate the estimation of a resulting multivariate probit model, a Bayesian reformulation in terms of latent variables is extended from the Chib and Greenberg (1998) paper that offers a convenient simulation procedure aimed at resolving the problems of evaluating the integral of multivariate normal density by classical methods. The essence of the method is to jointly simulate the parameters and the latent variables from conditional posterior distributions using a Markov Chain Monte Carlo algorithm. If there is any unobserved heterogeneity for each agent t, it is properly accounted for as a part of the disturbance terms by the covariance structure of the variance matrix resulting from a joint estimation of a system of equations.

[23] To obtain some of the statistics I used the MATLAB program *momentg.m* by James LeSage.

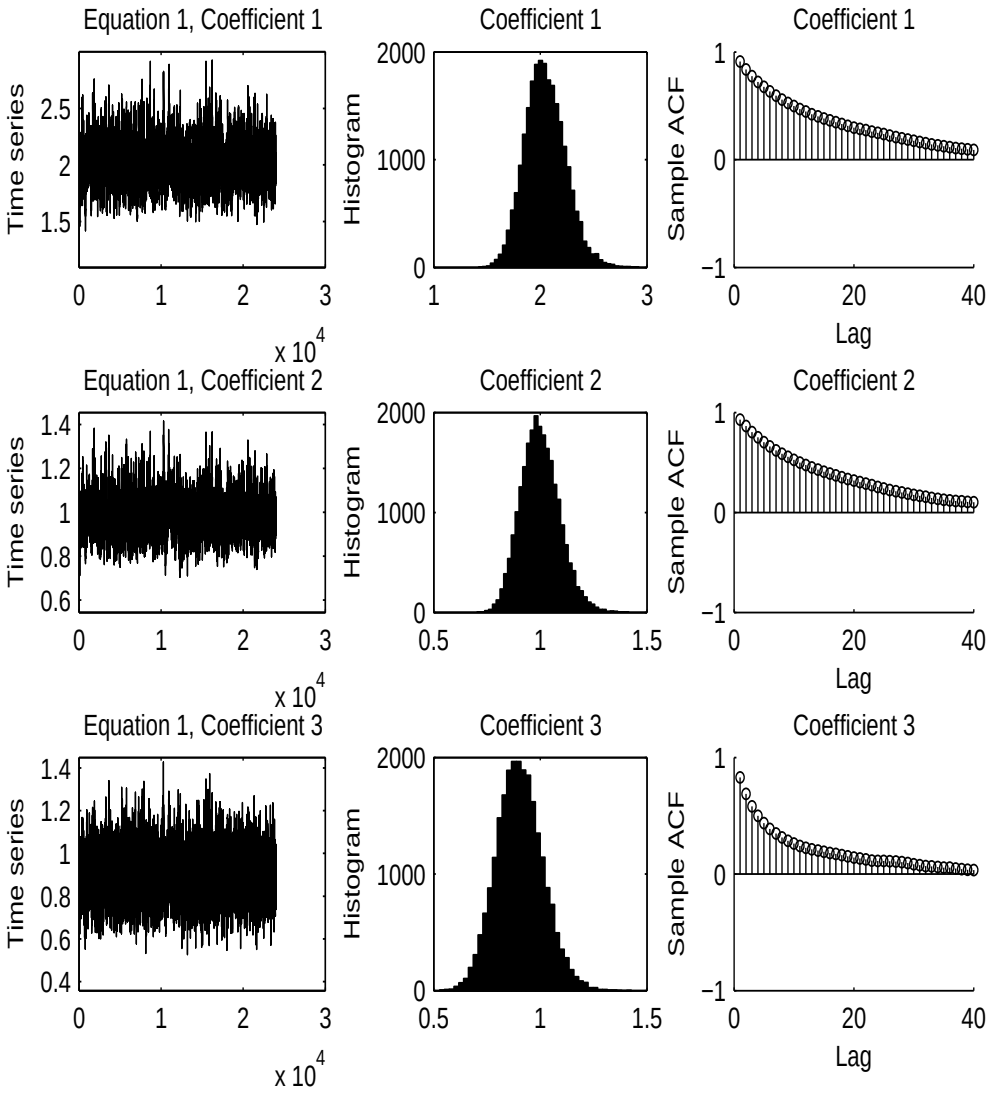

Figure 1.1: Posterior distributions — selection equation.

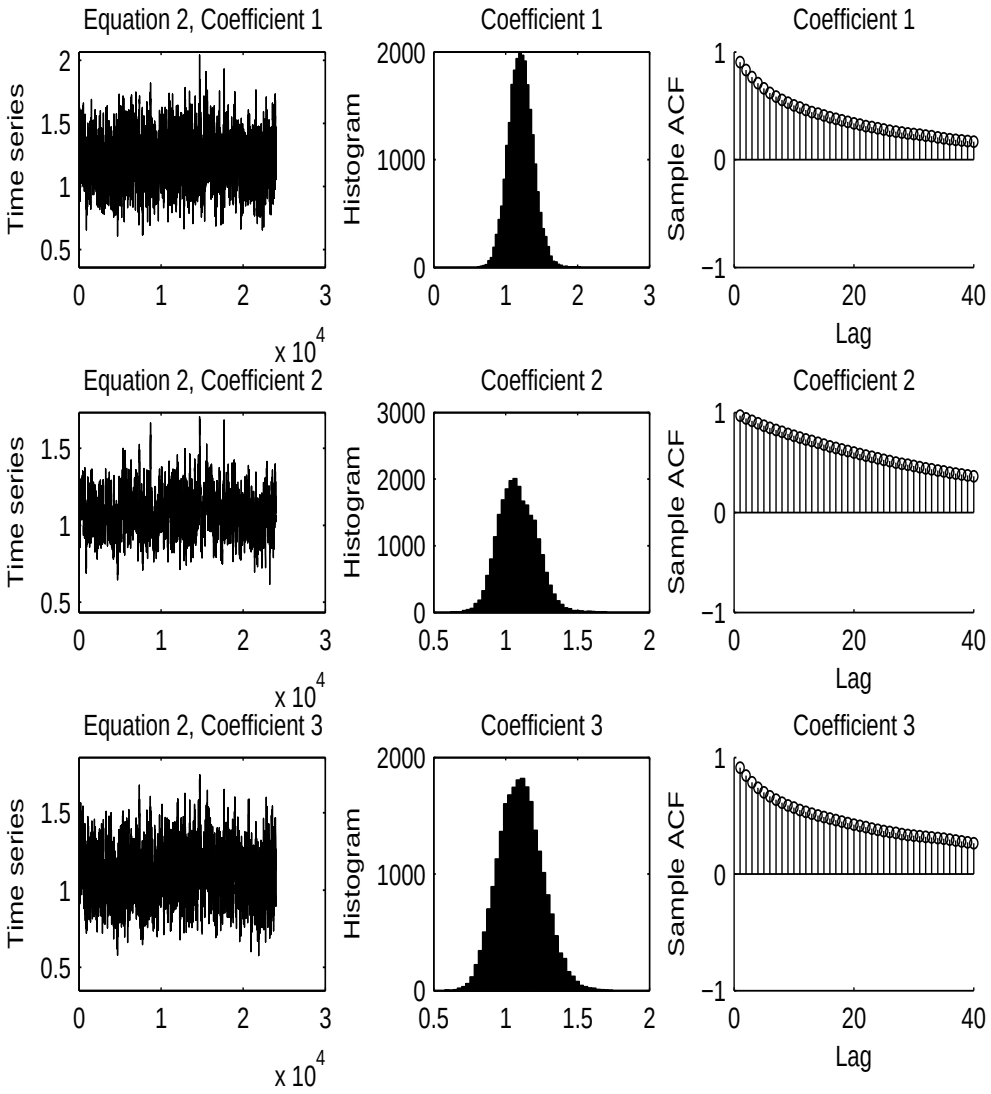

Figure 1.2: Posterior distributions — stroke equation.

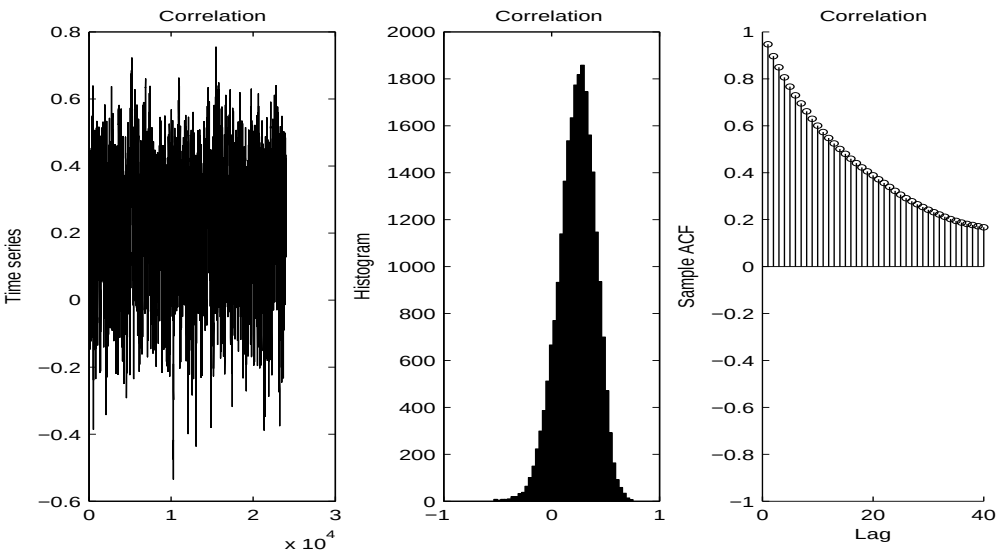

Figure 1.3: Posterior distribution — correlation coefficient.

This chapter also makes two technical advances to the Chib and Greenberg (1998) setup by (i) adding some missing binary responses and (ii) simplifying the estimation of the variance matrix via a multivariate normal representation of the elements in the lower triangular matrix from the Cholesky factorization of Σ^{-1}. I also discuss how the results on identification in Meng and Schmidt (1985) apply in the bivariate probit model with sample selection.

