ESSAYS ON UNOBSERVED HETEROGENEITY AND ENDOGENEITY IN HEALTH ECONOMETRICS

by
HIROAKI MASUHARA

Submitted in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in
Economics

Graduate School of Economics
Hitotsubashi University

2014

Contents

List of Tables

List of Figures

Acknowledgments

In writing this dissertation, I benefited from the help and support of a number of people whom I would like to thank here.

First of all, I am grateful to my supervisor, Prof. Dr. Motohiro Sato, for his support and guidance. I would also like to thank Prof. Dr. Masako Ii and Prof. Dr. Eiji Tajika for cosupervising my dissertation for thought-provoking discussions on my research. Besides my advisor, I would like to thank the rest of my thesis committee: Prof. Dr. Kazumitsu Nawata, Prof. Dr. Takashi Oshio, and Prof. Dr. Hiroyuki Kawaguchi, for their encouragement, insightful comments, and hard questions. I am grateful to my past colleagues at the graduate school of economics in Hitotsubashi University—Kei Hosoya, Yukinari Hayashi—for valuable comments and suggestions on my essays.

I am especially grateful to the late Dr. Tadahiko Tokita. He guided me to the field of health econometrics and encouraged me tremendously. Finally, special thanks go to my family and to my friends for their constant encouragement and support.

Introduction

Microdata in Health Economics

During the past two decades, applied econometric analysis has been widely adopted among health economists. Its adoption is accelerating, producing ever-richer research as electronic recording and collection make available more data about individual patients. In addition, computational power for analyzing large, complex datasets is increasing, facilitating econometric analysis involving latent variables, unobserved heterogeneity, and nonlinear models in the field now established as "health econometrics." Jones and O'Donnell (2002), Jones (2007), and Jones *et al.* (2007) review health econometrics comprehensively, and many textbooks examine microdata and related topics (Amemiya, 1985; Maddala, 1986; Cameron and Trivedi, 1998, 2005; Gouriéroux, 2000; Wooldridge, 2002; Winkelmann, 2004; Winkelmann and Boes, 2006; Greene, 2007a).

Extensive individual-, household-, and establishment-level microdata are available from cross-sectional and longitudinal sample surveys and the census. Health economics primarily employs cross-sectional data. That is, observations are independent of each other, and pure time series applications are excluded.[1] Microdata used in health econometrics have two notable features. First, they are often measured on a non-continuous scale: data are not only continuous and discrete variables but also on a non-continuous scale, such as quantitative and qualitative (or categorical) variables. This leads inconsistency of linear regression models. For example, analyzing expenditure data is complicated when samples feature a preponderance of observations with zero expenditures. The consistency of standard approaches to the problem relies on the validity of distributional assumptions. To analyze these data, health econometrics requires disparate nonlinear models, including binary responses, multinomial responses, limited dependent variables, integer counts, and measures of duration. Moreover, variables denoting health or quality of life are often unobservable and perhaps measurable only with

[1] Panel data, which contain both time series and cross-sectional properties, are regarded as microdata. However, this dissertation focuses on only cross-section microdata.

error (through subjective reports, for example). This situation induces latent variables and selection problems.

Second, health data are observational, i.e., they are neither experimental nor collected from surveys and administrative records through randomized experiment. Although availability of "experimental" data is increasing in the social sciences, their use is restricted, and empirical works continue to rely on non-experimental data. Accordingly, sample selection bias may pervade observational data in health econometrics. In analyzing smoking-related illness, for example, smokers acknowledge their risks and rationally select their behavior. Failing to consider self-selection distorts the estimated health effects of smoking based on comparisons between smoking and non-smoking samples.

Microdata used in health econometrics are quantitative and qualitative (categorical). Qualitative data are discrete and are of three types: binary, multinomial, and ordered. A *binary variable* addresses only two possible outcomes and indicates presence or absence of a property. It arises, for example, in answers to questions about the fulltime employment status (yes/no). A *multinomial variable* is a natural extension of binary data and has at least three possible outcomes. It indicates the quality of an object using a set of mutually exclusive and exhaustive non-ordered categories. For instance, it arises from questions about employment status featuring several alternatives (full-time · part-time · unemployed · not in a labor force). *Ordered variables* have three or more possible outcomes. They indicate qualitative features using sets of mutually exclusive and exhaustive ordered categories, but differences between categories are undefined. Ordered variables arise from questions such as "How satisfied are you with the medical system" (completely satisfied · somewhat satisfied · neutral · somewhat dissatisfied · completely dissatisfied)?

Quantitative data are discrete or continuous and also restricted or unrestricted. Discrete and unrestricted quantitative data are *count variables.* They take the form of non-negative integers $\{0, 1, 2, \dots\}$. The number of physician visits is an example. Count data fill an intermediate position between qualitative and quantitative data. If the number of counts is relatively low, responses are treated as categories.

Limited dependent variables are the continuous and restricted data and consist of three types: non-negative variables with frequent zeros, truncated variables, and censored variables. In *non-negative variables with frequent zeros*, there is a continuous positive variable with a discrete cluster of observations at zero. For example, data about monthly medical expenditures have many zeros and right-skewed distributions. Such data provide two kinds of information: how medical care is utilized and in what quantity.

The data are *truncated* if all observations with realizations above or below a specified threshold are excluded from the sample. For instance, if we observe data for medical expenditures above zero and the possibility of zero is not eliminated, the data are truncated. Consequently, observed data no longer represent the population even if the sampling is otherwise admitted.

Data are *censored* if only an interval rather than a numerical value is observed. Censored data are common in duration analysis. For example, when we observe the duration of leaving hospital but the study terminates one week later, we do not know complete spell of the sample leaving hospital after one week. This is a censored observation. Unlike truncated data, censoring does not exclude those observations from the sample and its proportion is known.

Microeconometrics in Health Economics

Linear and Nonlinear Regression Models

Applied econometric studies often employ standard linear regression models. These models assume that the relation between an outcome (dependent variable) y_i and explanatory variables (independent variables, regressors, or covariates) $\mathbf{x}_i$; is a linear function of the $\mathbf{x}_i$; variables and of a random error term ε_i. This relation can be noted in shorthand as

$$y_i = \mathbf{x}_i' \boldsymbol{\beta} + \varepsilon_i,$$

where $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{iK-1})'$ is a $K \times 1$ vector and $\boldsymbol{\beta}$ is a $K \times 1$ parameter vector. For simplicity, we drop the subscript i and write the model for typical observation as $y = \mathbf{x}'\boldsymbol{\beta} + \varepsilon$. The random error term ε captures all the variation in y not explained by the $\mathbf{x}$ variables. The classical model makes four assumptions about the error term: (i) its mean is zero; (ii) its variance σ^2 is the same across all observations (homoskedasticity); (iii) its values are independent across observations (serial independence); (iv) its values are independent of the values of the $\mathbf{x}$ variables (exogeneity).

Investigators often assume the error term has a normal distribution. This implies that, conditional on each $\mathbf{x}$, each observation of dependent variable y follows a normal distribution with mean $\mathrm{E}(y \mid \mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}$. This assumption has two implications. First, the ordinary least squares (OLS) estimator is asymptotically efficient among all possible estimators. Second, the small sample distribution of the OLS estimator is known, and exact inference can therefore be based on t- or F-statistics. This standard linear regression model is easily estimated and interpreted, and it provides optimal inference if standard regularity assumptions are fulfilled. Under these Gauss-Markov assumptions, the OLS estimator is the best linear unbiased estimator.

However, if the dependent variable is neither quantitative nor continuous, the OLS estimator may be inappropriate. First, we consider the case of a binary dependent variable that takes 0 or 1. In this case, the linear regression is interpreted as a probability model, since $\mathrm{E}\left(y \mid \mathbf{x}\right) = 0 \times P\left(y = 0 \mid \mathbf{x}\right) + 1 \times P\left(y = 1 \mid \mathbf{x}\right)$. Therefore, we obtain

$$P\left(y = 1 \mid \mathbf{x}\right) = \mathbf{x}'\boldsymbol{\beta}.$$

If we calculate the prediction using this model, it is required that $0 \leq P\left(\widehat{y} = 1 \mid \mathbf{x}_0\right) \leq 1$. However, the restriction on linearity is violated for certain values $\mathbf{x}_0$ of the regressors. Moreover, this model is not homoskedastic since the variance of a binary variable conditional on the regressors takes the values of $\mathrm{Var}\left(y = 1 \mid \mathbf{x}\right) = P\left(\widehat{y} = 1 \mid \mathbf{x}\right)\left[1 - P\left(\widehat{y} = 1 \mid \mathbf{x}\right)\right]$, which is a function of $\mathbf{x}$. A similar discussion applies to multinomial dependent variables. The computed expected value of a multinomial variable has no meaning using a linear model. Since the numerical coding of outcomes is qualitative and arbitrary, no ranking affects the analysis.

Second, consider a count dependent variable that takes the value of non-negative integer. Count data are quantitative and have well-defined expectations, but the linear regression model is again inappropriate. The expectation of a count must be non-negative, but this expectation is not assured by the functional form above. Moreover, variance in count data analysis generally depends on $\mathbf{x}$, and that dependence violates the assumption of homoskedasticity.

Third, examine the case of limited dependent variables. If the dependent variable is continuous with support over the real line, there is no argument against using the linear regression model, and it indeed is the best. However, it is inappropriate and other models are required if the dependent variable is limited to positive real numbers and zeros are important. Since the limited dependent variable is censored or truncated and it is undesirable to regard the observed sample as representative of the population, to estimate the linear regression model directly takes the biased estimator. The estimator fails because the assumption of mean independence between the error terms and regressors must fail under sample selection. Similar considerations apply to duration analysis.

In health econometrics, empirical analysis is complicated because outcomes of individual-level survey data often are based on qualitative or limited dependent variables and nonlinear models are necessary. Moreover, the discipline's theoretical models often involve unobservable (latent) concepts such as health endowments, physician agency and supplier inducement, or quality of life. Therefore, health econometrics requires nonlinear regression models such as binary responses, multinomial responses, limited dependent variables, duration, and count data.

Methods for modeling such data are interrelated and based on maximum likelihood estimation (MLE). The MLE method differs from the least squares method used to fit a regression line to data. It assumes a distribution of the data-generating process and the estimate parameters based on this distribution. Therefore, distributional assumptions dictate whether estimated parameters are strongly true, but many applications using maximum likelihood are parametric. This disadvantage has been discussed, and this dissertation addresses this problem later.

An Evaluation Problem in Regression Analysis

An evaluation problem is how to identify causal effects from empirical data. Consider an outcome y_{it} for individual i at time t—for example, the extent to which someone sought health care during the past year. If we analyze the influence of health maintenance activities, such as hours of exercise per month, on outcome y_{it}, it is difficult to identify the causal effect of treatment. The causal effect of interest is the difference between the outcome with treatment and without treatment. However, this pure treatment effect is not identifiable from empirical data because the counterfactual can never be observed, i.e., the patient cannot be two places simultaneously.

To analyze this problem, it is useful to estimate the average treatment effect using sample data by comparing the average outcome among those receiving treatment with the average outcome and with those who do not receive treatment. However, if unobserved factors influence both the selection of treatment and the response to it, this method promotes biased estimators of the treatment effect. It is best to use a randomized experimental design that randomly allocates individuals into treatments, and in some circumstances it is better to use natural experiment data. Because this method is prohibitively expensive, however, many empirical studies use non-experimental data. In the absence of experimental data, we require alternative estimation strategies, such as instrumental variables, corrections for selection bias, and longitudinal data.

Because health econometrics employs quantitative and qualitative (categorical) data, nonlinear models are necessary. Hence, the instrumental variables method based on linear regression is sometimes inappropriate for analyzing non-experimental health econometrics data. Here we use MLE based on a parametric distribution. We consider this problem later when introducing semiparametric distribution.

Outline

Part I of this dissertation analyzes a heterogeneity problem in nonlinear health econometrics. Part II considers an endogeneity problem in nonlinear health econometrics.

Unobserved heterogeneity causes problems in nonlinear regression models such as duration models and count data models. Here, heterogeneity means that data differ across observations. In linear regression models, when the heterogeneity is independent of regressors, the OLS estimator is not always efficient but consistent because the conditional mean is unchanged, the unobserved heterogeneity is absorbed into the error term, and omitted variables bias is absent. In nonlinear regression models, omitting unobserved heterogeneity causes spurious results, i.e., spurious negative (or positive) dependence in duration analysis or a greater (smaller) variance in count data analysis. Chapter 1 in Part I reviews unobserved heterogeneity in nonlinear health econometric models. First, we consider continuous heterogeneity and introduce gamma distributed heterogeneity, which is often used in duration and count data analysis. Second, we investigate discrete heterogeneity, which is referred as a *finite mixture model* and is semiparametric. Moreover, we show the limitations of nonlinear regression models to introduce heterogeneity and suggest alternatives to avoid these problems.