In addition to introducing the multivariate probit model with sample selection, this chapter also offers some interesting topics for further research. In particular, it might be of interest to further study the identification in the case of three and more equations, which clearly depends on the selection rule into a sample. The likelihood is different in each particular case and extensive study of this topic along the lines of Meng and Schmidt (1985) may be rewarding. Alternatively, some of the potentially

Table 1.1: Statistics based on posterior distribution

Coefficient	True Value	Posterior mean	Posterior std	NSE	CD
$\beta_{1,1}$	2	2.0470	0.1902	0.0068	-0.6033
$\beta_{1,2}$	1	0.9994	0.0912	0.0034	-0.5040
$\beta_{1,3}$	1	0.9036	0.1090	0.0032	-0.5625
$\beta_{2,1}$	1	1.2122	0.1707	0.0076	-0.3340
$\beta_{2,2}$	1	1.0834	0.1299	0.0076	-0.2923
$\beta_{2,3}$	1	1.1042	0.1509	0.0075	-0.3694
ρ	0.5	0.2332	0.1709	0.0069	-0.3500

Note: "True Value" — stands for the true value, "Posterior std" — posterior standard deviation, "NSE" — numerical standard error (4% autocovariance tapered estimate), "CD" — test statistics for Geweke's convergence diagnostics.

interesting topics in empirical health and labor economics outlined in the introduction can be done with little (or no) modification of the model in this chapter.

CHAPTER 2
SAMPLE SELECTION AND THE PROBABILITY OF STROKE AMONG THE OLDEST AMERICANS

2.1 Introduction

In this chapter I apply the multivariate probit model of sample selection developed in the first chapter to study the risk factors associated with stroke among the oldest Medicare-eligible Americans. The problem of missing data often complicates empirical work based on survey data in health economics and other social sciences. From a theoretical perspective, statistical methods typically assume that all the information is available for the observations included in the sample and, as such, the majority of statistical textbooks have little to say about how to deal with missing data (Allison 2001). From an empirical point of view, most survey data sets are characterized by global non-response when respondents refuse to participate, missing item-specific information for a particular respondent or attrition when some respondents are lost over time.

Given the severity of the issue, it is not surprising that many different methods have been developed to mitigate the problem of missing data. The choice of appropriate methods depends on assumptions of the underlying missing data mechanisms. For the purposes of this chapter, missing data mechanisms can be defined by the following three classes: data *missing completely at random*, data *missing at random* and *models of sample selection*.[1]

[1]Data are said to be *missing completely at random* if the probability of missing data on any variable is independent of the value of any variable in the data set. In this case it is sufficient to exclude any observations with missing data and estimate the model by any of

In this chapter I use the data from the Survey on Assets and HEAlth Dynamics among the Oldest Old (AHEAD). AHEAD is a large and nationally representative sample of Americans 70 years old or older at the time of their baseline interviews in 1993-1994. In 1998 AHEAD was merged with the Health and Retirement Study (HRS) to provide a common longitudinal data set (Leacock 2006). The HRS is sponsored by the National Institute of Aging (grant number NIA U01AG009740) and is conducted by the University of Michigan.

The issue of selection came up as a part of a larger study of health services use by the oldest old Americans.[2] Professor Wolinsky and his colleagues at The University of Iowa are among a handful of research groups that are approved to link the HRS/AHEAD survey data to a restricted data set of respondents' Medicare claims. This gives me a unique possibility to work with a very rich data set not available to many other researchers.

This data set is particularly convenient for research purposes as it allows an investigator to identify changes in the health status of Medicare eligible respondents, from their Medicare claims, for up to 12 years after the baseline interviews in 1993-1994. I broadly follow Wolinsky *et al.* (2009) in the way of defining selection of AHEAD respondents into the sample used in the empirical analysis. First of all, the AHEAD analytic sample is limited to participants that allowed access to their Medicare claims. Medicare claims are used to find out if a person has experienced

---

the available methods.

[2]National Institute on Aging: Health and Health Services Use in the HRS/AHEAD, NIH grant R01 AG-022913 to Professor Fredric Wolinsky.

stroke after the baseline in 1993-1994. The second restriction is that a person should not be in managed care at the baseline (and she is censored out if she enrolls past baseline). Selection occurs here because managed care plans do not have the same data reporting requirements as fee-for-service Medicare plans. Thus, selection may be a serious issue when dealing with the AHEAD data as only 5,983 respondents out of 7,367 who have complete data on independent variables meet both selection criteria.[3]

In the current application to health economics, it is of interest to model the risk factors associated with the probability of stroke among the AHEAD respondents prior to death or enrollment into managed care in up to 12 years after the baseline. This morbid event places a substantial burden on elderly Americans and identifying the key risk factors should reduce this burden as it informs health care professionals about specific prevention steps that can be targeted. The occurrence of stroke can be verified only if a person allowed linkage to her Medicare claims and she is not enrolled into managed care. If either of these two conditions is violated for respondent t, then the data on whether a health event occurred is missing. In order to obtain consistent coefficient estimates it is necessary to account for the missing data in the AHEAD analytic sample.

One way to proceed is to assume that data *are missing at random* and apply one of propensity score, multiple imputation or maximum likelihood (ML) methods

[3]Wolinsky *et al.* (2009) also exclude proxy respondents because they do not have survey data on their cognitive status. Proxy status does not prevent a researcher from obtaining the information on stroke occurrence. In addition, the total number of people in the AHEAD data set is slightly higher (7,447 respondents) in Wolinsky *et al.* (2009). Respondents that have some missing independent variables (80 people) were excluded from the sample in the current chapter.

to account for this type of missing data.[4]

This chapter uses re-weighting of observations based on propensity scores as in Wolinsky *et al.* (2009). The estimated probabilities of inclusion into the analytic sample can be obtained for all 7,367 AHEAD respondents from a multivariate logistic regression. The predicted probabilities are then divided into deciles and the average participation rate P (i.e., the percent of respondents in the analytic sample in each decile) is determined. The inverse of participation rate can now be used to re-weight the observations in the probit equation for stroke that is estimated only for 5,983 respondents in the analytic sample. This procedure accounts for ignorable selection if data are missing at random because it gives higher weight to participants that are more similar to those who are not included.[5]

It turns out that accounting for ignorable selection by using this propensity score method makes almost no difference in significance of the coefficients in the probit equation for the probability of stroke among the oldest old Americans.[6] It might be hard to say *a priori* how realistic is the assumption of data missing at random. This calls for some alternative selection mechanism, the underlying assumption of which

---

[4]Data are *missing at random* when the probability of missing data on variable Y is independent of the value of Y after controlling for all other variables X in the analysis $P(Y\ missing|Y, X) = P(Y\ missing|X)$. If, in addition, parameters governing missing data process are unrelated to the parameters that are estimated, the missing data mechanism is ignorable (Allison 2001).

[5]This is just one of a myriad of propensity score methods with some others reviewed in D'Agostino (1998), Rosenbaum and Rubin (1983) and Rubin (1979).

[6]This follows from comparing the point estimates in the univariate probit models estimated by maximum likelihood with and without reweighting by propensity scores.

may be tested.

In the first chapter of this dissertation I consider an alternative *sample selection model* using the multivariate probit setup in Chib and Greenberg (1998). This model extends Heckman's (1979) classic *Econometrica* paper on sample selection to the case when the outcome variable of interest is discrete. The Bayesian model in the first chapter extends the multivariate probit setup in Chib and Greenberg (1998) as it (i) permits missing outcome variables in the outcome equation and (ii) simplifies the parameterization of the variance matrix. The joint posterior distribution is obtained from combining the priors and *augmented* likelihood function based on *latent* (unobserved) variables. The simulation from this posterior distribution is made by means of *Gibbs* sampler, as described in the first chapter of the dissertation.[7]

I implement the model using g-prior (Zellner 1986) and perform prior predictive analysis as described in Geweke (2005) to learn if the model and the priors impose any unreasonable restrictions on the outcome. Prior predictive analysis indicates that the parameterization in the sample selection model allows virtually any outcome in the stroke equation. This implies that the prior is non-restrictive and the model is adequate for current purposes. The MCMC algorithm demonstrates good convergence properties, which makes it applicable in other empirical studies with a binary outcome variable. The sample selection model also does not indicate serious selection issues in the AHEAD data, which is consistent with my earlier model based

[7]Gibbs sampler iteratively draws from the full set of conditional distributions that have recognizable form conditional on the draws obtained in the previous run of the sampler. Geweke (2005) provides extensive explanation on this topic.

on propensity scores. Thus, relaxing the assumption from ignorable to non-ignorable selection does not detect additional sources of selectivity from unobserved variables.

This chapter is organized as follows. Section 2.2 introduces the AHEAD data set and the analytic sample as well as defines dependent and explanatory variables. Section 2.3 lays out the propensity score method based on the assumption of data missing at random and reports the results of univariate probit estimations for the observed AHEAD subsample using different weights. Section 2.4 deals with the prior predictive analysis and reports the results from the multivariate probit model. The last section concludes the discussion.

2.2 The Probability of Stroke in the AHEAD Data

In the current application I study the risk factors associated with the probability of stroke in the presence of possible sample selection in the AHEAD data set. The Asset and HEAlth Dynamics among the Oldest Old (AHEAD) started as a distinct data survey in 1993 and was merged in 1998 with the Health and Retirement Study (HRS) to provide a common longitudinal data set. The original AHEAD cohort can be identified in the current HRS data set, and it includes Americans born between 1890 and 1923 who have been interviewed in 1993, 1995, 1998 and every two years thereafter as a part of the present HRS study. As noted in Servais (2004, p. 1),

> "The study paints an emerging portrait of an aging America's physical and mental health, insurance coverage, financial status, family support systems, labor market status, and retirement planning."

Stroke among older Americans is a frequent and severe health event that often has devastating health consequences. Wolinsky *et al.* (2009) cite the following facts about the severity of the effects that stroke has on the health and assets of older Americans: (i) 780,000 Americans experienced stroke (first-ever or recurrent) in 2005; (ii) 150,000 people died from their stroke, making stroke the third leading cause of death in the US; (iii) a mean lifetime cost of a stroke is projected to reach $140,000 per stroke patient.[8] The first step in reducing the burden of this health event lies in identifying the stroke risk factors so that the necessary intervention points can be targeted.