Chapter 2 suggests generalized and semiparametric log-normal survival analysis using Hermite polynomials and Box-Cox transformation. It is empirically difficult to separate the effects of duration dependence from those of unobserved heterogeneity, so many survival models do not explicitly assume unobserved heterogeneity. However, omitted variables are inevitable, and controlling population heterogeneity is not always adequate. The model without unobserved heterogeneity overestimates (underestimates) the degree of negative (positive) duration dependence in the hazard. We propose new semiparametric (semi-nonparametric) survival models that generalize unobserved heterogeneity, as well as a dependent variable of the log-normal survival model. First, we generalize the log-transformed dependent variable using Box-Cox transformation, which contains various function forms. Second, we generalize the normally distributed unobserved heterogeneity using Hermite polynomials, which include a normal distribution as a special case. The General Social Survey in 2002 shows that the proposed model performs well in empirical application.

Chapter 3 proposes and demonstrates the identifiability of a finite mixture cross-sectional probit model in selected situations, i.e., a probit model with a single linear equation. Although finite mixture models are semiparametric and flexible, a cross-sectional finite mixture probit (binomial) model is not estimated for an identification problem. However, it is not enough to apply only a cross-sectional probit model because we do not know the true data-generating process of a binary variable. Therefore, this chapter investigates the possibility of estimating a cross-sectional finite mixture probit model. We show the identifiability of bivariate random variables using a natural expansion of Teicher's theorem. Using this result, the chapter then investigates the identifiability of a finite mixture *cross-sectional* probit model with one linear equation. We demonstrate that the class of all finite mixtures of a probit model with one linear equation is identifiable even if the number of components does not exceed three. That is, a finite mixture *cross-sectional* probit model sometimes can be estimated. Monte Carlo simulations support our demonstration.

It is known in microeconometrics, especially in health econometrics, that endogenous regressors may cause inconsistent parameter estimation. Health econometrics faces no endogeneity problem if data are randomly assigned or regressors are not the results of incentives, as in the experimental sciences. However, these conditions are seldom fulfilled in social sciences, and endogeneity bias is inevitable. Therefore, a method to treat it correctly is required. Focusing on health econometrics, Chapter 4 in Part II reviews the problem of endogeneity and explains the estimation of regression models with endogenous regressors. First, we analyze the problem of endogeneity using a simple linear regression model, explain the instrumental variable method that obtains the consistent estimator even if endogenous variables exist, and describe the two-stage least squares method (2SLS) often used in applied fields. Although the discussion of instrumental variable estimators is based on continuous endogenous regressors, we extend this discussion to a binary endogenous variable, referred to as *treatment effects*. Second, we explain, using examples of probit and count data models, that the two-stage method is applied in nonlinear models with endogenous continuous regressors. We demonstrate that, in nonlinear regression with endogenous discrete, censored, or truncated regressors, the two-stage method is insufficient, and the full information maximum likelihood method (FIML) is consistent. Third, we provide Monte Carlo simulations of the four cases and analyze the consistency of proposed models. We show the consistency of linear models with an endogenous continuous, discrete, censored, or truncated regressor and the inconsistency of probit models with an endogenous binary variable. Finally, we show the limitation of nonlinear health econometric regressions containing endogenous variables and propose more desirable analysis.

Chapter 5 proposes a semiparametric (semi-nonparametric) Poisson model with an endogenous binary variable, which generalizes bivariate correlated unobserved heterogeneity using Hermite polynomials, and compares this model with a parametric model. Health econometrics encounters occasions in which explanatory variables are simultaneously determined with the dependent variable. In such cases, Poisson or negative binomial models yield biased estimates of parameters of interest because they assume perfect explanatory variables are perfectly exogenous. Therefore, count data models with an endogenous binary variable are required, and many studies have analyzed this problem. Chapter 5 considers a Poisson model with one endogenous binary variable and the heterogeneity of both count dependent and binary variables. We propose a Poisson model that comprises a semiparametric joint distribution using Hermite polynomials. Our model is semiparametric and includes the natural extension of a bivariate normal distribution. In an example using 1990 National Health Interview Survey data, the semi-parametric model overcomes rival models in terms of the likelihood ratio test. Absolute values of the endogenous binary regressor coefficients of the semiparametric models are smaller than those of the parametric model, and those in the semiparametric model are the smallest among the three. Moreover, estimated densities of the semiparametric models have fatter tails than the parametric model.

Chapter 6 proposes a robust duration model with an endogenous binary variable. As with many nonlinear models, endogeneity in duration analysis is a problem because censored duration data lead to nonlinearity, prompting the two-stage method toward inconsistency. Studies have addressed endogeneity in duration analysis, but models based on a hazard rate do not explicitly assume heterogeneity. Chapter 6 proposes an alternative semiparametric duration model with an endogenous binary variable that generalizes the heterogeneity of both duration and endogeneity. Heterogeneity is generalized as follows. First, we consider a simple log-normal duration model with an endogenous binary variable. Second, we assume heterogeneity that follows a semiparametric bivariate distribution using Hermite polynomials. Under these setups, we investigate the difference between the endogenous binary variable's coefficients of the parametric and semiparametric models using Medical Expenditure Panel Survey (MEPS) data. When applied to the duration of hospital stays in MEPS data, the estimated results of non-censored and artificially censored semiparametric (semi-nonparametric) models show good performance. The absolute values of the endogenous binary regressor coefficients of the semiparametric models are larger than in parametric models whether data are censored or not. This introduces the interpretation of the binary endogenous variable, that is, the variable denoting insurance coverage. The parametric model underestimates the effect of a survey respondent's insurance coverage in our example. The difference of the estimated endogenous coefficients in the two models is smaller than in parametric models. This means that the parametric model has a large inconsistency if the data are censored. Moreover, estimated densities of the semiparametric models have twin peak distributions.

The main contributions of this dissertation are as follows. First, since the true distribution of heterogeneity is usually unknown, to generalize a distribution leads to the consistent estimators of coefficients. In linear regression models, independent unobserved heterogeneity to regressors causes no complications. However, in nonlinear regression models, an observed positive or negative relationship may be spurious due to unobserved heterogeneity across samples. The only way to obtain consistent estimators of coefficients is to use semiparametric models that generalize unobserved heterogeneity.

Second, using generalized unobserved heterogeneity, average treatment effects in health economics is correctly estimated and is accurately tested. In general, large-scale cross-section or panel microdata (or survey datasets) are applied to measuring the effect of some treatment to health. Since these data are based on a face-to-face interview or a self-completion postal questionnaire, there are many discrete, censored, or truncated variables and is a few continuous variables. Moreover, these data contain a potential problem of endogeneity because it is difficult to assume some variable as an exogenous variable. Therefore, to evaluate average treatment effects in health economics is inevitable to use nonlinear regression models with discrete endogenous variables. This dissertation obtains robust and more accurate methods to estimate and test the average treatment effects.

The methods that this dissertation investigates may be cumbersome and conventional methods that are often used in health economics are tractable. However, if and only if the method discussed in this dissertation demonstrates that unobserved heterogeneity follows some specific distribution, then to utilize conventional methods is justified. Therefore, it is very important to generalize unobserved heterogeneity for robust and correct estimators of coefficients.

Part I

Unobserved Heterogeneity in Health Econometrics

Chapter 1

Nonlinear Estimation and Heterogeneity in Health Econometrics

1.1 Introduction

Many statistical and econometric studies investigate unobserved heterogeneity, also known as *frailty* in biostatistics. Roughly speaking, heterogeneity means that data differ across observations. Either regressors (observables) or unobservables cause heterogeneity in regression analysis. *Observed* heterogeneity is inter-individual differences that are measured by regressors, and *unobserved* heterogeneity is all other differences. Unobserved heterogeneity causes no complications in linear regression models if heterogeneity is independent of regressors. In such cases, the conditional mean is unchanged, the unobserved heterogeneity is absorbed into the error term, and there is no omitted variables bias. Unobserved heterogeneity does cause problems in nonlinear models, such as duration models and count data analysis. For example, in survival analysis, especially in Cox proportional hazard models, the baseline hazard contains observed and unobserved heterogeneity and thus even individuals presenting the same values for all covariates may have different hazards.

We consider this problem using a well-known empirical example discussed by Cameron and Trivedi (2005) and Winkelmann and Boes (2006). Assume the population is composed of two groups in a 50/50 proportion. Group 1 has a hazard rate of 0.5 and Group 2 a hazard rate of 0.1. For 100 people in Group 1, we observe 50 transitions in the first period, 25 in the second, and 12.5 in the third. For Group 2, we observe 10, 9, and 8.1 transitions in the first, second, and third periods, respectively. Therefore, an aggregate hazard rate is $(50 + 10)/200 = 0.3$, $(25 + 9)/140 = 0.24$, and $(12.5 + 8.1)/106 = 0.19$. The average hazard rate drops from 30% to 24%, and then to 19% in the next period. If the empirical evidence shows negative duration dependence—that is, the hazard rate falls over time—this character may be spurious. The declining aggregate hazard is a consequence of aggregation across heterogeneous groups that have constant but different hazard rates.

Count data analysis encounters the same problem. Assume two homogeneous groups of equal size in the population under study. Each is characterized by a Poisson distributed random variable, denoted by y_1 and y_2, with parameters $\lambda_1 = 0.5$ and $\lambda_2 = 1.5$, respectively. The analyst cannot distinguish between both groups because of the lack of data. The results above obtains

$$\begin{aligned}\mathrm{E}(y) =& 0.5 \times \mathrm{E}(y_1) + 0.5 \times \mathrm{E}(y_2)\\ =& 0.5 \times 0.5 + 0.5 \times 1.5 = 1,\end{aligned}$$

$$\begin{aligned}\mathrm{Var}(y) =& 0.5 \times \mathrm{Var}(y_1) + 0.5 \times \mathrm{Var}(y_2)\\ &+ 0.5 \times [\mathrm{E}(y_1) - \mathrm{E}(y)]^2 + 0.5 \times [\mathrm{E}(y_2) - \mathrm{E}(y)]^2\\ =& 0.5 \times 0.5 + 0.5 \times 1.5 + 0.5 \times (-0.5)^2 + 0.5 \times 0.5^2 = 1.25.\end{aligned}$$

The unconditional variance of y in the population is greater than its unconditional mean, and the population cannot be Poisson distributed.

It is important to consider the consequences of this unavoidable misspecification. It is known from ordinary linear multiple regression analysis that such an omission can lead to omitted variable bias. However, analysis of unobserved heterogeneity is more complex in nonlinear models. Introducing unobserved heterogeneity leads to an important class of models called *mixture models*. This chapter explains unobserved heterogeneity in nonlinear health econometric models. Section 2 analyzes continuous heterogeneity. Section 3 investigates discrete heterogeneity, using examples of duration and count data models common in health econometrics. Moreover, the following chapters establish the limitations of nonlinear regression models to introduce heterogeneity and suggest alternatives that avoid these problems.

1.2 Continuous Heterogeneity

1.2.1 Mixtures of Distribution

To understand the effect of unobserved heterogeneity in nonlinear econometric models, we first consider an exponential duration model. In the exponential regression without heterogeneity, the distribution of non-censored spells t_i is specified conditional on observable exogenous covariates $\mathbf{x}_i$. We formulate the hazard function $\lambda\left(t_i \mid \mathbf{x}_i, \varepsilon_i\right)$ with an additive error term ε_i as

$$\ln \lambda\left(t_i \mid \mathbf{x}_i, \varepsilon_i\right) = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i, \tag{1.1}$$

where $\boldsymbol{\beta}$ is a $K \times 1$ vector of parameters. For notational simplicity, we omit subscript i in the following analysis. The random error ε shows the unobserved heterogeneity and can arise if additional information affects the hazard but is unobserved.