The model of sample selection with dichotomous dependent variables developed in the first part of this thesis is applied to the sample selection equation and outcome equation (whether a respondent has had a stroke). Respondent t is selected into the sample if (i) she has allowed access to her Medicare claims and (ii) she has not been enrolled into managed care at the baseline interview in 1993-1994. Medicare claims are used to identify whether stroke occurred after the baseline. Managed care plans do not have the same data reporting requirements as fee-for-service Medicare plans. If either condition is violated, the data on stroke occurrence is missing because

[8]For the sources of these and other stroke related facts please refer to Wolinsky *et al.* (2009).

there is no reliable way of identifying whether this morbid event has happened. The AHEAD data are characterized by the following outcomes for $T = 7,367$ respondents who have complete data on independent variables:

Sample selection equation. Is respondent selected into sample?

$$y_{1t} = \begin{cases} \text{Yes} = 1 & \text{Medicare claims and not in managed care (5,983 respondents),} \\ \text{No} = -1 & \text{If either condition violated (1,384 respondents).} \end{cases}$$

The outcome (whether stroked occurred) is observed only for the $5,983$ respondents that are selected into a sample:

Outcome equation. Has a stroke occurred to respondent t after the baseline interview?

$$y_{2t} = \begin{cases} \text{Yes} = 1 & \text{if a stroke occurred (606 respondents),} \\ 0 & \text{missing, if the occurrence cannot be verified (1,384 respondents),} \\ \text{No} = -1 & \text{if a stroke did not occur (5,377 respondents).} \end{cases}$$

This work follows the Wolinsky *et al.* (2009) definition of high sensitivity low specificity stroke (minimal false negatives but excessive false positives) as an indicator for the occurrence of a stroke.[9]

The full AHEAD sample in Wolinsky *et al.* (2009) includes 7,447 respondents, 80 of whom were excluded here due to some missing independent variables.[10]

[9]Throughout this study of stroke I borrow the data definitions from the paper by Wolinsky *et al.* (2009).

[10]These include 77 observations with missing body mass index and 3 observations with missing smoker status. One way to extend the current model is by endogenizing missing independent variables.

The independent variables can be organized in the following broad categories as in Wolinsky *et al.* (2009):

· Sociodemographic factors — age, gender, race, marital status

· Socieconomic factors — education, income, number of health insurance policies, neighborhood safety

· Residence characteristics — population density, region of the US, type of residence

· Health behavior factors — body mass index, smoking and drinking history

· Disease history — whether the respondent was diagnosed to have a health condition prior to the baseline interview, the number of doctor visits in the past year

· Functional status — self-rated health, number of difficulties with Activities of Daily Living (ADLs) and Instrumental Activities of Daily Living (IADLS).[11]

The exact definitions of the independent variables, as well as their means and standard deviations, are reported in Appendix B.1.

2.3 Results of Univariate Probit Estimation

2.3.1 Propensity Score Method

Before considering the results of the multivariate sample selection model it is worthwhile to describe the results under the alternative assumption of data missing at random. In particular, three univariate probit equations are estimated for the same

[11]In some studies the cognitive status factors are also included in the analysis, which are observed only for self-respondents. This chapter deals only with missing endogenous variables, so these variables are not considered. An extension to the current model may endogenize missing independent variables. Proxy status does not prevent the researcher from observing the occurrence of stroke.

independent variables as in the multivariate case, but the sample is restricted only to the individuals that have data on stroke occurrence. These equations include unweighted observations, observations weighted by $WTRNORM$, which is the *centered respondent weight* from HRS, and, finally, observations weighted by $WTRNORM/P$, where P is the average participation rate as explained below. $WTRNORM$ adjusts for the unequal probabilities of selection due to the multi-stage cluster and oversampling of African Americans, Hispanics, and Floridians. The minimum weight in my sample is 0.238 and the maximum value is 2.857.[12]

The dependent variable in these models is whether or not a stroke occurred, y_{2t}, which now can take only values of "Yes=1" and "No=-1" and is observed only if $y_{1t} = 1$. The explanatory variables for the stroke equation include all those found to be significant in the Wolinsky *et al.* (2009) paper (with some modifications), as well as some additional variables.

The propensity score is the conditional probability of being assigned to a risk group, given the observed covariates. To estimate the propensity scores, some identifying assumptions about the distribution of the selection binary variable must be adopted. Once the propensity score is estimated, it can be used to reduce bias through matching, stratification, regression adjustment or some combination of the above (D'Agostino 1998, Rosenbaum and Rubin 1983, Rubin 1979).

This chapter follows Wolinsky *et al.* (2009) in the way of using propensity

[12]The distribution of $WTRNORM$ is skewed to the right with a 5^{th} percentile equal to 0.517 and the 95^{th} percentile of 1.592.

Table 2.1: The results of univariate probit for stroke (unweighted)

Parameter	Coefficient	Standard Error	χ^2 p-value
Intercept	-2.030	0.368	< 0.001
Age	0.004	0.004	0.354
Men	-0.049	0.052	0.342
African American	0.105	0.068	0.124
Hispanic	0.112	0.104	0.284
Widowed	0.071	0.053	0.180
Divorced/Separated	-0.111	0.109	0.306
Never Married	0.210	0.125	0.093
Religion not Important	0.089	0.072	0.219
Grade School	-0.054	0.057	0.343
College	-0.085	0.057	0.133
Mobile Home	0.014	0.095	0.883
Multiple Story Home	0.085	0.047	0.073
BMI	0.009	0.005	0.072
Diabetes	0.147	0.066	0.026
Heart	0.071	0.051	0.160
Hypertension	0.105	0.046	0.024
Previous Stroke	0.384	0.070	< 0.001
Poor Self-Rated Health	0.075	0.075	0.317
Fair Self-Rated Health	0.096	0.056	0.083
ADL Sum	-0.028	0.030	0.345
IADL Sum	-0.003	0.025	0.910
Picking up a Dime	0.143	0.079	0.069

Note: "Coefficient" stands for the coefficient estimate, "Standard Error" — for the standard error of the estimate and "χ^2 p-value" — for the p-value of the chi-square test that the coefficient is zero.

scores to adjust for ignorable selection. A multivariable logistic regression model of inclusion in the analytic sample is estimated for all 7,367 AHEAD participants that have no missing explanatory variables. Predictors include all of the available independent variables as well as some interaction terms. The resulting model is used to estimate the predicted probabilities of being in the analytic sample. The predicted probabilities are then divided into deciles and the average participation rate P (i.e., the percent of respondents in the analytic sample in each decile) is determined. The original AHEAD weights $WTRNORM$ are re-weighted by the inverse of participation rate $(1/P)$ and then re-scaled so that the sum of weights equals the number of participants in the analytic sample. This procedure gives greater influence to participants in the analytic sample most like those not included.

2.3.2 The Probability of Stroke

Table 2.1 reports the results of univariate probit with unweighted observations for the stroke equation. The probability of stroke increases for respondents that were never married, living in multiple story home, of those with higher body mass index, of patients with diabetes, hypertension and previous stroke at the baseline, for people that reported fair self-rated health and having difficulty picking up a dime. These results somewhat differ from findings in a recent paper by Wolinksy *et al.* (2009) because I used slightly different definitions of the independent variables.

The results of univariate probit with observations reweighted by propensity score are given in the last three columns of Table 2.2. In terms of significant predictors of stroke there are certain differences with the case of unweighted probit. Widowed

respondents and patients with prior heart disease are more likely to have a stroke, while body mass index is no longer a significant predictor. The reader might think that accounting for missing data by using propensity scores changed the results. In fact, all the differences come from using HRS weight $WTRNORM$ as the first three columns in Table 2.2 indicate. Indeed, the same risk factors remain significant and even their estimates are very close. It seems that the HRS team did a really good job in terms of developing the $WTRNORM$ weight and there is virtually no difference if those weights are adjusted by propensity scores.

Judging from those comparisons it follows that accounting for ignorable selection in the AHEAD data does not substantially affect the estimates in the univariate probit equation for the probability of a stroke. The next model I consider — the bivariate probit model with sample selection — is based on a less restrictive assumption of non-ignorable selection.

2.4 Sample Selection Model

The multivariate probit model with sample selection is developed in the first chapter of this dissertation. The Gibbs sampler is run over the full conditional set of posterior normal distributions for the coefficient vector β, the element of variance matrix decomposition F and the multivariate truncated normal distribution for each respondent t. The explanatory variables in the stroke equation are the same as in the univariate probit models above. It is believed that socioeconomic characteristics and place of living affect the probability of being selected into a sample. Functional status variables are included in both equations. The convergence properties of the

Table 2.2: The results of univariate probit for stroke using two weights

Model Weights	$WTRNORM$			$WTRNORM/P$		
Parameter	Coef.	St.er.	χ^2 p-value	Coef.	St.er.	χ^2 p-value
Intercept	-1.973	0.375	< 0.001	-1.865	0.374	< 0.001
Age	0.003	0.004	0.566	0.001	0.004	0.823
Men	-0.035	0.053	0.502	-0.030	0.053	0.570
African American	0.107	0.077	0.162	0.113	0.076	0.136
Hispanic	0.138	0.122	0.255	0.132	0.119	0.267
Widowed	0.104	0.055	0.056	0.119	0.055	0.029
Divorced/Separated	-0.011	0.111	0.918	-0.004	0.110	0.968
Never Married	0.242	0.122	0.046	0.238	0.122	0.051
Religion not Important	0.115	0.072	0.108	0.097	0.071	0.176
Grade School	-0.064	0.059	0.277	-0.065	0.059	0.269
College	-0.069	0.056	0.220	-0.063	0.056	0.259
Mobile Home	-0.004	0.100	0.970	0.013	0.098	0.896
Multiple Story Home	0.111	0.047	0.018	0.118	0.047	0.013
BMI	0.008	0.005	0.102	0.008	0.005	0.120
Diabetes	0.162	0.067	0.016	0.159	0.067	0.017
Heart	0.097	0.051	0.058	0.101	0.051	0.048
Hypertension	0.112	0.047	0.017	0.119	0.047	0.011
Previous Stroke	0.354	0.072	< 0.001	0.349	0.072	< 0.001
Poor Self-Rated Health	0.062	0.078	0.425	0.051	0.078	0.513
Fair Self-Rated Health	0.131	0.056	0.019	0.129	0.056	0.020
ADL Sum	-0.022	0.031	0.479	-0.017	0.030	0.571
IADL Sum	-0.016	0.027	0.546	-0.015	0.026	0.558
Picking up a Dime	0.150	0.079	0.059	0.160	0.080	0.046

Note: "Coef." stands for the coefficient estimate, "St.er." — for the standard error of the estimate and "χ^2 p-value" — for the p-value of the chi-square test that the coefficient is zero. The first three columns use $WTRNORM$ and the last three columns use $WTRNORM/P$ as weights.

correlation coefficient are not affected even if the same set of variables is used in both equations.

2.4.1 Prior predictive analysis

Before the estimation can proceed, it is necessary to set the prior hyperparameters for the vector of coefficients β as well as F used for variance decomposition. One way to achieve this is by means of prior predictive analysis, as described in Geweke (2005). The purpose of prior predictive analysis is to ascertain the prior distribution of functions of interest that are relevant to the problem.

In the current setup, prior predictive analysis can accomplished by means of forward simulation for iterations $n = 1, ..., N$ from the prior distributions of parameters

$$\begin{aligned} \beta^{(n)} &\sim N(\underline{\beta}, \underline{B}^{-1}) \\ F^{(n)} &\sim N(\underline{F}, \underline{H}^{-1}) \end{aligned} \tag{2.1}$$

as well as the data $y_{.t}^{(n)}$ for respondents $t \leq T$. To obtain the latter the latent data $\widetilde{y}_{.t}^{(n)}$ are first drawn from the untruncated multivariate normal distribution

$$\widetilde{y}_{.t}^{(n)} \sim N\big(Z_t\beta^{(n)}, ([F^{(n)}]')^{-1}(F^{(n)})^{-1}\big). \tag{2.2}$$

After that, the latent data $\widetilde{y}_{.t}^{(n)}$ are converted to binary data $y_{.t}^{(n)}$, taking into account the selection rule into the sample. That is, if $y_{1t}^{(n)}$ is equal to "-1" (i.e. Medicare claims are not available for respondent t or she has been enrolled in managed care at the baseline) then $y_{2t}^{(n)}$ (if stroke occurred) is missing and set to zero.

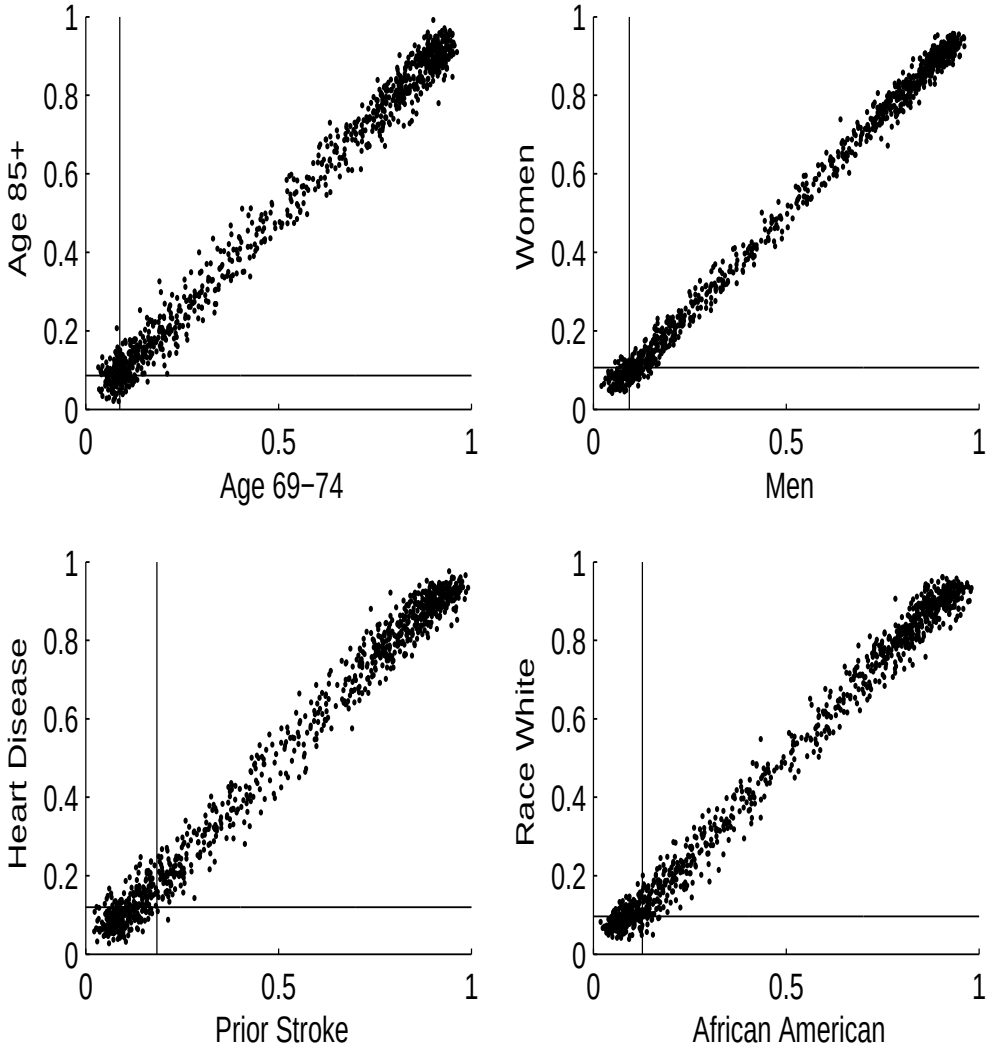

Figure 2.1: Proportion of stroke from prior predictive analysis

In each panel the scatterplot shows the proportion of stroke in different risk groups for 1000 independent draws from the prior — the observed value is the intersection of the vertical and horizontal lines. For each of 1000 draws from the prior, I generate the sample of 7367 artificial observations of the dependent variables. Each point represents the sample statistic in one of those 1000 samples.

After repeating the forward simulation N times, it is possible to look at the prior distribution of functions of interest $h(y)$. Forward simulation can reveal de-

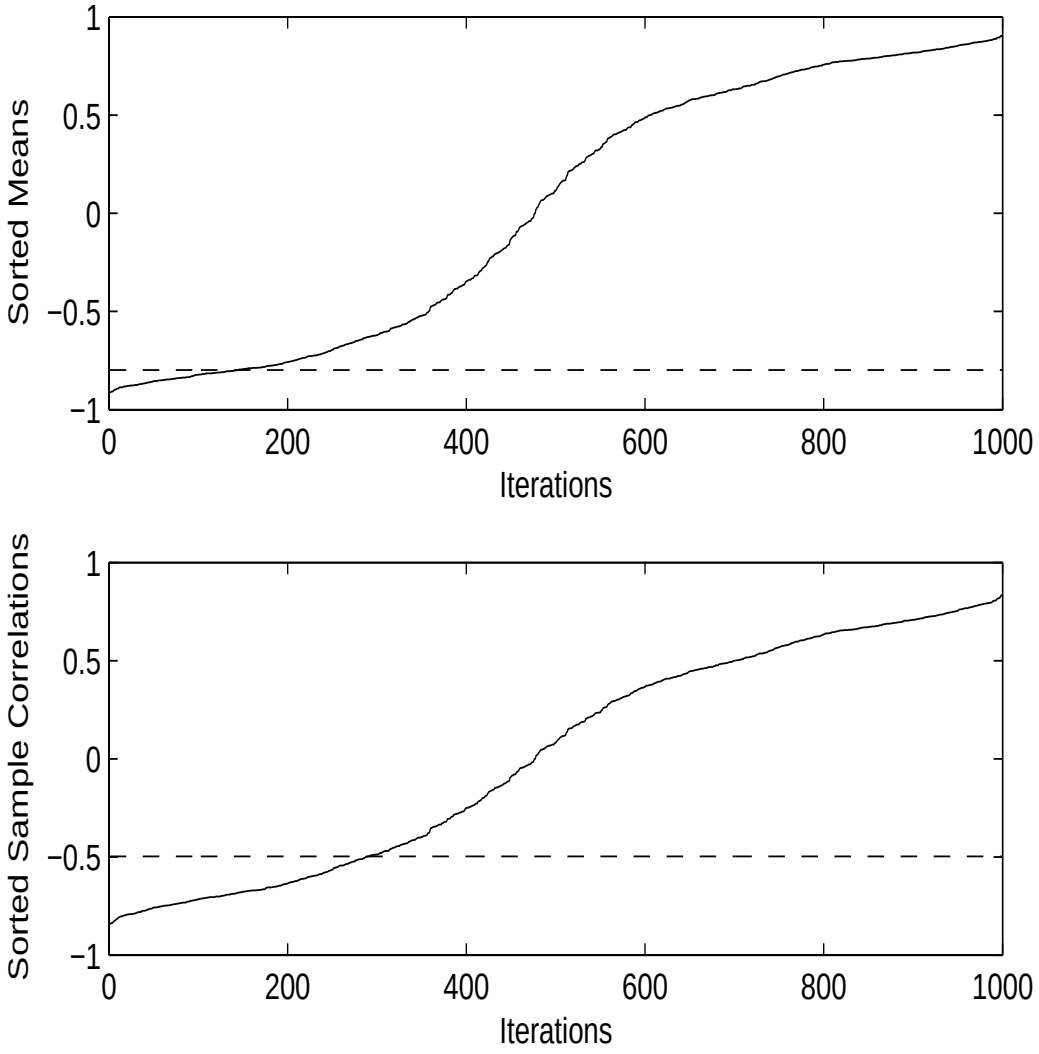

Figure 2.2: Means and correlation coefficients from prior predictive analysis

Top panel shows the sorted means for the observed probability of stroke in each of 1000 iterations sorted in ascending order. Bottom panel shows the sample correlation coefficients with responses converted into -1,0 and 1 sorted in the ascending order. The horizontal line represents the observed value in both cases. For each of 1000 draws from the prior, I generate the sample of 7367 artificial observations of the dependent variables. Each point represents the sample statistic in one of those 1000 samples.

ficiencies in the model if the distribution of $h(y)$ coincides poorly with the prior beliefs about this function. In general, prior predictive analysis interprets the model specification in terms of observables that are usually easier to understand than the parameters themselves (Geweke 2005). If there any deficiencies in the model, then prior hyperparameters may be adjusted or a new specification can be chosen.