Equation (1.1) can be rewritten as $\lambda\left(t \mid \mathbf{x}, \varepsilon\right) = \exp\left(\mathbf{x}'\boldsymbol{\beta}\right)\nu$, where $\nu = \exp\left(\varepsilon\right)$. This model is said to have multiplicative heterogeneity. Without loss of generality, let $\mathrm{E}\left(\nu \mid \mathbf{x}\right) = 1$. Thus, $\mathrm{E}\left[\lambda\left(t \mid \mathbf{x}, \nu\right) \mid \mathbf{x}\right] = \exp\left(\mathbf{x}'\boldsymbol{\beta}\right)\mathrm{E}\left[\nu \mid \mathbf{x}\right] = \exp\left(\mathbf{x}'\boldsymbol{\beta}\right)$. The cumulative distribution function conditional on ν for the exponential model takes the form

$$F\left(t \mid \mathbf{x}, \nu\right) = 1 - \exp\left(-\exp\left(\mathbf{x}'\boldsymbol{\beta}\right)\nu t\right). \tag{1.2}$$

The unconditional cumulative distribution function is obtained by averaging the conditional function. The technique for doing so is to marginalize (1.2) with respect to ν, (e.g., to integrate out the unobservables). Let $g\left(\nu \mid \mathbf{x}\right)$ denote the density function of ν given $\mathbf{x}$. We obtain the marginal distribution function by integrating the product of $F\left(t \mid \mathbf{x}, \nu\right)$ and $\nu\left(u \mid \mathbf{x}\right)$ over ν:

$$F\left(t \mid \mathbf{x}\right) = \int_0^\infty F\left(t \mid \mathbf{x}, \nu\right) g\left(\nu \mid \mathbf{x}\right) \mathrm{d}\,\nu. \tag{1.3}$$

A parametric density function of $g\left(\nu \mid \mathbf{x}\right)$ is usually specified. The gamma distribution is a suitable candidate since it satisfies $\nu > 0$ and leads to relatively simple derivations and *closed-form* solutions. When ν is gamma distributed with parameters $\theta > 0$ and $\gamma > 0$, the density function is obtained by

$$g\left(\nu \mid \mathbf{x}\right) = \frac{\gamma^\theta}{\Gamma\left(\theta\right)}\nu^{\theta-1}e^{-\gamma\nu}, \tag{1.4}$$

where $\mathrm{E}\left[\nu \mid \mathbf{x}\right] = \theta/\gamma$, $\mathrm{Var}\left[\nu \mid \mathbf{x}\right] = \theta/\gamma^2$, and $\Gamma\left(\theta\right)$ is the gamma function defined as $\int_0^\infty z^{\theta-1}e^{-z}\mathrm{d}\,z$. For the normalization condition of $\mathrm{E}\left[\nu \mid \mathbf{x}\right] = 1$, we impose the restriction $\theta = \gamma$. Figure 1.1 shows two probability density functions (PDF) of the gamma distribution with $\mathrm{E}\left[\nu \mid \mathbf{x}\right] = 1$, $\mathrm{Var}\left[\nu \mid \mathbf{x}\right] = 0.5$ and $\mathrm{Var}\left[\nu \mid \mathbf{x}\right] = 0.25$. Therefore, the unconditional cumulative distribution function takes the following form:

$$\begin{aligned} F\left(t \mid \mathbf{x}\right) &= \int_0^\infty \left[1 - \exp\left(-\exp\left(\mathbf{x}'\boldsymbol{\beta}\right)\nu t\right)\right]\frac{\theta^\theta}{\Gamma\left(\theta\right)}\nu^{\theta-1}e^{-\theta\nu}\mathrm{d}\,\nu \\ &= 1 - \frac{\theta^\theta}{\Gamma\left(\theta\right)}\int_0^\infty \nu^{\theta-1}e^{-\left(\theta+\exp\left(\mathbf{x}'\boldsymbol{\beta}\right)t\right)\nu}\mathrm{d}\,\nu \\ &= 1 - \left(\frac{\theta}{\theta+\exp\left(\mathbf{x}'\boldsymbol{\beta}\right)t}\right)^\theta. \end{aligned} \tag{1.5}$$

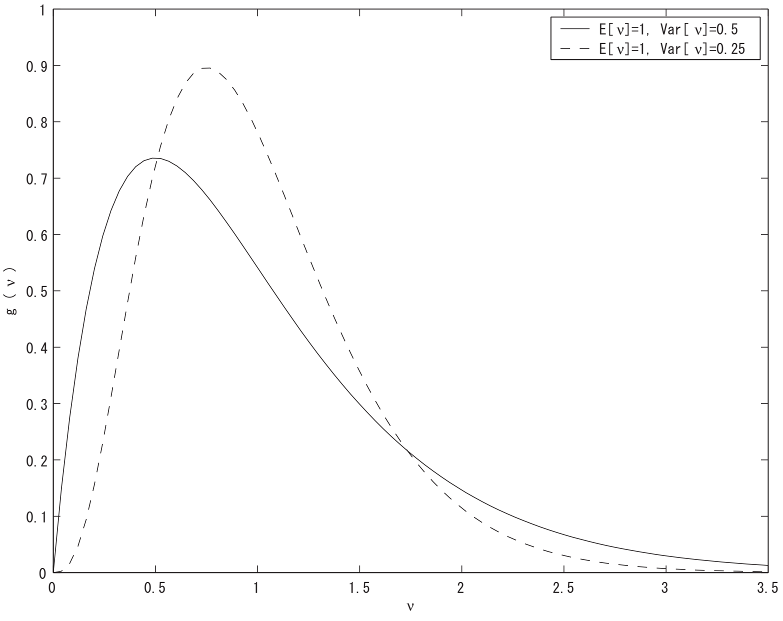


Figure 1.1: PDFs of Gamma Distributions

Because the integrand equals the density function of a Gamma $(\theta, \theta + \exp(\mathbf{x}'\boldsymbol{\beta})\, t)$ distribution and integrate to unity, the last equality is obtained. The unconditional duration density function is given by differentiating with respect to t, which yields

$$f(t \mid \mathbf{x}) = \exp(\mathbf{x}'\boldsymbol{\beta}) \left[1 + \theta^{-1} \exp(\mathbf{x}'\boldsymbol{\beta})\, t\right]^{-\theta-1}. \tag{1.6}$$

The unconditional hazard function in the exponential model with gamma distributed unobserved heterogeneity is written as

$$\lambda(t \mid \mathbf{x}) = \frac{f(t \mid \mathbf{x})}{1 - F(t \mid \mathbf{x})} = \exp(\mathbf{x}'\boldsymbol{\beta}) \left[1 + \theta^{-1} \exp(\mathbf{x}'\boldsymbol{\beta})\, t\right]^{-1}. \tag{1.7}$$

For $\theta^{-1} \to 0$, the hazard function is the simple exponential model; for $\theta^{-1} > 0$ the hazard rate becomes a decreasing function of t.

Poisson Regression Analysis

The same discussion applies to unobserved heterogeneity when analyzing count data, but first we explain count data models.[1] Count data use non-negative integers $\{0, 1, 2, \dots\}$ with no explicit upper limit to describe how many times an event occurs within a fixed interval. The natural stochastic model for counts is a Poisson point process for occurrence of the event. Let

[1] Winkelmann and Zimmermann (1995), Cameron and Trivedi (1998), and Winkelmann (2003) comprehensively explain various count data models.

$y = 0, 1, 2, \ldots$ denote a random variable that takes the non-negative integer. The PDF of a Poisson distribution takes

$$\mathrm{P}\left(Y = y\right) = \frac{\exp\left(-\mu\right)\left(\mu\right)^{y}}{y!}, \tag{1.8}$$

where μ is the intensity parameter. The Poisson distribution is a one-parameter distribution and parameter μ uniquely determines mean and variance: $\mathrm{E}\left(y\right) = \mathrm{Var}\left(y\right) = \mu$. The Poisson regression model is derived from the Poisson distribution by parameterizing the relation between the mean parameter μ_i and covariates $\mathbf{x}_i$, introducing the observation subscript i, attached to both y_i, μ_i, and $\mathbf{x}_i$.[2] It is convenient to specify μ as a log-linear function of covariates $\mathbf{x} \sim K \times 1$ that account for observed sample heterogeneity: $\mu = \exp\left(\mathbf{x}'\boldsymbol{\beta}\right)$, where $\boldsymbol{\beta}$ denote a $K \times 1$ vector of unknown parameters. Since $\mathrm{Var}\left(y\right) = \exp\left(\mathbf{x}'\boldsymbol{\beta}\right)$, the Poisson regression is intrinsically heteroskedastic.

The Poisson regression has the property of *equidispersion*, i.e., expectation and variance are equal. However, equidispersion is frequently violated in empirical applications through *overdispersion* (variance exceeds the mean) or *underdispersion* (variance smaller than the mean). Therefore, by introducing unobserved heterogeneity ε, the Poisson regression specifies as follows:

$$f\left(y\right) = \int \frac{\exp\left(-\exp\left(\mathbf{x}'\boldsymbol{\beta} + \varepsilon\right)\right)\left(\exp\left(\mathbf{x}'\boldsymbol{\beta} + \varepsilon\right)\right)^{y}}{y!} g\left(\varepsilon\right) \mathrm{d}\,\varepsilon, \tag{1.9}$$

where $g\left(\varepsilon\right)$ is an unknown PDF. We assume that random term ν enters the conditional mean function multiplicatively like the above duration analysis. That is, $\nu = \exp\left(\varepsilon\right)$. The marginal distribution of y is obtained by integrating out ν,

$$f\left(y \mid \mathbf{x}\right) = \int_{0}^{\infty} f\left(y \mid \mathbf{x}, \nu\right) g\left(\nu \mid \mathbf{x}\right) \mathrm{d}\,\nu, \tag{1.10}$$

where $g\left(\nu \mid \mathbf{x}\right)$ is the density function of unobserved heterogeneity ν. As in exponential-gamma distribution, we specify that ν has a gamma distribution obtained by (1.4). Setting $\theta = \gamma$ for normalization of $\mathrm{E}\left[\nu \mid \mathbf{x}\right] = 1$, the marginal distribution of y is given by

$$\begin{aligned} f\left(y \mid \mathbf{x}\right) &= \frac{\mu^{y}}{y!} \int_{0}^{\infty} \exp\left(-\mu\nu\right) \nu^{y} \frac{\theta^{\theta}}{\Gamma\left(\theta\right)} \nu^{\theta - 1} e^{-\theta\nu} \mathrm{d}\,\nu \\ &= \frac{\mu^{y}}{y!} \frac{\theta^{\theta}}{\Gamma\left(\theta\right)} \int_{0}^{\infty} e^{-\left(\mu + \theta\right)\nu} \nu^{y + \theta - 1} \mathrm{d}\,\nu \\ &= \frac{\mu^{y}}{\Gamma\left(y + 1\right)} \frac{\theta^{\theta}}{\Gamma\left(\theta\right)} \frac{\Gamma\left(y + \theta\right)}{\left(\mu + \theta\right)^{y + \theta}} \\ &= \frac{\Gamma\left(y + \theta\right)}{\Gamma\left(y + 1\right)\Gamma\left(\theta\right)} \left(\frac{\mu}{\mu + \theta}\right)^{y} \left(\frac{\theta}{\mu + \theta}\right)^{\theta}. \end{aligned} \tag{1.11}$$

[2] We omit subscript i in the following analysis for notational simplicity.

Since the integrand equals the density function of a $\text{Gamma}\left(y+\theta, \theta+\mu\right)$ distribution and integrate to unity, the third equality is obtained. This PDF is the mixed distribution of Poisson and gamma-distributed unobserved heterogeneity and is a type II negative binomial distribution with $\mathrm{E}\left[y\right]=\mu$ and $\mathrm{Var}\left[y\right]=\mu\left(1+\mu/\theta\right)>\mu$ if $\theta>0$.[3]

1.2.2 Interpreting the Mixture Hazard Function

Duration dependence is the process whereby a hazard is not constant over time. If $\mathrm{d}\,\lambda\left(t\right)/\mathrm{d}\,t>0$, it is positive at time t; if $\mathrm{d}\,\lambda\left(t\right)/\mathrm{d}\,t<0$, it is negative at time t. Positive or negative, duration dependence is important when considering economic duration data. For example, the probability of quitting smoking may decline because of rational addiction as a term of unemployment increases. However, if unobserved heterogeneity is present, it is difficult to distinguish hazard rates that decline over time from simple variation in rates across individuals.

Consider the hazard function when unobserved heterogeneity is present in the exponential-gamma mixture model. From (1.7), even if individual hazard is constant at $\mu=\exp\left(\widetilde{\mathbf{x}}'\boldsymbol{\beta}\right)$, where $\widetilde{\mathbf{x}}=\mathbf{x}_i$; for all i, the average or aggregate hazard $\lambda\left(t\right)$ is declining in t. This suggests not negative duration dependence in individual hazard rates but aggregation across individuals whose hazard rates differ randomly. That is, when each person has a constant but different hazard rate, raw data show a decreasing hazard rate because of aggregation across individuals. It is difficult to distinguish hazard rates that decline over time from simple variations across individuals. Moreover, neglecting unobserved heterogeneity may promote underestimating the slope of the hazard function.

1.2.3 Specification of the Heterogeneity Distribution

It is important but difficult to consider unobserved heterogeneity because health econometrics often employs nonlinear models. The preceding analysis assumed parametric unobserved heterogeneity for the sake of computational tractability. However, Heckman and Singer (1984b) point out that parametric specifications of unobserved heterogeneity can be arbitrary and that these assumptions distort parameters to be estimated. Since imposing *ad hoc* restrictions on the functional form of unobserved heterogeneity causes the estimator to be inconsistent, quasi-likelihood methods are inefficient. A parametrically flexible or nonparametric specification is desirable.

[3] Andrews (1988), Cameron and Windmeijer (1996), and Santos Silva (2001) discuss useful specification tests for fully parametric models. Wedel *et al.* (1993), Winkelmann (1996, 2000), Santos Silva (1997b), and Greene (2007b) analyze count data models with various degrees of heterogeneity.

To deal with *ad hoc* parametric distribution of unobserved heterogeneity, Gurmu (1997) and Gurmu *et al.* (1999) propose semiparametric Poisson estimation based on series expansions for the unknown unobserved heterogeneity.[4] They approximate $g(\cdot)$ by Laguerre polynomials, derive the corresponding moment-generating function, and use it to estimate $\boldsymbol{\beta}$ together with additional approximation parameters by maximum likelihood. Doing so avoids the a priori specification of a density function for the unobserved heterogeneity component. They show that the resulting estimator is consistent. We rewrite the marginal PDF of the Poisson mixture model as

$$f(y \mid \mathbf{x}) = \frac{\mu^y}{y!} \int_0^\infty \exp(-\mu\nu)\, \nu^y g(\nu \mid \mathbf{x})\, \mathrm{d}\,\nu \equiv \frac{\mu^y}{y!} M^{(y)}(-\mu), \tag{1.12}$$

where $M^{(y)}(-\mu)$ is regarded as the yth-order moment-generating function of ν. Recall the domain of a moment-generating function

$$M(s) = \int e^{sz} h(z)\, \mathrm{d}\, z. \tag{1.13}$$

Taking y-th order derivatives with respect to s yields

$$M^{(y)}(s) = \int e^{sz} z^y h(z)\, \mathrm{d}\, z = \mathrm{E}\left[e^{sz} z^y\right]. \tag{1.14}$$

For $s = -\mu$ and $z = \nu$, this is precisely the expectation on the right side of (1.12). To specify semiparametric unobserved heterogeneity of $g(\nu)$, let

$$g(\nu) = \frac{1}{S_P} \left[P_K(\nu)\right]^2 w(\nu), \tag{1.15}$$

where $w(\nu)$ is a gamma distribution with parameter (θ, γ) presented in (1.4). Further, $P_K(\nu)$ are the following polynomials that contain a Laguerre series,

$$P_K(\nu) = \sum_{j=0}^{K} c_j h_j L_j^{\alpha-1}(\nu), \tag{1.16}$$

$$L_j^{\alpha-1}(\nu) \equiv \sum_{\ell=0}^{j} \binom{j}{\ell} \frac{\Gamma(j+\alpha)}{\Gamma(\ell+\alpha)\Gamma(j+1)} (-1)^\ell (\gamma\nu)^\ell, \tag{1.17}$$

$$h_j \equiv \frac{\Gamma(j+\alpha)}{\Gamma(\alpha)\Gamma(j+1)}, \tag{1.18}$$

and $S_P = \int_0^\infty P_K(\nu)\, w(\nu)\, \mathrm{d}\,\nu$ ensures integration to 1 by scaling density. Gurmu (1997) and Gurmu *et al.* (1999) demonstrate the analytical solution of (1.12):

$$\begin{aligned} M^{(y)}(-\mu) = {} & \left(1 + \frac{\mu}{\gamma}\right)^{-\theta} (\gamma+\mu)^{-y} \frac{\Gamma(\theta)}{\sum_{j=0}^{K} c_j^2} \sum_{j=0}^{K} \sum_{k=0}^{K} \sum_{\ell=0}^{j} \sum_{m=0}^{k} c_j c_k (h_j h_k)^{\frac{1}{2}} \\ & \times \binom{j}{\ell} \binom{k}{m} \frac{\Gamma(\theta+\ell+m+y)}{\Gamma(\theta+\ell)\Gamma(\theta+m)} \left(-1 - \frac{\mu}{\gamma}\right)^{-(\ell+m)}, \end{aligned} \tag{1.19}$$

[4] This discussion of duration analysis is essentially the same as for count data models.

where $h_j \equiv \Gamma(j+\theta) / (\Gamma(\theta)\Gamma(j+1))$, and $(\boldsymbol{\beta}, \theta, \gamma, c_j), j = 0, \ldots, K$ are parameters to be estimated. The normalization restriction $c_0 = 1$ and the restriction $M^{(1)}|_{-\mu=0} = 1$ are imposed to assure that the mean of unobserved heterogeneity is unity. When $c_j = 0$, $\forall j \geq 1$ (that is, $\gamma = \theta$), the above model results in the preceding Poisson gamma mixture model. Although this model is structurally complex, maximization of the log-likelihood is not complicated.

Another specification of unobserved heterogeneity is Hermite polynomials, which resemble Laguerre polynomials and are a natural extension of a normal distribution. Gallant and Nychka (1987) proposed a semiparametric series estimator that approximates an unknown error term. Based on Gallant and Nychka (1987), we generalize a normally distributed error term using a Hermite series.[5] Let $\varepsilon = \ln \nu$. Then unobserved heterogeneity takes this form:

$$g(\varepsilon) = \frac{1}{S_H}\left[H_K(\varepsilon)\right]^2 w(\varepsilon), \tag{1.20}$$

$$H_K(\varepsilon) \equiv \sum_{k=0}^{K} c_k \varepsilon^k, \tag{1.21}$$

$$w(\varepsilon) \equiv \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{\varepsilon}{\sigma}\right)^2\right), \tag{1.22}$$

where $H_K(\varepsilon)$ are polynomials that contain an Hermite series, c_k is the parameter of the Hermite series to be estimated, and $S_H = \int_{-\infty}^{\infty}\left[H_K(\varepsilon)\right]^2 w(\varepsilon)\,\mathrm{d}\,\varepsilon$ assures integration to 1 by scaling density. Figure 1.2 graphs examples of generalized normal distributions with Hermite series; dashed lines show the standard normal distributions; solid lines show generalized normal distributions $K = 2$. In Figure 1.2, we find that the generalized normal distribution is occasionally skewed and has twin- or triplet-peaks.

Moreover, the generalized normal distribution approximates other famous distributions, such as exponential and gamma distributions. In Figure 1.3 and 1.4, dashed lines show the log-exponential distributions $\mathcal{E}(\gamma)$ and log-gamma distributions $\mathcal{G}(\theta, \gamma)$[6]; solid lines show generalized normal distributions with Hermite series ($K = 5$). The generalized normal distribution is a good candidate for the approximation of the exponential or gamma distribution. This means that the generalized normal distribution is flexible semiparametric one and contains many distributions as a special case. Moreover, since this heterogeneity generalizes additive separable normal distributed heterogeneity, it is easy to interpret in health economics. We briefly discuss unobserved heterogeneity of survival analysis using this Hermite polynomials in Chapter 2.

[5] See also Gabler *et al.* (1993) and van der Klaauw and Koning (2003).

[6] A gamma distribution with $\theta = 1$ results in an exponential distribution $\mathcal{E}(\gamma)$.

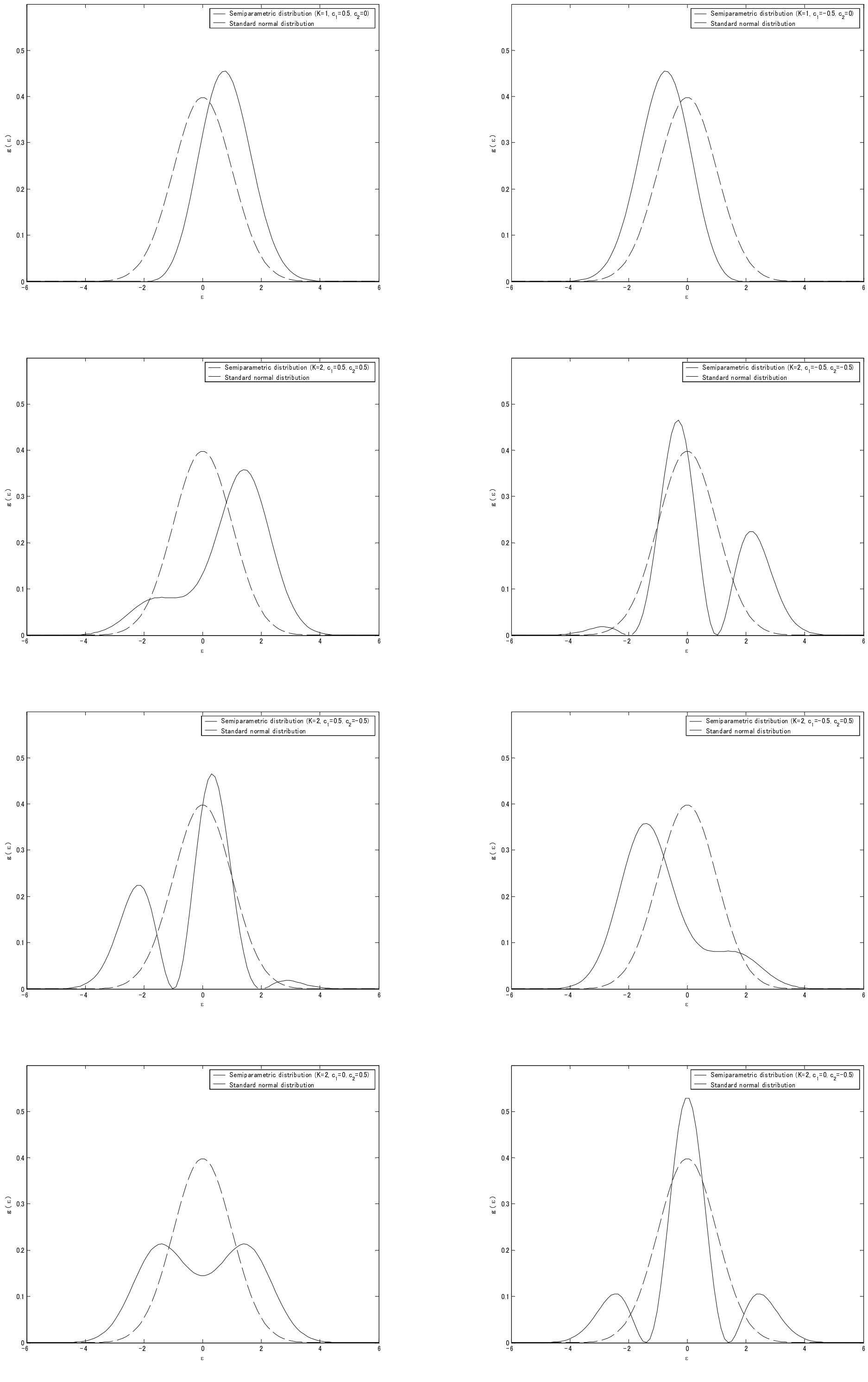


Figure 1.2: PDFs of Generalized Normal Distributions

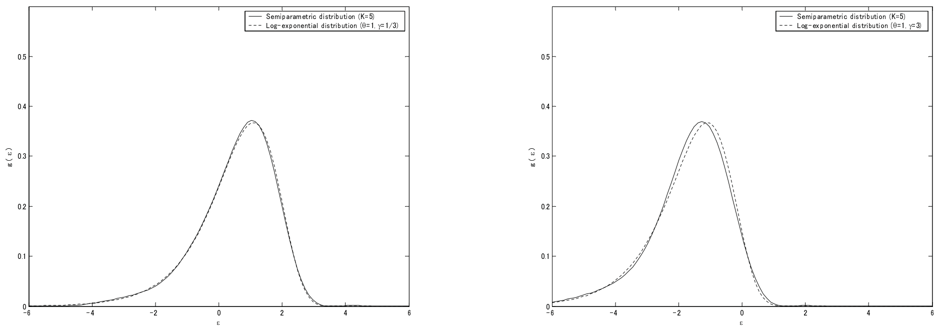


Figure 1.3: PDFs of Generalized Normal and Log-exponential Distributions

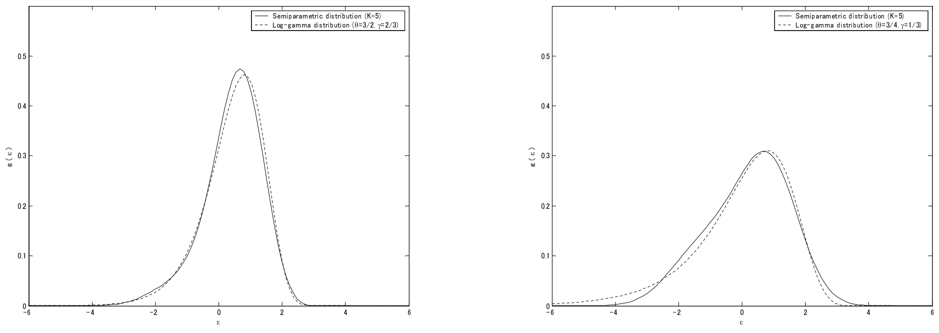


Figure 1.4: PDFs of Generalized Normal and Log-gamma Distributions

1.3 Discrete Heterogeneity

1.3.1 Finite Mixture Models

The preceding discussion assumes that unobserved heterogeneity follows a (parametric or semiparametric) continuous distribution. An alternative approach to consider unobserved heterogeneity is known as a *finite mixture* or *latent class model*. It assumes the sample of individuals is extracted from a population containing a finite number of latent classes j and that each element is drawn from one of these j latent subpopulations or strata. Heckman and Singer (1984a) point out that its attractive features are semiparametric heterogeneity and a flexible parametric distribution.