A set of prior hyperparameters should not be informative in the sense that it should not favor any particular outcome *a priori*. On the contrary, prior should allow for a wide array of reasonable outcomes without ruling out any result that is remotely plausible. Prior should not also impose any unreasonable restrictions on the distributions of functions of interest $h(y)$. For example, if the prior distribution of the probability of stroke assigns a 100% chance of a stroke for females in all N simulations, then the prior hyperparameters or the model are not adequate and must be changed. On the other hand, if in some proportion of simulations all, some or no females experience a stroke, then the model and the prior parameters can be adopted. In the current setup it is a prior belief that, conditional on the set of exogenous variables, there is a reasonable probability that the outcome could be either "-1" or "+1" for virtually any respondent in the sample.

The model hyperparemeters are set using g-prior (Zellner 1986) for the variance matrix $\underline{B}^{-1}$. In particular, $\underline{B}^{-1}$ is block diagonal with each block $i = 1, 2$ defined as $g_B \cdot (Z_i' Z_i)^{-1}$ of the corresponding dimension $k_i \times k_i$. In this expression Z_i is a $T \times k_i$ set of explanatory variables in equation i. The prior variance for F is set equal to $\underline{H}^{-1} = g_H \cdot I_{m(m-1)/2}$. One simple reason to use this prior is to reduce a

rather large number of parameters to only three hyperparameters which facilitates prior predictive analysis and estimation. A more subtle and fundamental reason is to ensure the sensible magnitudes of variances for continuous *versus* binary covariates that constitute the majority in the data set.

I set the prior means for β and F to 0. When the prior is specified as above, the problem is reduced to choosing only two hyperparameters g_B and g_H. I found that with $g_B = 5$ and $g_H = 9$, this g-prior is relatively non-informative in the sense that it allows for virtually any plausible outcome for the functions of interest $h(y)$. Three functions of interest $h(y)$ are considered: the proportion of stroke occurrences in different risk groups, correlation coefficient for the discrete outcomes $y_{.t}^{(n)}$ in two equations and the mean predicted value of $y_{.t}^{(n)}$ in the stroke equation.

All the results that follow are the *prior* probabilities based on 1000 simulations from the prior distributions. Again, the prior should not rule out any plausible outcome implying that each function $h(y)$ should have prior range that is wide enough to incorporate the observed outcome as well as almost any other reasonable value. Figure 2.1 shows that the proportion of strokes in various risk groups takes values in the range from almost 0 to almost 1. Similarly, Figure 2.2 indicates that the prior (i) allows sufficient variability in the average probability of stroke in the sample and (ii) is consistent with a variety of correlation patterns between the observed outcomes. Table 2.3 shows that the empirical 95% interval from the prior distribution ranges from almost 0 to almost 1 in various risk groups, but always includes the observed proportion.

Table 2.3: The prior probability of stroke in various risk groups

Parameter	Observed $h(y^o)$	$P^{-1}[h(y^o)]$	2.5 %	97.5 %
Age 69-74	0.089	0.105	0.059	0.939
Age 85+	0.086	0.087	0.054	0.945
Men	0.094	0.122	0.055	0.937
Women	0.107	0.159	0.060	0.939
Prior Stroke	0.185	0.282	0.054	0.956
Heart Disease	0.120	0.185	0.057	0.938
African American	0.127	0.195	0.050	0.945
Race White	0.096	0.137	0.061	0.939

Note: The second column represents the observed proportion of stroke in the corresponding risk group. The third column is the fraction of 1000 iterations from the prior distribution that were less than the observed proportion $h(y^o)$. The last two columns are correspondingly the values of $h(y)$ that leave 2.5% and 97.5% of 1000 iterations below.

These figures and table show that the outcome observed in the AHEAD data is well accommodated by the selected prior. It is also the case that the prior is not restrictive, as it is consistent with a variety of other plausible outcomes.

2.4.2 Results of multivariate probit

The results reported herein are based on 50,000 Gibbs iterations (after dropping the first 20% burn-in iterations).[13]

The results do not show any problems with stability if the Gibbs sampler is

[13]I describe the implementation of the algorithm in the first chapter. It takes between 7.4 and 7.8 seconds to obtain 1,000 draws from the Gibbs sampler when I use MATLAB 7.6.0 (R2008a) with a 64-bit Windows Vista operational system. I use Dell Precision workstation with a dual-quad core processor Intel(R) Xeon (R) E5430 @ 2.66 GHz and a Memory (RAM) of 8.00 GB.

run 500,000 times.[14] Consider Table 2.4 with the F parameter and the corresponding ρ coefficient which is constructed as

$$\rho(j) = \frac{-F(j)}{\sqrt{1 + F(j)^2}} \tag{2.3}$$

for each iteration $j = 1, ..., 60,000$.[15] The t-ratio reported in column 4 does not have the same interpretation as in classical econometrics but it can be used as a quick guidance on whether the coefficient's highest density posterior interval (HDPI) contains zero.

Table 2.4: The results of 50,000 Gibbs draws for F and ρ

Parameter	pmean	pstd	t-ratio	Geweke's CD
F	0.8171	0.0433	18.8907	0.3694
ρ	-0.6321	0.0201	-31.4981	-0.3609

Note: "pmean" and "pstd" stand for the posterior mean and standard deviation of the sample from 50,000 Gibbs draws (not including 20% initial burn-in draws), "t-ratio" is their ratio and "Geweke's CD" stands for Geweke's (1992) convergence diagnostic statistics.

Tables 2.5 and 2.6 report the results for the coefficients in the stroke and selection equation correspondingly.[16] The Geweke's (1992) convergence diagnostic test does not indicate any convergence problems in any of the coefficients in the two

[14]The results are not reported, in order to save space, but are available upon request.

[15]The formula for ρ can be obtained from the variance matrix given my parameterization.

[16]The coefficients in the first (selection) equation are normalized by $\sqrt{1 + F^2(j)}$ in each draw $j = 1, ..., 60,000$ to have variances comparable with the second equation.

Table 2.5: The results for β coefficients (stroke equation)

Parameter	pmean	pstd	t-ratio	Geweke's CD
Intercept	-0.841	0.239	-3.523	0.570
Age	0.001	0.003	0.464	-0.030
Men	-0.028	0.033	-0.839	-0.777
African American	0.060	0.046	1.325	-0.637
Hispanic	0.085	0.067	1.261	0.347
Widowed	0.041	0.034	1.207	0.307
Divorced/Separated	-0.031	0.068	-0.462	-0.424
Never Married	0.104	0.084	1.227	-0.178
Religion not Important	0.061	0.047	1.305	-0.292
Grade School	-0.023	0.038	-0.609	-0.833
College	-0.038	0.037	-1.022	-0.661
Mobile Home	0.018	0.061	0.299	0.851
Multiple Story Home	0.041	0.031	1.320	0.391
BMI	0.003	0.003	1.030	-1.271
Diabetes	0.076	0.045	1.709	-0.276
Heart	0.029	0.033	0.880	-0.390
Hypertension	0.046	0.030	1.513	0.873
Previous Stroke	0.196	0.049	3.988	-0.055
Poor Self-Rated Health	0.025	0.050	0.497	0.917
Fair Self-Rated Health	0.043	0.036	1.192	-0.564
ADL Sum	-0.014	0.021	-0.675	-0.896
IADL Sum	0.011	0.017	0.640	0.255
Picking up a Dime	0.066	0.054	1.231	-0.173

Note: "pmean" and "pstd" stand for the posterior mean and standard deviation of the sample from 50,000 Gibbs draws (not including 10,000 initial burn-in draws), "t-ratio" is their ratio and "Geweke's CD" stands for Geweke's (1992) convergence diagnostic statistics.

equations. It appears that the multivariate probit model with sample selection works well in terms of its convergence properties.

A closer look at the posterior means in Table 2.5 shows that they are often about half of the mean obtained in Table 2.1 for the coefficients significant at the 10% level. This indicates the strong prior centered at zero — it places the posterior means

Table 2.6: The results for β coefficients (selection equation)

Parameter	pmean	pstd	t-ratio	Geweke's CD
Intercept	0.359	0.213	1.683	-0.282
Grade School	-0.018	0.031	-0.557	1.094
College	0.035	0.032	1.107	0.931
Income Zero	0.377	0.259	1.454	-0.696
Log of Income	0.031	0.016	1.968	0.053
Home Value Zero	-0.084	0.190	-0.440	0.428
Log of Home Value	-0.007	0.017	-0.384	0.403
# of Health Insurance Policies	0.049	0.022	2.231	-0.649
Long Term Care Insurance	-0.045	0.039	-1.132	-0.339
Neighborhood Safety Poor/Fair	0.006	0.036	0.174	-0.717
Population over 1,000,000	-0.184	0.027	-6.751	0.061
Northeast region of US	0.034	0.036	0.950	-1.596
North Central region of US	0.020	0.032	0.627	-0.025
West region of US	-0.028	0.039	-0.722	-0.075
ADL Sum	0.002	0.016	0.150	-1.481
IADL Sum	-0.031	0.014	-2.296	0.292
Fall	0.029	0.029	1.010	-1.225

Note: "pmean" and "pstd" stand for the posterior mean and standard deviation of the sample from 50,000 Gibbs draws (not including 10,000 initial burn-in draws), "t-ratio" is their ratio and "Geweke's CD" stands for Geweke's (1992) convergence diagnostic statistics.

approximately half the way from the estimates obtained using only the data. This observation is confirmed by the prior sensitivity analysis. As I start relaxing the prior by setting g_B to 10, 100 and 1,000, I find that the posterior mean of F decreases to 0.633, 0.439 and 0.337 correspondingly. The posterior standard deviation increases at the same time to 0.045, 0.089 and 0.157. Table 2.7 shows how the correlation coefficient ρ changes with g_B. The posterior means of the coefficients in the stroke equation are getting closer to those in Table 2.1 as g_B increases. Thus, prior sensitivity analysis reveals that prior drives most of the results even though it is not restrictive,

Table 2.7: The results of 50,000 Gibbs draws for ρ.