To simplify discussion, consider a two-component finite mixture model. The mixture density is obtained by

$$f\left(y \mid \mathbf{x}\right) = \pi_1 f_1\left(y \mid \mu_1\left(\mathbf{x}\right)\right) + \left(1 - \pi_1\right) f_2\left(y \mid \mu_2\left(\mathbf{x}\right)\right), \tag{1.23}$$

where $f_1\left(\cdot\right)$ and $f_2\left(\cdot\right)$ are subpopulations, $0 \leq \pi_1 \leq 1$ is a mixing proportion, and $\left(\pi_1, \mu_1, \mu_2\right)$

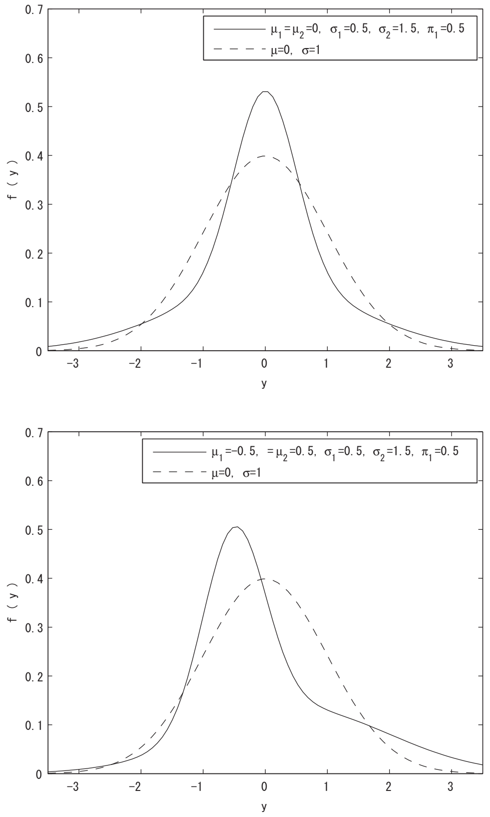
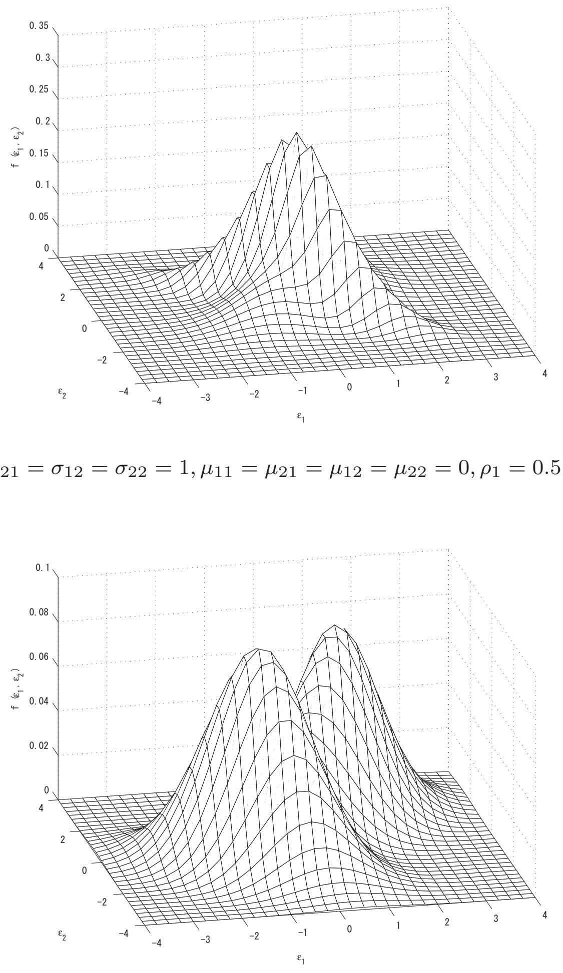


Figure 1.5: PDFs of Finite Mixture Normal Distributions

are parameters to be estimated. Observations are drawn from f_1 and f_2, with probabilities π_1 and $1-\pi_1$, respectively. This means that the finite mixture model has two types of individuals: those extracted from $f_1(\cdot)$ and those extracted from $f_2(\cdot)$.[7]

Figure 1.5 graphs examples of two-component finite mixture normal distributions. Dashed lines show the standard normal distributions. Solid lines show finite mixture normal distributions that differ from variance parameters $(\sigma_j, j=1,2)$ and both the mean $(\mu_j, j=1,2)$ and variance parameters. Figure 1.5 indicates that although these two finite mixture distributions have the same mean and variance as a standard normal distribution, their shapes differ from it; a finite mixture distribution is sometimes fat-tailed and sometimes skewed (or has twin peaks). Moreover, Figure 1.6 shows examples of two-component finite mixture bivariate normal distributions $\mathcal{N}_j\left(\left(\mu_{1j}, \mu_{2j}\right), \left(\sigma_{1j}^2, \rho_j\sigma_{1j}\sigma_{2j}, \sigma_{2j}^2\right)\right)$, $j=1,2$. In the bivariate case, a finite

[7] Finite mixture models are not exclusive to microeconometrics. A Markov switching model used in time series analysis is a type of finite mixture model. Alfò *et al.* (2008) apply finite mixture models to economic growth analysis. See McLachlan and Peel (2000) for statistical features of finite mixture models and a comprehensive review.



(a) $\sigma_{11} = \sigma_{21} = \sigma_{12} = \sigma_{22} = 1, \mu_{11} = \mu_{21} = \mu_{12} = \mu_{22} = 0, \rho_1 = 0.5, \rho_2 = -0.9$

(b) $\sigma_{11} = \sigma_{21} = \sigma_{12} = \sigma_{22} = 1, \mu_{11} = \mu_{21} = 1, \rho_1 = 0.5, \mu_{12} = \mu_{22} = -1, \rho_2 = -0.5$

Figure 1.6: PDFs of Finite Mixture Bivariate Normal Distributions

mixture distribution is sometimes fat-tailed and has twin peaks.

Parameter π_1 is constant or further parameterized using, for example, the logit function $\pi_1 = \exp(\lambda) / [1 + \exp(\lambda)]$. Since the property of J components finite mixture models is essentially that of two components finite mixture models, it is easy to generalize finite mixture models to J components. However, extending finite mixtures to three or more components may interfere with identifiability of the components.

Although the distribution of the unobserved heterogeneity was infinite and continuous in all previous examples, Heckman and Singer (1984a) propose the discrete approximation of population heterogeneity using finite mixture models in duration analysis. Theirs is another interpretation of finite mixture models. If the continuous distribution $g(\nu_i)$ is approximated by a discrete distribution, denoted by π_j $(j = 1, \ldots, J)$ with a finite number of support points

J, then the marginal distribution takes the following form:

$$h\left(y \mid \mathbf{x}, \pi_j, \boldsymbol{\beta}\right) = \sum_{j=1}^{J} f\left(y \mid \mathbf{x}, \nu_j, \boldsymbol{\beta}\right) \pi\left(\nu_j\right), \tag{1.24}$$

where ν_j is an estimated support point and π_j $\left(\sum \pi_j = 1\right)$ is the associated probability.

Moreover, finite mixture models cause another discussion in health econometrics. Recent empirical studies of count data models for medical care demand compare the performance of the two most common approaches: the hurdle (two-part) model and the finite mixture model. The hurdle model that is first discussed by Mullahy (1986) distinguishes the decision to seek care from the level of utilization, and the finite mixture model contains *the frequent and infrequent users' behavior*. The hurdle model focuses on the difference between *users and non-users for medical care demand*. The first part of the hurdle model specifies the decision to seek care, and the second part models the level of utilization. The probability density function of the hurdle model is given by

$$f\left(y\right) = f_1\left(0\right)^d \left[\left(1 - f_1\left(0\right)\right) \times f_T\left(y \,|\, y > 0\right)\right]^{1-d},$$

where y is a count dependent variable that takes a non-integer value and $d = 1 - \min\left(1, y\right)$. The first part of the hurdle model is generally specified as the logistic type; the zero-truncated distribution $f_T\left(y\right)$ assumes the 0 truncated Poisson (or negative binomial) distribution. Because the hurdle model is occasionally regarded as the approximation of the principal-agent hypothesis (Pohlmeier and Ulrich, 1996), and the finite mixture model is considered as standard demand theory, there are many studies to analyze medical care demand using both the models. Gerdtham (1997) analyzes medical care demand using the hurdle model; Deb and Trivedi (1997, 2001, 2002), Deb and Holmes (2000), and Gerdtham and Trivedi (2001) find that the finite mixture model is a better approach for medical care demand. Jemernéz-Martín *et al.* (2002) criticize that the finite mixture model is not based on economics but on statistical reasoning, and find good performance of the hurdle model in EU countries. Bago d'Uva (2005, 2006) analyzes demand for medical care in Britain using the panel finite mixture and panel finite mixture hurdle models. Santos Silva and Windmeijer (2001) propose a multi-episode model for medical care demand and compare with the finite mixture and hurdle models. Winkelmann (2004b) shows that the hurdle model based on bivariate normally distributed heterogeneity surpasses the finite mixture model.[8]

[8] Many studies involve a hurdle model and/or discrete analysis of demand for medical care. For detailed discussions, see Cameron *et al.* (1988), Deb (2001), Deb and Holmes (1998), Manning *et al.* (1987), Mullahy (1997b), Santos Silva (1997a), Santos Silva and Covas (2000), van Outri (2004), Winkelmann (2004a, 2006), and Wang and Alba (2006). Gurmu and Trivedi (1996) and Gurmu (1997, 1998) propose the generalized version of a hurdle model. DeSarbo and Choi (1998) and Martínez-Espiñeira (2006) discuss a double hurdle

1.3.2 Estimation of Finite Mixture Models

Computational problems arise when estimating finite mixture models. Because their log-likelihood is cumbersome and flat, the traditional Newton-Raphson algorithm sometimes works poorly. Even though it produces satisfactory results using numerical derivatives for estimating finite mixture models (Cameron and Trivedi, 2005), the Newton-Raphson algorithm is sensitive to initial values and features problems of local maxima. Thus, the expectation and maximization (EM) algorithm (Dempster *et al.*, 1977) based on latent class models, is often used to estimate finite mixture models.[9] In presenting the EM algorithm, we first explain the latent class interpretation of finite mixture models.

Let $d = (d_1, \ldots, d_J)$ define an indicator (dummy) variable such that $\sum_{j=1}^{J} d_j = 1$ indicates y was drawn from the jth (latent) group for each observation. That is, each observation may be regarded as a sample from one of the j latent subpopulations, classes, or "types." The following discussion assumes the model is identified.

The model specifies that $(y \mid d_j, \mu_j, \pi_j)$ are independently distributed with densities

$$\sum_{j=1}^{J} d_j f\left(y \mid \mu_j\right) = \prod_{j=1}^{J} f\left(y \mid \mu_j\right)^{d_j}, \tag{1.25}$$

where $\mu_j = \mu\left(\mathbf{x}, \boldsymbol{\beta}\right)$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_J)$, and $(d \mid \boldsymbol{\mu}, \pi_j)$ are identically and independently distributed with multinomial distribution

$$\prod_{j=1}^{J} \pi_j^{d_j}, \quad 0 < \pi_j < 1, \quad \sum_{j=1}^{J} \pi_j = 1. \tag{1.26}$$

Therefore, the likelihood function is obtained by

$$\mathcal{L}\left(\boldsymbol{\beta}, \boldsymbol{\pi} \mid y\right) = \prod_{i=1}^{N} \sum_{j=1}^{J} \pi_j^{d_j} f_j\left(y \mid \mu_j\right)^{d_j}, \tag{1.27}$$

which is called a *latent class model.*

Using Bayes' theorem, if π_j is given, the posterior probability that observation y belongs to population j, which denotes z_j, takes this form:

$$z_j = \frac{\pi_j f_j\left(y \mid \mu_j\right)}{\sum_{j=1}^{J} \pi_j f_j\left(y \mid \mu_j\right)}. \tag{1.28}$$

model. Terza and Wilson (1990), Cameron and Johansson (1998), Munkin and Trivedi (1999), Gurmu and Elder (2000), Riphahn *et al.* (2003), Wang (2003), Alfò and Trovato (2004), Cameron *et al.* (2004), and Hellström (2006) analyze bivariate (or multivariate) count data models. The discussion of hurdle (two part) and sample-selection models appears in Duan *et al.* (1983, 1984), Newey *et al.* (1990), Leung and Yu (1996), Melenberg and van Soest (1996), Lopez-Nicolas (1998), Mullahy (1998, 2001), Martins (2001), Chen and Khan (2003), Christofides *et al.* (2003), Dow and Norton (2003), Buntin and Zaslavsk (2004), Lahiri and Xing (2004), Raikou and McGuire (2004), Lee (2005b), and Cantoni and Ronchetti (2006).