Parameter	pmean	pstd	t-ratio	Geweke's CD
$g_B = 1$	-0.9057	0.0082	-110.8968	-2.7330
$g_B = 10$	-0.5341	0.0272	-19.6113	-1.0393
$g_B = 100$	-0.3988	0.0677	-5.8886	-0.8000
$g_B = 1000$	-0.3500	0.1299	-2.6944	-0.4746

Note: "pmean" and "pstd" stand for the posterior mean and standard deviation of the sample from 50,000 Gibbs draws (not including 20% initial burn-in draws), "t-ratio" is their ratio and "Geweke's CD" stands for Geweke's (1992) convergence diagnostic statistics.

as it was shown by the prior predictive analysis.

An interesting question is why the prior plays such an important role in the stroke application. One possible reason is that most of the independent variables are also binary: only age, body-mass index and self-reported income measures are continuous and even those are not always important predictors in stroke or selection equation. Experiments with artificial data show that variation in continuous covariates is indeed important for model performance.[17] When I perform principal component analysis for each of the six groups of explanatory variables, I find that the first principal component explains about a quarter of the total variance, which shows the lack of information in the independent variables.

The fact that the sample selection model with binary outcome did not find strong selection effects in the data is consistent with two recent papers by Munkin

[17]This seems to be consistent with the observation in Leung and Yu (1996), who find that the sample selection model works well only when there is enough variation in the independent variables.

and Trivedi (2003) and also Preget and Waelbroeck (2006). This does not undermine the validity of the model, which can be used to test for the presence of selection in other applications in health or empirical economics. Thus, neither propensity score nor sample selection model indicate serious selection issues in the AHEAD data set when applied in the study of stroke predictors.

2.5 Concluding Remarks

This chapter considers two different methods of dealing with the problem of missing binary outcome variable in the context of the stroke occurrence among the oldest Americans. The propensity score model based on the assumption of data missing at random does not generate substantial differences in the significance of the important risk predictors compared to using $WTRNORM$ weight from the HRS. The multivariate probit model with sample selection also does not find any strong correlation in the data when the outcome and selection equations are estimated jointly. Thus, the main substantive contribution of the paper is that there is no evidence of selection in the AHEAD data based on either propensity score or sample selection model when applied in the study of stroke predictors. In addition, this work is the first application of the multivariate probit model of sample selection, developed in the first chapter of the thesis, to the real data set. The model shows reasonable variability in the prior distribution of the stroke occurrence and fast convergence.

APPENDIX A
DERIVATIONS OF RESULTS IN CHAPTER 1

A.1 Conditional Posterior Distributions

This Appendix derives the conditional posterior distributions in the Gibbs sampler. The posterior density kernel is the product of the prior for β, prior for F_{vector} and augmented likelihood for $\widetilde{y}'_{.t}s$

$$
\begin{aligned}
& |\underline{B}|^{1/2} \exp\Big\{ -\frac{1}{2}(\beta-\underline{\beta})'\underline{B}(\beta-\underline{\beta})\Big\} \qquad (A.1)\\
\cdot\ & |\underline{H}|^{1/2} \exp\Big\{ -\frac{1}{2}(F_{vector}-\underline{F}_{vector})'\underline{H}(F_{vector}-\underline{F}_{vector})\Big\}\\
\cdot\ & |\Sigma|^{-T/2}\prod_{t=1}^{T}\exp\Big\{ -\frac{1}{2}(\widetilde{y}_{.t}-Z_t\beta)'\Sigma^{-1}(\widetilde{y}_{.t}-Z_t\beta)\Big\} I(\widetilde{y}_{.t}\ \epsilon\ B_t).
\end{aligned}
$$

The three conditional posterior distributions can be obtained as follows.

(i) The conditional posterior kernel for β can be obtained from equation (A.1) by collecting the terms that contain β and completing the square

$$
\begin{aligned}
p(\beta|\Sigma,\widetilde{y}) \ &\propto\ \exp\Big\{ -\frac{1}{2}(\beta'\underline{B}\beta - 2\beta'\underline{B}\underline{\beta} + \underline{\beta}'\underline{B}\underline{\beta})\Big\} \qquad (A.2)\\
&\quad\cdot\ \prod_{t=1}^{T}\exp\Big\{ -\frac{1}{2}(\widetilde{y}_{.t}\Sigma^{-1}\widetilde{y}_{.t} - 2\beta' Z_t'\Sigma^{-1}\widetilde{y}_{.t} + \beta' Z_t'\Sigma^{-1}Z_t\beta)\Big\}\\
&\propto\ \exp\Big\{ -\frac{1}{2}\Big(\beta'(\underline{B}+\sum_{t=1}^{T}Z_t'\Sigma^{-1}Z_t)\beta - 2\beta'(\underline{B}\underline{\beta}+\sum_{t=1}^{T}Z_t'\Sigma^{-1}\widetilde{y}_{.t})\Big)\Big\}\\
&\propto\ \exp\Big\{ -\frac{1}{2}(\beta-\overline{\beta})'\overline{B}(\beta-\overline{\beta})\Big\},
\end{aligned}
$$

where $\overline{B} = \underline{B} + \sum_{t=1}^{T} Z_t'\Sigma^{-1}Z_t$ is the posterior precision and

$$
\overline{\beta} = \overline{B}^{-1}(\underline{B}\underline{\beta} + \sum_{t=1}^{T} Z_t'\Sigma^{-1}\widetilde{y}_{.t})
$$

is the posterior mean for β.

(ii) The alternative expression for the density of $\widetilde{y}$

$$p(\widetilde{y}|y,\beta,F,D) \;\propto\; \prod_{i=1}^{m}\exp\Big\{-\frac{1}{2}\sum_{t=1}^{T}(\varepsilon_{t,i}+F'_{i+1:m,i}\varepsilon_{t,i+1:m})^2\Big\}, \tag{A.3}$$

is derived in the text. Remembering that $F_{vector}=[F'_{2:m,1},...,F'_{m,m-1}]'$ one can collect the terms in the posterior density kernel (A.1) as

$$\begin{aligned}
p(F_{vector}|\beta,\widetilde{y}) &\propto \prod_{i=1}^{m-1}\exp\Big\{-\frac{1}{2}(F_{i+1:m,i}-\underline{F}_{i+1:m,i})'\underline{H}_i(F_{i+1:m,i}-\underline{F}_{i+1:m,i})\Big\} \qquad (A.4)\\
\cdot\quad & \prod_{i=1}^{m-1}\exp\Big\{-\frac{1}{2}\sum_{t=1}^{T}(\varepsilon_{t,i}+F'_{i+1:m,i}\varepsilon_{t,i+1:m})^2\Big\}\\
\propto\quad & \prod_{i=1}^{m-1}\exp\Big\{-\frac{1}{2}\Big(F'_{i+1:m,i}\underline{H}_iF_{i+1:m,i}-2F'_{i+1:m,i}\underline{H}_i\underline{F}_{i+1:m,i}+\underline{F}_{i+1:m,i}\underline{H}_i\underline{F}_{i+1:m,i}\Big)\\
\cdot\quad & \prod_{i=1}^{m-1}\exp\Big\{-\frac{1}{2}\Big(\sum_{t=1}^{T}\varepsilon_{t,i}^2+2F'_{i+1:m,i}\sum_{t=1}^{T}\varepsilon_{t,i+1:m}\varepsilon_{t,i}\\
& +F'_{i+1:m,i}\big(\sum_{t=1}^{T}\varepsilon_{t,i+1:m}\varepsilon'_{t,i+1:m}\big)F_{i+1:m,i}\Big)\Big\}\\
\propto\quad & \prod_{i=1}^{m-1}\exp\Big\{-\frac{1}{2}\Big(F'_{i+1:m,i}\big(\underline{H}_i+\sum_{t=1}^{T}\varepsilon_{t,i+1:m}\varepsilon'_{t,i+1:m}\big)F_{i+1:m,i}\\
& -2F'_{i+1:m,i}\big(\underline{H}_i\underline{F}_{i+1:m,i}-\sum_{t=1}^{T}\varepsilon_{t,i+1:m}\varepsilon_{t,i}\big)\Big)\Big\}\\
\propto\quad & \prod_{i=1}^{m-1}\exp\Big\{-\frac{1}{2}\Big(F_{i+1:m,i}-\overline{F}_{i+1:m,i})'\overline{H}_i(F_{i+1:m,i}-\overline{F}_{i+1:m,i}\Big)\Big\}
\end{aligned}$$

where $\overline{H}_i=\underline{H}_i+\sum_{t=1}^{T}\varepsilon_{t,i+1:m}\varepsilon'_{t,i+1:m}$ is the posterior precision and

$$\overline{F}_{i+1:m,i}=\overline{H}_i^{-1}\underline{H}_i\;\underline{F}_{i+1:m,i}-\overline{H}_i^{-1}\sum_{t=1}^{T}\varepsilon_{t,i+1:m}\varepsilon_{t,i}$$

is the posterior mean. It is understood that $\underline{H}_i$ is the ith element of the block-diagonal prior precision matrix $\underline{H}$ with dimensions decreasing from $(m-1)\times(m-1)$ for the first block to 1×1 for the last block. The final step is to organize $i=1:m-1$

multivariate normal distributions in the last line of equation (A.4) into one

$$\begin{aligned} p(F_{vector}|\beta,\widetilde{y}) &\propto \prod_{i=1}^{m-1} \exp\Big\{-\frac{1}{2}\Big(F_{i+1:m,i}-\overline{F}_{i+1:m,i})'\overline{H}_i(F_{i+1:m,i}-\overline{F}_{i+1:m,i}\Big)\Big\} \\ &= \exp\Big\{-\frac{1}{2}\Big(F_{vector}-\overline{F}_{vector})'\overline{H}(F_{vector}-\overline{F}_{vector}\Big)\Big\}, \end{aligned} \tag{A.5}$$

which is used in the text. Since $F_{vector} = (F'_{2:m,1}, F'_{3:m,2}, ..., F'_{m,m-1})'$ with $F'_{j+1:m,j} = (f_{j+1,j}, ..., f_{m,j})$ for $j = 1, ..., m-1$ being the vectors under the main diagonal of F

$$F^{(n)} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ f_{21}^{(n)} & 1 & 0 & \cdots & 0 \\ f_{31}^{(n)} & f_{32}^{(n)} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ f_{m1}^{(n)} & f_{m2}^{(n)} & f_{m3}^{(n)} & \cdots & 1 \end{pmatrix}.$$

one can construct the covariance matrix Σ as

$$\Sigma = (F')^{-1}F^{-1}. \tag{A.6}$$

(iii) Finally, the latent data $\widetilde{y}_{.t}$ are drawn for each respondent $t \leq T$ from the truncated multivariate normal distribution as in Geweke (1991) conditional on Z_t, β and F, as well as $\widetilde{y}_{.t}$ obtained in the previous draw. The multivariate normal distribution is truncated to the region defined by the $m \times 2$ matrix $[a, b]$ with a typical row i equal to $(0, \infty)$ if $y_{it} = 1$ and $(-\infty, 0)$ if $y_{it} = -1$. If y_{it} is not observed, then row i is $(-\infty, \infty)$

A.2 Identification

Meng and Schmidt (1985) apply the general principle developed in Rothenberg (1971), that the parameters are (locally) identified if and only if the information

matrix is nonsingular, to the censored bivariate probit model. I extend their result to the sample selection model with one binary selection and one binary outcome equation.