[9] McLachlan and Krishnan (1996) show examples of EM algorithms. Ueda and Nakano (1998) propose a modified version of the EM algorithm that attains the global optimum with a high probability

The advantage value of z_j is the probability that a randomly chosen individual belongs to subpopulation j and equals π_j. That is, $\mathrm{E}\left[z_j\right] = \pi_j$.

Now we turn to the EM algorithm. Let the parameter vectors of each density be $\Theta = \left[\mu_1, \ldots, \mu_J, \pi_1, \ldots, \pi_J\right]'$ and $\Theta^{(t)}$ be a parameter vector at the t-th iteration step. In the E-step, given an initial parameter $\Theta^{(0)}$, the EM algorithm directly maximizes the likelihood function (1.27), in which the variable d_j is treated as missing data. Replacing d_j by its expected value $\mathrm{E}\left[d_j\right] = z_j$ yields the conditional expected log-likelihood:

$$Q^{(t)}\left(\Theta \mid \Theta^{(t)}\right) = \sum_{i=1}^{N} \prod_{j=1}^{J} z_j \ln\left[\pi_j f_j\left(y \mid \mu_j\right)\right].$$

The M-step of the algorithm maximizes the $Q^{(t)}$ function, given z_j, by solving the first-order conditions. Further, setting $t \leftarrow t+1$, we calculate new values of z_j obtained by (1.28) and iterate through the E- and M-steps until Θ converges.[10] The EM algorithm is slow to convergence but is easily evaluated because the log-likelihood is separable.

1.3.3 Limitations of Finite Mixture Models

Finite mixture models have limitations. First, their number of components J is unknown a priori, and no theory aids in determining it. Since the dimension of parameters to be estimated is $J \dim[\boldsymbol{\beta}] + J - 1$, where $\boldsymbol{\beta} = \left[\boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_J'\right]'$, we can estimate many parameters. To start with $J = 2$ and then to check the fit of the model using diagnostic tests, an additional component is determined if the fit is poor. When the difference of two or more components is small, the additional component is optional. Moreover, additional components may simply reflect the presence of outliers. The likelihood ratio test is insufficient to select the number of components because of the parameter boundary hypothesis problem. Therefore, we determine the number of components based on the Akaike or the Bayesian information criterion, although many applications use $J = 2$. The parameters are not identified if the model is overparameterized. Overparameterization is revealed by the presence of multiple optima or a flat likelihood surface.

Second, the finite mixture model presents difficulty in maximization. Although the EM algorithm is helpful in understanding the computational structure and theoretically attains the global maxima, it is often slow in attaining the local maxima in practice. As noted, the Newton-Raphson algorithm based on numerical derivatives may work well in practice, but it is cumbersome for checking local maxima. Other algorithms estimate finite mixture models, such as the simulated annealing EM and stochastic evolution algorithms. Both achieve global

[10] For application of the EM algorithm to econometrics, see Ruud (1991), Nielsen (2000), Arcidiacono and Jones (2003), and Ferrall (2005).

maxima with a high probability, but convergence is slow.

Third and most important, some finite mixture models cannot be estimated for an identification problem. Health econometrics often employs cross-sectional binary dependent variable models, such as the probit or logit model. However, the following finite mixture *binary* model is not identified:

$$f(y) = \sum_{j=1}^{J} \pi_j \begin{pmatrix} T \\ y \end{pmatrix} \left[F\left(\mathbf{x}'\boldsymbol{\beta}_j\right) \right]^y \left[1 - F\left(\mathbf{x}'\boldsymbol{\beta}_j\right) \right]^{T-y}, \tag{1.29}$$

where T is the number of Bernoulli trials and $0 < F(\cdot) < 1$ is a cumulative density function, usually specified as a standard normal or logit distribution. When $T = 1$, the result is a cross-sectional finite mixture binary model; when $T \geq 2$, the result is a panel data finite mixture binary model. Teicher (1960, 1963) and Blischke (1978) obtain necessary and sufficient conditions for identifiability of a finite mixture binomial model. Their results summarize as follows: The J-component finite mixture binary model is identifiable if and only if

$$J \leq (T + 1)/2. \tag{1.30}$$

Therefore, even if $J = 2$, *any cross-sectional finite mixture binary model is not identified* and panel finite mixture binary models with $T \geq 3$ are identified.

Consequently, it is impossible to analyze a cross-sectional binary outcome using finite mixture models in health econometrics. This problem has remained unsolved since Teicher's demonstration and is troublesome because available data often are limited to cross-sectional binary variables. In this case, we must apply probit or logit estimation even if a true data-generating process is a finite mixture. Of course, the estimated covariates are inconsistent and the results lack value. Therefore, we take up estimating cross-sectional finite mixture binary models in Chapter 3. We demonstrate that finite mixture binomial models sometimes are identifiable and that this type of model has a consistent estimator.

Chapter 2

Semiparametric Estimation of Regression-Based Survival Models

2.1 Introduction[1]

Survival analysis is widespread in applied econometrics such as labor economics, industrial organizations, health economics, and population economics. Many survival models, including the most popular Cox's proportional hazard model, do not explicitly assume unobserved heterogeneity. It is empirically difficult to separate the effect of duration dependence from those of unobserved heterogeneity. However, in social science, the existence of omitted variables is inevitable, and it is always inadequate to control population heterogeneity. Therefore, the model without unobserved heterogeneity overestimates (or underestimates) the degree of negative (positive) duration dependence in the hazard, even if this model has consistency of coefficients.

One way to introduce unobserved heterogeneity into survival models is to assume that the heterogeneity is multiplicative to the hazard function and follows a gamma distribution. This method is simple and has a closed form solution. However, this is a parametric method, and the distribution assumption of heterogeneity is rather important. Furthermore, we do not determine the econometric interpretation of this method if the probability density function is not an exponential or Weibull distribution.

In this chapter, we propose new semiparametric (semi-nonparametric) survival models

[1] This chapter is the modified version of Masuhara (2007).

that generalize unobserved heterogeneity, as well as a dependent variable of the log-normal survival model. First, we generalize the log-transformed dependent variable using the Box-Cox transformation, which contains various function forms. Second, we generalize the normally distributed unobserved heterogeneity using Hermite polynomials, which include a normal distribution as a special case. In the empirical application, this chapter compares the performance of the proposed models, using the General Social Survey (GSS) in 2002 following Winkelmann and Boes (2006).

This chapter is organized as follows. Section 2 proposes the semiparametric regression-based survival model. Section 3 depicts the application of the fertility data, and Section 4 presents our concluding remarks.

2.2 The Model

We consider the standard survival analysis. Suppose that a random variable, $T_i, i = 1, \ldots, N$, has a continuous probability distribution $f(t_i)$, where t_i is a realization of T_i. The cumulative distribution function of this variable takes the following form: $F(t_i) = \int_0^{t_i} f(s_i)\,\mathrm{d}\,s_i$. We define a censoring indicator $c_i = 1$ if the observation is censored, and $c_i = 0$ if the observation is uncensored. Further, the log-likelihood function is obtained as follows: $\ln L = \sum_{i=1}^{N} (1 - c_i) \ln f(t_i) + c_i \ln [1 - F(t_i)]$.

In the log-normal survival model, a logarithmic survival variable consists of a linear index of independent variables and an additive normally distributed error term: $\ln t_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i$, where $\mathbf{x}_i$ is a $K \times 1$ covariate vector of covariates and $\boldsymbol{\beta} \sim K \times 1$ is a parameter to be estimated. Although the log-normal model clearly assumes heterogeneity, it has some disadvantages. First, the transformation of the dependent variable t_i is quite arbitrary. The log transformation may not always be the best choice. Second, this model is parametric; we assume the normally distributed unobserved heterogeneity and estimate the parameters using the maximum likelihood (ML) method. However, the incorrect specification of the error term causes the inconsistency of the ML. Therefore, we require a more flexible method to estimate survival data.

First, we generalize the log-transformed dependent variable using the Box-Cox transformation, such as $\left(t_i^{\lambda} - 1\right)/\lambda = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i$, where λ is a parameter of the Box-Cox transformation. In this model, when $\lambda \to 0$, the left-hand side (LHS) is $\ln(t_i)$, and when $\lambda = 1$, the LHS is $t_i - 1$. The Box-Cox transformation contains the log-linear and linear models as a special case.[2]

Second, we generalize the error term. If the relation between the error term and the

[2] Cai *et al.* (2005) also use the Box-Cox transformation in duration analysis.

survival random variable is quasi-linear, like the log-normal model, we obtain the following conditions using the Jacobian of the transformation $\mathrm{d}\,t_i/\mathrm{d}\,\varepsilon_i$: $f(\varepsilon_i) = f(t_i)\,\mathrm{d}\,t_i/\mathrm{d}\,\varepsilon_i$ and $F(\varepsilon_i) = F(t_i)$. Therefore, we concentrate the distribution of ε_i. Following Gallant and Nychka (1987), who proposed the semiparametric (semi-nonparametric) series based on a normal distribution, we approximate the unknown error term using Hermite polynomials.[3] The approximated density is obtained as follows:

$$f(\varepsilon_i) = \frac{1}{P}\left(\sum_{k=0}^{K} \alpha_k \varepsilon_i^k\right)^2 \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{\varepsilon_i}{\sigma}\right)^2\right) \equiv \frac{f^*(\varepsilon_i)}{P}, \tag{2.1}$$

where σ is a standard deviation parameter, α_k is a parameter of a Hermite series to be estimated, and $P = \int_{-\infty}^{\infty} f^*(\varepsilon_i)\,\mathrm{d}\,\varepsilon_i$ ensures integration to 1 by scaling density. Moreover, we impose $\alpha_0 = 1$ to ensure the identification of the parameters. This Hermite series density contains not only a normal distribution as a special case but also fat-tailed and twin peak distributions.

We discuss two generalizations for survival data and term the above model as the Box-Cox transformed semi-nonparametric (BC-SN), or semiparametric model. This model generalizes not only the error terms using Hermite polynomials but also the dependent variable using the transformation of t_i, and includes the log-normal model as a special case. In estimating survival data, the calculation of $F(\varepsilon_i) = \int_{-\infty}^{\varepsilon_i} f(\nu_i)\,\mathrm{d}\,\nu_i$ is required. Fortunately, this integral has a closed-form solution and is easy to evaluate.

The log-likelihood function is usually maximized by gradient-based methods such as Newton, BFGS, or BHHH algorithms. However, these algorithms are sensitive to initial values and contain the problem of local maxima. It is difficult to estimate parameters α_k of the semiparametric model correctly, and it is practically impossible to attempt several initial conditions. To avoid the problem of local maxima, we use the stochastic evolution algorithm (StocE), which attains the global maxima with a high probability (Saab and Rao, 1990; Sait and Youssef, 2000).

Let the parameter vector of this density be $\boldsymbol{\theta} = [\boldsymbol{\beta}', \sigma, \lambda, \alpha_1, \alpha_2, \ldots]'$. Further, the StocE algorithm obtains the parameters as follows:

1) Set the initial parameters p_0, $\boldsymbol{\theta}_0$, R, p_{up}, r and $t \leftarrow 0$. Calculate the log-likelihood $\ln L_0$ under $\boldsymbol{\theta}_0$.

2) Set $\boldsymbol{\theta}_{\text{best}} = \boldsymbol{\theta}_0$ and $\ln L_{\text{best}} = \ln L_0$.

3) Generate $\hat{\boldsymbol{\theta}}$ in the neighborhood of $\boldsymbol{\theta}$; e.g., calculate $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_t + a \cdot u_1$ and $\ln \hat{L}\left(\hat{\boldsymbol{\theta}}\right)$, where a is a positive constant number and u_1 is a uniform random vector on $[-1,\ 1]$.

[3] For more discussion, see Gallant (1981), Gallant and Tauchen, Gabler *et al.* (1993), Coppejans (2001), Coppejans and Gallant (2002), van der Klaauw and Koning (2003), and Stewart (2004, 2005).

4) If $\ln \hat{L} - \ln L_t > -p_t \cdot u_2$, where u_2 is a uniform random number, then $\boldsymbol{\theta}_{t+1} = \hat{\boldsymbol{\theta}}$; otherwise, $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t$. Calculate $\ln L(\boldsymbol{\theta}_{t+1})$.

5) If $\ln L(\boldsymbol{\theta}_{t+1}) = \ln L(\boldsymbol{\theta}_t)$, then set $p_{t+1} = p_t + p_{\text{up}}$; otherwise, $p_{t+1} = p_0$.

6) If $\ln L(\boldsymbol{\theta}_{t+1}) > \ln L_{\text{best}}$, then $\boldsymbol{\theta}_{\text{best}} = \boldsymbol{\theta}_{t+1}$, $\ln L_{\text{best}} = \ln L_{t+1}$, and $r = r - R$; otherwise, $r = r + 1$. Set $t \leftarrow t + 1$.

7) Repeat steps 3 to 6 until $r > R$.

The StocE algorithm resembles the simulated annealing (SA) method [4], which is a stochastic technique using the Metropolis algorithm. If the annealing process is slow, the SA algorithm is similar to a random search and its convergence speed is slow. The StocE algorithm eliminates the inefficient path with a probability $p_t u_2$. Therefore, in general, the StocE algorithm is faster than the SA. However, the StocE algorithm does not ensure the convergence of the global maximum, because the sequence of the StocE is not a perfect Markov chain. Nonetheless, in practice, the StocE is effective.

2.3 An Application to Fertility

We present the results of the application of the proposed model. In this example, we analyze the factors that affect the time until a woman bears her first child, using the data obtained from the GSS in 2002, an annual or biannual cross-section survey that began in 1972.Following Winkelmann and Boes (2006), we regard the variable age at the time of the first child's birth as duration and employ women under the age of 40. The total number of observations is 1,371. There are two types of women: Type A includes women who have had their first child (the number of uncensored sample is 1,154); Type B includes women who are childless under the age of 40 (the number of right-censored samples is 217). Moreover, this survey considers the number of years of formal schooling, the number of siblings, four dummy variables, namely, those in the low-income group at age 16 (less than average income), those who were urban residents at age 16, those who are white, and those who are immigrants. The summary statistics of these data are obtained by Table 2.1.

We investigate the result of the BC-SN model but find that the parameter λ is nearly zero (and not significant). Further, we estimate the semi-nonparametric (LN-SN), or semiparametric model with log transformation ($\lambda = 0$). For comparison purposes, we include the standard log-normal survival model and the log-normal model with gamma distributed unobserved heterogeneity (LN-G). Table 2.2 shows the estimated results of the four models

[4] See Goffe *et al.* (1994).

Table 2.1: Age at First Birth: Variable Description

Variable	Mean	Std. Dev.	Min.	Max.
age at first birth	23.463	5.340	11	39
dummy of right censored sample	0.158	0.365	0	1
years of education	13.241	2.783	0	20
number of siblings	3.694	2.947	0	23
white	0.761	0.427	0	1
immigrant	0.123	0.328	0	1
low income at age 16	0.199	0.400	0	1
lived in city at age 16	0.458	0.498	0	1
number of kids	2.144	1.604	0	8
age at 2002	45.692	17.720	19	89

Data: GSS 2002. The data are downloadable from
http://www.uzh.ch/sts/research/publications/microdata/index.html.

and also presents the values of the log-likelihood, Akaike's information criteria (AIC), and Bayesian information criteria (BIC). The maximum value of the log-likelihood and the minimum value of the AIC is the LN-SN model (the reason for the positive log-likelihood is $\sigma < 1$). The minimum value of the BIC is the LN-G model. Further, we use Vuong's (1989) test to select the unique model between the LN-SN and LN-G models. Let $f^{(1)}$ be the likelihood of model 1 and $f^{(2)}$ be that of model 2. Under the null hypothesis that both the models are equivalent ($\mathrm{E}\left[\ln f^{(1)} - \ln f^{(2)}\right] = 0$), the test statistic

$$\frac{\sum_{i=1}^{N}\left[\ln f^{(1)} - \ln f^{(2)}\right]}{\sqrt{\sum_{i=1}^{N}\left[\left(\ln f^{(1)} - \ln f^{(2)}\right)^2 - \left(\left(\sum_{i=1}^{N}\ln f^{(1)} - \ln f^{(2)}\right)/N\right)^2\right]}} \tag{2.2}$$

follows a standard normal distribution. If the statistic exceeds the critical value c, model $f^{(1)}$ is better than model $f^{(2)}$. If the statistic is smaller than $-c$, model $f^{(2)}$ is better than model $f^{(1)}$. The test statistic for the LN-SN model against the LN-G model is 1.252, and its value at the significant level is 0.105.[5] Hence, there is (weak) evidence that the LN-SN model is the best of the four.

In Table 2.2, we find three features of the estimated parameters. First, the estimated parameters of the four models closely resemble each other. Second, the significant level of the estimated parameters also resemble each other. Third, the values of the estimated parameters

[5] The tests for the LN-SN model, the BC-SN model, and the LN-G model against the log-normal model are 6.327, 6.320, and 4.999, respectively. This means that the log-normal assumption is strongly rejected.

Table 2.2: Estimation Results of Age at First Birth

	log-normal	LN-G	LN-SN	BC-SN
years of education	0.031	0.032	0.031	0.031
	(0.002)	(0.002)	(0.002)	(0.002)
number of siblings	-0.005	-0.003	-0.004	-0.004
	(0.002)	(0.002)	(0.002)	(0.002)
white	0.083	0.091	0.088	0.088
	(0.015)	(0.013)	(0.012)	(0.012)
immigrant	0.056	0.060	0.054	0.054
	(0.019)	(0.018)	(0.016)	(0.016)
low income at age 16	-0.021	-0.038	-0.031	-0.031
	(0.016)	(0.015)	(0.013)	(0.013)
lived in city at age 16	0.008	-0.004	-0.004	-0.004
	(0.013)	(0.011)	(0.011)	(0.011)
constant	2.696	2.615	2.693	2.693
	(0.038)	(0.036)	(0.032)	(0.032)
σ	0.219	0.150	0.214	0.214
	(0.005)	(0.007)	(0.004)	(0.004)
γ^{-1}		0.673		
		(0.085)		
λ				0.000
				(0.001)
α_1			-1.555	-1.555
			(0.065)	(0.065)
α_2			-0.171	-0.166
			(0.211)	(0.208)
α_3			15.489	15.493
			(0.379)	(0.379)
α_4			1.458	1.447
			(0.728)	(0.735)
α_5			-19.030	-19.039
			(1.139)	(1.139)
log-likelihood	-72.933	-8.699	0.534	0.440
AIC	161.865	35.398	24.932	27.119
BIC	203.652	82.407	92.835	100.245
number of regressors	8	9	13	14

Notes: Standard errors are in parentheses; LN-G, LN-SN, and BC-SN denote the log-normal model with gamma-distributed unobserved heterogeneity, the semi-nonparametric model with the log transformation, and the Box-Cox transformed semi-nonparametric, respectively; AIC $= -2\ln L + 2K$, BIC $= -2\ln L + K\ln N$, where L is the maximized likelihood, K is the number of parameters, and N is the number of observations (N = 1,371); γ is the parameter of the gamma frailty.

of the BC-SN and LN-SN models are midway of those of the log-normal and LN-G models. The reason for these features is that the BC-SN, LN-SN, and LN-G models are based on and extended to the standard log-normal model.

However, there are large differences among the four models. For example, in the log-normal

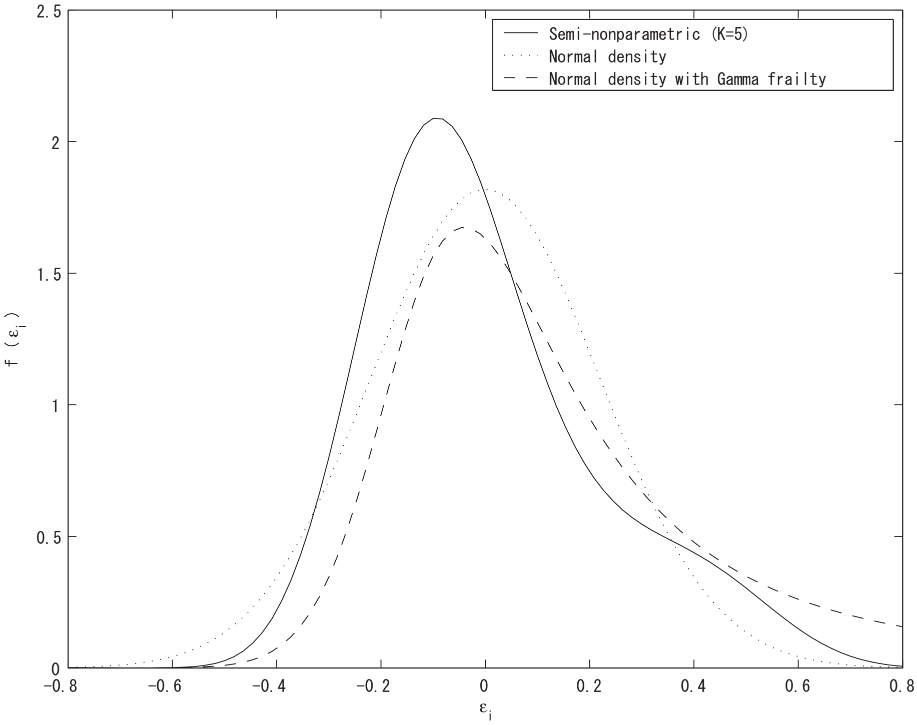


Figure 2.1: The Estimated Density Function of ε_i

model, the value of *low income at age 16* is -0.021 and is not statistically significant; in all the remaining models, its value ranges from -0.031 to -0.038 and is statistically significant at the 5% level. Moreover, the value of *white* is 0.091 in the LN-G model and 0.088 in the LN-SN model. The difference between these two models is small but negligible in these data.

Figure 2.1 shows the estimated density of ε_i. The solid line is the LN-SN model,[6] the dotted line is the log-normal model, and the dashed line is the LN-G model. The density of the LN-SN model is skewed to the left and has a fatter tail. However, the density of the LN-G model has the fattest tail among the three models.

2.4 Conclusion

This chapter proposes new semiparametric survival models that generalize both an explanatory variable and unobserved heterogeneity. The former is Box-Cox transformation and the latter is a Hermite series. In an example using the GSS data, the Box-Cox transformation does not work well. However, the LN-SN model overcomes the other models, except for the BIC,

[6] The densities of the LN-SN and BC-SN models are almost identical. Thus, we omit the latter in Figure 2.1.

and shows a good performance. Therefore, there is (weak) evidence that the LN-SN model is the best of the four. Moreover, the coefficients of the models, except for the log-normal model, resemble each other but the difference is not negligible.

In both econometric and statistic interpretations, the results may show a small difference between the LN-G and semiparametric models. This view is incorrect. The LN-G model does not have a clear econometric interpretation because this model assumes multiplicative heterogeneity not to the probability density function but to the *hazard* function. That is, the LN-G model explicitly assumes multi-heterogeneity. Therefore, in applied econometrics, the semiparametric model proposed in this chapter is an important method to estimate a demand or supply function because it assumes only single and additively separable heterogeneity.

Appendix

A. The Box-Cox Transformed Semi-Nonparametric Survival Model

Following Gabler *et al.* (1993), (2.1) takes

$$f\left(\varepsilon_i\right) = \frac{1}{P}\left(\sum_{k=0}^{K} \alpha_k \varepsilon_i^k\right)^2 \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{\varepsilon_i}{\sigma}\right)^2\right) \equiv \frac{f^*\left(\varepsilon_i\right)}{P}, \tag{A2.1}$$

where $\varepsilon_i = \left(t_i^\lambda - 1\right)/\lambda - x_i'\boldsymbol{\beta}$ and

$$P = \int_{-\infty}^{\infty} \left(\sum_{k=0}^{K} \alpha_k \varepsilon_i^k\right)^2 \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{\varepsilon_i}{\sigma}\right)^2\right) \mathrm{d}\,\varepsilon_i. \tag{A2.2}$$

We require algebraic computations of (A2.2). Van der Klaauw and Koning (2003) show the following recursion formulas:

$$I_k\left(a, b\right) = \int_a^b u^k \exp\left(-\left(\frac{u}{\delta}\right)^2\right) \mathrm{d}\,u. \tag{A2.3}$$

Equation (A2.3) obtains

$$I_j\left(-\infty, \infty\right) = \begin{cases} \delta\sqrt{\pi}, & j = 0. \\ 0, & j = 1, 3, 5, \ldots \\ \frac{(j-1)\delta^2}{2} I_{j-2}\left(-\infty, \infty\right), & j = 2, 4, 6, \ldots \end{cases} \tag{A2.4}$$

Substituting $\delta = \sqrt{2}\sigma$ into (A2.4) and calculating up to $j = 10$ yields

$$I_0 = \sqrt{2}\sigma\sqrt{\pi}, \qquad I_2 = \frac{1}{2}\left(\sqrt{2}\sigma\right)^3\sqrt{\pi},$$

$$I_4 = \frac{3}{4}\left(\sqrt{2}\sigma\right)^5\sqrt{\pi}, \qquad I_6 = \frac{15}{8}\left(\sqrt{2}\sigma\right)^7\sqrt{\pi},$$

$$I_8 = \frac{105}{16}\left(\sqrt{2}\sigma\right)^9\sqrt{\pi}, \qquad I_{10} = \frac{945}{32}\left(\sqrt{2}\sigma\right)^{11}\sqrt{\pi}.$$

When $K = 5$, we obtain the following relation after some algebraic manipulation.