Let $\theta = [\beta_1, \beta_2; f_{21}]$ be the vector of parameters that is used to construct the information matrix

$$I(\theta) = E\Big[\Big(\frac{\partial \ln L}{\partial \theta}\Big)\Big(\frac{\partial \ln L}{\partial \theta'}\Big)\Big]. \tag{A.7}$$

The information matrix in the censored bivariate probit is

$$I(\theta) = C_1' C_1 + C_2' C_2 + C_6' C_6, \tag{A.8}$$

as shown in Meng and Schmidt (1985). In this expression each matrix C_j' for $j = 1, 2, 6$ is of dimension $(2k + 1) \times T$ and

the tth column of C_1' is $\frac{1}{\sqrt{F^t}} \frac{\partial F^t}{\partial \theta}$

the tth column of C_2' is $\frac{1}{\sqrt{\Phi_1^t - F^t}} \frac{\partial [\Phi_1^t - F^t]}{\partial \theta}$

the tth column of C_6' is $\frac{1}{\sqrt{1-\Phi_1^t}} \frac{\partial [1-\Phi_1^t]}{\partial \theta}$.

After the simplification the information matrix $I(\theta)$ takes the form

$$\begin{pmatrix}
\sum_{t=1}^{T}\left(\frac{\partial F^t}{\partial \beta_{1,1}}\right)^2\omega_1^t & & & & & & \\
+\left(\frac{\partial \Phi_1^t}{\partial \beta_{1,1}}\right)^2\omega_2^t & \cdots & I_{(1,k_1)} & I_{(1,k_1+1)} & \cdots & I_{(1,k)} & I_{(1,k+1)} \\
-\frac{2}{\Phi_1^t-F^t}\frac{\partial \Phi_1^t}{\partial \beta_{1,1}}\frac{\partial F^t}{\partial \beta_{1,1}} & & & & & & \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
\sum_{t=1}^{T}\frac{\partial F^t}{\partial \beta_{1,1}}\frac{\partial F^t}{\partial \beta_{1,k_1}}\omega_1^t & & & & & & \\
+\frac{\partial \Phi_1^t}{\partial \beta_{1,1}}\frac{\partial \Phi_1^t}{\partial \beta_{1,k_1}}\omega_2^t-\frac{1}{\Phi_1^t-F^t}\cdot & \cdots & I_{(k_1,k_1)} & I_{(k_1,k_1+1)} & \cdots & I_{(k_1,k)} & I_{(k_1,k+1)} \\
\cdot\frac{\partial \Phi_1^t}{\partial \beta_{1,1}}\frac{\partial F^t}{\partial \beta_{1,k_1}}\frac{\partial F^t}{\partial \beta_{1,1}}\frac{\partial \Phi_1^t}{\partial \beta_{1,k_1}} & & & & & & \\
\sum_{t=1}^{T}\frac{\partial F^t}{\partial \beta_{1,1}}\frac{\partial F^t}{\partial \beta_{2,1}}\omega_1^t & \cdots & I_{(k_1+1,k_1)} & \left(\frac{\partial F^t}{\partial \beta_{2,1}}\right)^2\omega_1^t & \cdots & I_{(k_1+1,k)} & I_{(k_1+1,k+1)} \\
-\frac{1}{\Phi_1^t-F^t}\frac{\partial \Phi_1^t}{\partial \beta_{1,1}}\frac{\partial F^t}{\partial \beta_{2,1}} & & & & & & \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
\sum_{t=1}^{T}\frac{\partial F^t}{\partial \beta_{1,1}}\frac{\partial F^t}{\partial \beta_{2,k_2}}\omega_1^t & \cdots & I_{(k,k_1)} & \sum_{t=1}^{T}\frac{\partial F^t}{\partial \beta_{2,1}}\frac{\partial F^t}{\partial \beta_{2,k_2}}\omega_1^t & \cdots & I_{(k,k)} & I_{(k,k+1)} \\
-\frac{1}{\Phi_1^t-F^t}\frac{\partial \Phi_1^t}{\partial \beta_{1,1}}\frac{\partial F^t}{\partial \beta_{2,k_2}} & & & & & & \\
\sum_{t=1}^{T}\frac{\partial F^t}{\partial \beta_{1,1}}\frac{\partial F^t}{\partial f_{21}}\omega_1^t & \cdots & I_{(k+1,k_1)} & \sum_{t=1}^{T}\frac{\partial F^t}{\partial \beta_{2,1}}\frac{\partial F^t}{\partial f_{21}}\omega_1^t & \cdots & I_{(k+1,k)} & \left(\frac{\partial F^t}{\partial f_{21}}\right)^2\omega_1^t \\
-\frac{1}{\Phi_1^t-F^t}\frac{\partial \Phi_1^t}{\partial \beta_{1,1}}\frac{\partial F^t}{\partial f_{21}} & & & & & &
\end{pmatrix},$$

where I use the weights $\omega_1^t = \left[\frac{1}{F^t}+\frac{1}{\Phi_1^t-F^t}\right]$ and $\omega_2^t = \left[\frac{1}{\Phi_1^t-F^t}+\frac{1}{1-\Phi_1^t}\right]$. Since the

information matrix is symmetric, I use $I_{(h,j)}$ to denote the corresponding $I_{(j,h)}$ mirror elements in $I(\theta)$. Some entries in the matrix are not shown to save space and take the values as below

$I_{(k_1,k_1)} = \left(\frac{\partial F^t}{\partial \beta_{1,k_1}}\right)^2 + \left(\frac{\partial \Phi_1^t}{\partial \beta_{1,k_1}}\right)^2 \omega_2^t - \frac{2}{\Phi_1^t - F^t}\frac{\partial \Phi_1^t}{\partial \beta_{k_1,k_1}}\frac{\partial F^t}{\partial \beta_{k_1,k_1}}$,

$I_{(k,k_1)} = \frac{\partial F^t}{\partial \beta_{1,k_1}}\frac{\partial F^t}{\partial \beta_{2,1}}\omega_1^t - \frac{1}{\Phi_1^t - F^t}\frac{\partial \Phi_1^t}{\partial \beta_{1,k_1}}\frac{\partial F^t}{\partial \beta_{2,1}}$,

$I_{(k,k_1)} = \sum_{t=1}^{T} \frac{\partial F^t}{\partial \beta_{1,k_1}}\frac{\partial F^t}{\partial \beta_{2,k_2}}\omega_1^t - \frac{1}{\Phi_1^t - F^t}\frac{\partial \Phi_1^t}{\partial \beta_{1,k_1}}\frac{\partial F^t}{\partial \beta_{2,k_2}}$,

$I_{(k+1,k_1)} = \sum_{t=1}^{T} \frac{\partial F^t}{\partial \beta_{1,k_1}}\frac{\partial F^t}{\partial f_{21}}\omega_1^t - \frac{1}{\Phi_1^t - F^t}\frac{\partial \Phi_1^t}{\partial \beta_{1,k_1}}\frac{\partial F^t}{\partial f_{21}}$,

$I_{(k,k+1)} = \sum_{t=1}^{T} \frac{\partial F^t}{\partial \beta_{1,k_2}}\frac{\partial F^t}{\partial f_{21}}\omega_1^t$,

and $I_{(k,k)} = \left(\frac{\partial F^t}{\partial \beta_{2,k_2}}\right)^2 \omega_1^t$.

Consider the $(k_1+1)th$ row in the information matrix corresponding to the second intercept and the last $(k+1)th$ row corresponding to the variance parameter f_{21}. If $\frac{\partial F^t}{\partial f_{21}}$ equals $c\frac{\partial F^t}{\partial \beta_{2,1}}$ for all t then the two rows are the same up to that constant c which is independent of t. In this case the information matrix is singular and the parameters in the bivariate sample selection model are not identified.

Meng and Schmidt (1985) show that

$$\frac{\partial F^t}{\partial \beta_{2,1}} = \phi(Z_{2t}\beta_2)\Phi\left(\frac{Z_{1t}\beta_1 - \rho Z_{2t}\beta_2}{\sqrt{1-\rho^2}}\right), \tag{A.9}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the standard univariate normal density and distribution functions. If $Z_{1t}\beta_1 = \rho Z_{2t}\beta_2$ then this derivative is equal to $\phi(Z_{2t}\beta_2)/2$. They also show that if $Z_{1t}\beta_1 = \rho Z_{2t}\beta_2$ then $\frac{\partial F^t}{\partial \rho} = \frac{\phi(Z_{2t}\beta_2)}{\sqrt{2\pi(1-\rho^2)}}$. In my formulation $\rho = \frac{-f_{21}}{\sqrt{1+f_{21}^2}}$ and the model parameters are not identified if $Z_{1t}\beta_1 = \frac{-f_{21}}{\sqrt{1+f_{21}^2}} Z_{2t}\beta_2$.

APPENDIX B
DESCRIPTIVE STATISTICS OF THE SAMPLE IN CHAPTER 2

Table B.1: Means and standard deviations of the independent variables

Variable	Description	Mean	St.Dev.
Age	Age at baseline	77.624	5.895
Men	1 if Men	0.393	0.488
African American	1 if African American	0.134	0.341
Hispanic	1 if Hispanic	0.055	0.228
Widowed	1 if widowed	0.407	0.491
Divorced/Separated	1 if divorced or separated	0.054	0.226
Never Married	1 if never married	0.031	0.173
Religion not Important	1 if religion not important	0.109	0.312
Grade School	1 if completed grade school	0.278	0.448
College	1 if completed some college	0.258	0.438
Income Zero	1 if income zero	0.003	0.058
Log of Income	Log of (positive) income	9.737	1.040
# of Health Insurance Policies	# of policies	0.833	0.620
Long Term Care Insurance	1 if available	0.112	0.316
Neighborhood Safety Poor/Fair	1 if safety poor or fair	0.146	0.353
Home Value Zero	1 if homevalue zero	0.267	0.442
Log of Home Value	Log of homevalue if > 0	8.131	4.969
Population over 1,000,000	1 if population $>$ 1 million	0.487	0.500
Northeast region of US	1 if Northeast	0.197	0.398
North Central region of US	1 if Central	0.259	0.438
West region of US	1 if Mountain/Pacific	0.156	0.363
Mobile Home	1 if lives in mobile home	0.068	0.252
Multiple Story Home	1 if multiple story home	0.384	0.486
BMI	Body mass index	25.365	4.502
Diabetes	1 if diabetes (1993)	0.127	0.333
Heart	1 if heart disease (1993)	0.288	0.453
Hypertension	1 if hypertension (1993)	0.457	0.498
Previous Stroke	1 if stroke (1993)	0.099	0.299
Poor Self-Rated Health	1 if poor self-rated health	0.134	0.340
Fair Self-Rated Health	1 if fair self-rated health	0.234	0.424
ADL Sum	# of difficulties	0.386	0.928
IADL Sum	# of difficulties	0.487	1.110
Fall	1 if fell down	0.254	0.436
Picking up a Dime	1 if has difficulty	0.087	0.282

REFERENCES

Albert, J. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* 88(422):669–679.

Allison, P. (2001). *Missing Data*. No. 07-136 in Series on Quantitative Applications in the Social Sciences. Sage University Papers.

Amemiya, T. (1974). Bivariate probit analysis: Minimum chi-square methods. *Journal of the American Statistical Association* 69(348):940–944.

Ashford, J. and Sowden, R. (1970). Multi-variate probit analysis. *Biometrics* (26):535–546.

Baker, S. (1995). Marginal regression for repeated binary data with outcome subject to non-ignorable non-response. *Biometrics* 51(3):1042–1052.

Baker, S. and Laird, M. (1988). Regression analysis for categorical variables with outcome subject to nonignorable nonresponse. *Journal of the American Statistical Association* 83(401):62–69.

Bayarri, M. and Berger, J. (1998). Robust bayesian analysis of selection models. *Annals of Statistics* (26):645–659.

Bayarri, M. and DeGroot, M. (1987). Bayesian analysis of selection models. *The Statistician* (36):137–146.

Boyes, W., Hoffman, D. and Low, S. (1989). An econometric analysis of the bank credit scoring problem. *Journal of Econometrics* (40):3–14.

Chakravarty, S. and Li, K. (2003). A bayesian analysis of dual trader informativeness in future markets. *Journal of Empirical Finance* 10(3):355–371.

Chib, S. and Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika* 85(2):347–361.

Conaway, M. (1993). Non-ignorable non-response models for time-ordered categorical variables. *Applied Statistics* 42(1):105–115.

D'Agostino, R. (1998). Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine* (17):2265 – 2281.

Deb, P., Munkin, K. and Trivedi, P. (2006). Bayesian analysis of the two-part model with endogeneity: Application to health care expenditure. *Journal of Applied Econometrics* (21):1081–1099.

Diggle, P. and Kenward, M. (1994). Informative drop-out in longitudinal data analysis. *Applied Statistics* 43(1):49–93.

Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica* (57):1317–1339.

Geweke, J. (1991). Efficient simulation from the multivariate normal and Student-t distributions subject to linear constraints. In: E. Keramidas (ed.) *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*. Fairfax: Interface Foundation of North America, Inc.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of the posterior moments. In: J. Berger, J. Dawid and A. Smith (eds.) *Bayesian Statistics 4*. Oxford: Clarendon Press.

Geweke, J. (2004). Getting it right: Joint distribution tests of posterior simulators. *Journal of the American Statistical Association* (99):799–804.

Geweke, J. (2005). *Contemporary Bayesian Econometrics and Statistics*. Hoboken, NJ: Wiley.

Greene, W. (1992). A statistical model for credit scoring. WP No. EC-92-29, Department of Economics, Stern School of Business, New York University.

Greene, W. (2003). *Econometric Analysis*. 5th ed. Upper Saddle River, NJ: Prentice Hall.

Greene, W. (2006). Censored data and truncated distributions. In: T. Mills and K. Patterson (eds.) *Palgrave Handbook of Econometrics*, vol. 1: Econometrics Theory. Hampshire: Palgrave.

Greene, W. (2008). *Econometric Analysis*. 6th ed. Upper Saddle River, NJ: Pearson Education, Inc.

Gronau, R. (1974). Wage comparisons - a selectivity bias. *The Journal of Political Economy* 82(6):1119–1143.

Hajivassiliou, V. and McFadden, D. (1998). The method of simulated scores for the estimation of LDV models. *Econometrica* (66):863–896.

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica* 47(1):153–161.

Huang, H.C. (2001). Bayesian analysis of the SUR tobit model. *Applied Economics Letters* (8):617–622.

Keane, M. (1992). A note on identification in the multinomial probit model. *Journal of Business and Economic Statistics* 10(2):193–200.

Keane, M. (1994). A computationally practical simulation estimator for panel data. *Econometrica* (62):95–116.

Kenkel, D. and Terza, J. (2001). The effects of physician advice on alcohol consumption: Count regression with an endogenous treatment effect. *Journal of Applied Econometrics* 16(2):165–184.

Koop, G., Poirier, D. and Tobias, J. (2007). *Bayesian Econometric Methods*. New York, NY: Cambridge University Press.

Leacock, C. (2006). *Getting Started with the Health and Retirement Study*. Survey Research Center, Health and Retirement Study, University of Michigan, Ann Arbor, MI.

Lee, J. and Berger, J. (2001). Semiparametric bayesian analysis of selection models. *Journal of the American Statistical Association* (96):1269–1276.

Lee, L. (2003). Self-selection. In: B. Baltagi (ed.) *A Companion to Theoretical Econometrics*, chap. 18. Blackwell Publishing.

Leung, S. and Yu, S. (1996). On the choice between sample selection and two-part models. *Journal of Econometrics* (72):197–229.

Li, K. (1998). Bayesian inference in a simultaneous equation model with limited dependent variables. *Journal of Econometrics* (85):387–400.

Manning, W., Duan, N. and Rogers, W. (1987). Monte Carlo evidence on the choice between sample selection and two-part models. *Journal of Econometrics* (35):59–82.

McCulloch, R., Polson, N. and Rossi, P. (2000). A bayesian analysis of the multinomial probit model with fully identified parameters. *Journal of Econometrics* (99):173–193.

McCulloch, R. and Rossi, P. (1994). An exact likelihood analysis of the multinomial probit model. *Journal of Econometrics* (64):207–240.

Meng, C. and Schmidt, P. (1985). On the cost of partial observability in the bivariate probit model. *International Economic Review* (26):71–86.

Mohanty, M. (2002). A bivariate probit approach to the determination of employment: a study of teen employment differentials in Los Angeles county. *Applied Economics* 34(2):143–156.

Munkin, M. and Trivedi, P. (2003). Bayesian analysis of a self-selection model with multiple outcomes using simulation-based estimation: An application to the demand for healthcare. *Journal of Econometrics* (114):197–220.

Ochi, Y. and Prentice, R. (1984). Likelihood inference in a correlated probit regression model. *Biometrika* 71(3):531–543.

Park, T. (1998). An approach to categorical data with nonignorable nonresponse. *Biometrics* 54(4):1579–1590.

Preget, R. and Waelbroeck, P. (2006). Sample selection with binary endogenous variable: A bayesian analysis of participation to timber auctions. Working Paper ESS-06-08, Telecom Paris.

Rosenbaum, P. and Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* (70):41–55.

Rothenberg, T. (1971). Identification in parametric models. *Econometrica* (39):577–591.

Rubin, D. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association* (74):318–324.

Servais, M. (2004). *Overview of HRS Public Data Files for Cross-sectional and Longitudinal Analysis*. Survey Research Center, Health and Retirement Study, University of Michigan, Ann Arbor, MI.

Tanner, M. and Wong, W. (1987). The calculation of posterior distribution by data augmentation. *Journal of American Statistical Association* (82):528–540.

Terza, J. (1998). Estimating count data models with endogenous switching: Sample selection and endogenous treatment effects. *Journal of Econometrics* 84(1):129–154.

AHEAD Core 1993 public use dataset (1998). *Produced and distributed by the University of Michigan with funding from the National Institute on Aging (grant number NIA U01AG009740)*. Health and Retirement Study, Ann Arbor, MI.

Train, K. (2003). *Discrete Choice Methods with Simulation*. New York, NY: Cambridge University Press.

van Hasselt, M. (2005). Bayesian sampling algorithms for the sample selection and two-part models. Working paper, Department of Economics, Brown University.

van Hasselt, M. (2008). Bayesian inference in a sample selection model. Working paper, Department of Economics, The University of Western Ontario.

Vella, F. (1998). Estimating models with sample selection bias: A survey. *The Journal of Human Resources* 33(1):127–169.

Waelbroeck, P. (2005). Computational issues in the sequential probit model: A Monte Carlo study. *Computational Economics* 26(2):141–161.

Wolinsky, F., Bentler, S., Cook, E., Chrischilles, E., Liu, L., Wright, K., Geweke, J., Obrizan, M., Pavlik, C., Ohsfeldt, R., Jones, M., Wallace, R. and Rosenthal, G. (2009). A 12-year prospective study of stroke risk in older medicare beneficiaries. *BMC Geriatrics* pp. 9–17.

Wooldridge, J. (2002). *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: The MIT Press.

Wynand, P. and Praag, B.V. (1981). The demand for deductibles in private health insurance. *Journal of Econometrics* (17):229–252.

Zellner, A. (1986). On assessing prior distributions and bayesian regression analysis with g-prior distributions. In: P. Joel and A. Zellner (eds.) *Bayesian Inference and Decision Techniques: Essays in Honour of Bruno de Finetti*. Amsterdam: North Holland.