$$\begin{aligned}\left(\sum_{k=0}^{5}\alpha_k\varepsilon_i^k\right)^2 =& \left(\alpha_0^2\right) + \varepsilon_i\left(2\alpha_0\alpha_1\right) + \varepsilon_i^2\left(2\alpha_0\alpha_2 + \alpha_1^2\right) + \varepsilon_i^3\left(2\alpha_0\alpha_3 + 2\alpha_1\alpha_2\right)\\ &+ \varepsilon_i^4\left(2\alpha_0\alpha_4 + 2\alpha_1\alpha_3 + \alpha_2^2\right) + \varepsilon_i^5\left(2\alpha_0\alpha_5 + 2\alpha_1\alpha_4 + 2\alpha_2\alpha_3\right)\\ &+ \varepsilon_i^6\left(2\alpha_1\alpha_5 + 2\alpha_2\alpha_4 + \alpha_3^2\right) + \varepsilon_i^7\left(2\alpha_2\alpha_5 + 2\alpha_3\alpha_4\right)\\ &+ \varepsilon_i^8\left(2\alpha_3\alpha_5 + \alpha_4^2\right) + \varepsilon_i^9\left(2\alpha_4\alpha_5\right) + \varepsilon_i^{10}\left(\alpha_5^2\right). \end{aligned} \tag{A2.5}$$

Substituting (A2.3), (A2.4), and (A2.5) into (A2.2) yields

$$\begin{aligned}P =& \alpha_0^2 + \left(2\alpha_0\alpha_2 + \alpha_1^2\right)\frac{1}{2}\left(\sqrt{2}\sigma\right)^2 + \left(2\alpha_0\alpha_4 + 2\alpha_1\alpha_3 + \alpha_2^2\right)\frac{3}{4}\left(\sqrt{2}\sigma\right)^4\\ &+ \left(2\alpha_1\alpha_5 + 2\alpha_2\alpha_4 + \alpha_3^2\right)\frac{15}{8}\left(\sqrt{2}\sigma\right)^6\\ &+ \left(2\alpha_3\alpha_5 + \alpha_4^2\right)\frac{105}{16}\left(\sqrt{2}\sigma\right)^8 + \left(\alpha_5^2\right)\frac{945}{32}\left(\sqrt{2}\sigma\right)^{10}. \end{aligned} \tag{A2.6}$$

Therefore, the probability density function of heterogeneity consists of (A2.1) and (A2.6). To ensure a zero mean, we impose the restriction $\alpha_0 = 1$ and $\mathrm{E}\left[\varepsilon_i\right] = 0$. The expectation term $\mathrm{E}\left[\varepsilon_i\right]$ takes the form as follows:

$$\begin{aligned}\mathrm{E}\left[\varepsilon_i\right] =& \frac{1}{P}\int_{-\infty}^{\infty}\varepsilon_i f^*\left(\varepsilon_i\right)\mathrm{d}\,\varepsilon_i\\ =& \frac{1}{P}\left[2\alpha_1\left(2\alpha_0 I_2 + 2\alpha_2 I_4 + 2\alpha_4 I_6\right)\right.\\ &+ 2\alpha_3\left(2\alpha_0 I_4 + 2\alpha_2 I_6 + 2\alpha_4 I_8\right)\\ &\left.+2\alpha_5\left(2\alpha_0 I_6 + 2\alpha_2 I_8 + 2\alpha_4 I_{10}\right)\right]. \end{aligned} \tag{A2.7}$$

From the (A2.7), we obtain the following relation:

$$\alpha_5 = -\frac{\alpha_1\left(\alpha_0\delta^3 + \frac{3}{2}\alpha_2\delta^5 + \frac{15}{4}\alpha_4\delta^7\right) + \alpha_3\left(\frac{3}{2}\alpha_0\delta^5 + \frac{15}{4}\alpha_2\delta^7 + \frac{105}{8}\alpha_4\delta^9\right)}{\left(\frac{15}{4}\alpha_0\delta^7 + \frac{105}{8}\alpha_2\delta^9 + \frac{945}{16}\alpha_4\delta^{11}\right)}. \tag{A2.8}$$

Next, we present the calculation of $F\left(\varepsilon_i\right) = \int_{-\infty}^{\varepsilon_i} f\left(\nu_i\right)\mathrm{d}\,\nu_i$. When $K = 5$, we obtain the following relation using (A2.3):

$$\begin{aligned}
I_0\left(-\infty,\varepsilon_i\right) &= \sqrt{2\pi}\sigma\Phi\left(\frac{\varepsilon_i}{\sigma}\right),\\
I_1\left(-\infty,\varepsilon_i\right) &= -\sigma^2\exp\left(-\frac{1}{2}\left(\frac{\varepsilon_i}{\sigma}\right)^2\right),\\
I_2\left(-\infty,\varepsilon_i\right) &= -\sigma^2\varepsilon_i\exp\left(-\frac{1}{2}\left(\frac{\varepsilon_i}{\sigma}\right)^2\right)+\sigma^2 I_0\left(-\infty,\varepsilon_i\right),\\
I_3\left(-\infty,\varepsilon_i\right) &= -\sigma^2\varepsilon_i^2\exp\left(-\frac{1}{2}\left(\frac{\varepsilon_i}{\sigma}\right)^2\right)+2\sigma^2 I_1\left(-\infty,\varepsilon_i\right),\\
I_4\left(-\infty,\varepsilon_i\right) &= -\sigma^2\varepsilon_i^3\exp\left(-\frac{1}{2}\left(\frac{\varepsilon_i}{\sigma}\right)^2\right)+3\sigma^2 I_2\left(-\infty,\varepsilon_i\right),\\
I_5\left(-\infty,\varepsilon_i\right) &= -\sigma^2\varepsilon_i^4\exp\left(-\frac{1}{2}\left(\frac{\varepsilon_i}{\sigma}\right)^2\right)+4\sigma^2 I_3\left(-\infty,\varepsilon_i\right),\\
I_6\left(-\infty,\varepsilon_i\right) &= -\sigma^2\varepsilon_i^5\exp\left(-\frac{1}{2}\left(\frac{\varepsilon_i}{\sigma}\right)^2\right)+5\sigma^2 I_4\left(-\infty,\varepsilon_i\right),\\
I_7\left(-\infty,\varepsilon_i\right) &= -\sigma^2\varepsilon_i^6\exp\left(-\frac{1}{2}\left(\frac{\varepsilon_i}{\sigma}\right)^2\right)+6\sigma^2 I_5\left(-\infty,\varepsilon_i\right),\\
I_8\left(-\infty,\varepsilon_i\right) &= -\sigma^2\varepsilon_i^7\exp\left(-\frac{1}{2}\left(\frac{\varepsilon_i}{\sigma}\right)^2\right)+7\sigma^2 I_6\left(-\infty,\varepsilon_i\right),\\
I_9\left(-\infty,\varepsilon_i\right) &= -\sigma^2\varepsilon_i^8\exp\left(-\frac{1}{2}\left(\frac{\varepsilon_i}{\sigma}\right)^2\right)+8\sigma^2 I_7\left(-\infty,\varepsilon_i\right),\\
I_{10}\left(-\infty,\varepsilon_i\right) &= -\sigma^2\varepsilon_i^9\exp\left(-\frac{1}{2}\left(\frac{\varepsilon_i}{\sigma}\right)^2\right)+9\sigma^2 I_8\left(-\infty,\varepsilon_i\right).
\end{aligned} \tag{A2.9}$$

Therefore,

$$\begin{aligned}
F\left(\varepsilon_i\right) =& \int_{-\infty}^{\varepsilon_i} f\left(\nu_i\right)\mathrm{d}\,\nu_i\\
=&\Big[\alpha_0^2 I_0\left(-\infty,\varepsilon_i\right)+2\alpha_0\alpha_1 I_1\left(-\infty,\varepsilon_i\right)\\
&+\left(2\alpha_0\alpha_2+\alpha_1^2\right)I_2\left(-\infty,\varepsilon_i\right)+\left(2\alpha_0\alpha_3+2\alpha_1\alpha_2\right)I_3\left(-\infty,\varepsilon_i\right)\\
&+\left(2\alpha_0\alpha_4+2\alpha_1\alpha_3+\alpha_2^2\right)I_4\left(-\infty,\varepsilon_i\right)\\
&+\left(2\alpha_0\alpha_5+2\alpha_1\alpha_4+2\alpha_2\alpha_3\right)I_5\left(-\infty,\varepsilon_i\right)\\
&+\left(2\alpha_1\alpha_5+2\alpha_2\alpha_4+\alpha_3^2\right)I_6\left(-\infty,\varepsilon_i\right)\\
&+\left(2\alpha_2\alpha_5+2\alpha_3\alpha_4\right)I_7\left(-\infty,\varepsilon_i\right)+\left(2\alpha_3\alpha_5+\alpha_4^2\right)I_8\left(-\infty,\varepsilon_i\right)\\
&+2\alpha_4\alpha_5 I_9\left(-\infty,\varepsilon_i\right)+\alpha_5^2 I_{10}\left(-\infty,\varepsilon_i\right)\Big]\times\frac{1}{P}\frac{1}{\sqrt{2\pi}\sigma}.
\end{aligned} \tag{A2.10}$$

Chapter 3

An Alternate Approach to Estimate a Finite Mixture Cross-Sectional Probit Model [1]

3.1 Introduction

Because finite mixture models are semiparametric and flexible, they enjoy widespread use in applied econometrics, including health economics. The Markov switching model in time series analysis is a specific version of finite mixture models. McLachlan and Peel (2000) present a comprehensive account of finite mixture models and examples of their applications. However, some finite mixture models—a finite mixture *cross-sectional* probit (binomial) model for example—are not estimated for an identification problem.

Teicher (1960, 1963) and Blischke (1964) indicated this problem and presented necessary and sufficient conditions for the identifiability of a finite mixture probit model. Their results summarize as follows: The J component probit mixture with T Bernoulli trials is identifiable if and only if $J \leq (T+1)/2$. In other words, a two-component finite mixture *cross-sectional* probit model is not identifiable and the *panel* probit model with $T \geq 3$ is.[2] Hettmansperger and Thomas (2000), Cruz-Medina *et al.* (2004), and Elmore and Wang (2003) present the same results in a binomial model. Hall *et al.* (2005) show that the sufficient condition for identification in a binomial model requires $T \geq (1 + o(1)) 6J \ln J$ as $J \to \infty$. Kasahara and

[1] This chapter was supported by JSPS Grant-in-Aid for Young Scientists (B) Number 26780137.

[2] If unrealistic restrictions are imposed on the covariates, this finite mixture cross-sectional probit model is identifiable. See Follmanna and Lambert (1989) and Brooks *et al.* (1997).

Shimotsu (2010) demonstrate that in finite mixture binomial models the number of components is nonparametrically identified if $T \geq 2$ and the mixing proportions and distribution of components are identified when $T \geq 3$. That is, the above results show it is impossible to estimate a finite mixture cross-sectional probit model.

However, many applied econometric analyses have only cross-sectional data with a binary endogenous variable. In this situation, even if a *true data generating process* is a finite mixture, we need only apply the cross-sectional probit (or logit) analysis to these data because we do not estimate a finite mixture probit model for an identification problem. Of course, the estimated coefficients of this analysis are inconsistent and the result is unreliable. This problem is cumbersome. Hence, this chapter analyzes the possibility of estimating a finite mixture cross-sectional probit model. First, we show the identifiability of bivariate random variables using a natural expansion of Teicher's (1963) theorem. Second, using this result, this chapter investigates the identifiability of a finite mixture *cross-sectional* probit model with a linear single equation. This chapter demonstrates that the class of all finite mixtures of a probit model with a linear single equation is identifiable even if the number of components exceeds three. That is, sometimes a finite mixture cross-sectional probit model can be estimated.

This chapter is organized as follows. Section 2 proposes three corollaries according to a finite mixture probit model using Teicher's (1963) theorem. Section 3 depicts the results of Monte Carlo experiments. Section 4 concludes.

3.2 Identification of Probit Finite Mixtures

Teicher (1963) has shown the necessary and sufficient conditions of identifiability for univariate finite mixtures. The bivariate version of Theorem 2 in Teicher (1963) takes the following form:

Corollary 1. *Let $\mathscr{F} = \{F\}$ be a family of the cumulative density function (CDF) with the moment-generating function (MGF) $M(t_1, t_2)$ defined for $t_1 \in S_{M_{\cdot,1}}$ and $t_2 \in S_{M_{\cdot,2}}$, where $S_{M_{\cdot,1}}$ and $S_{M_{\cdot,2}}$ are the domains of definition of M and the mapping $M : F \to M$ is linear and one-to-one. Suppose there is a total ordering $\preceq$ of $\mathscr{F}$. The relation $F_1 \prec F_2$ implies: (i) $S_{M_{1,1}} \subseteq S_{M_{2,1}}$ and $S_{M_{1,2}} \subseteq S_{M_{2,2}}$, (ii) there exists some $\widetilde{t}_1 \in \widetilde{S}_{M_{1,1}}$ and $\widetilde{t}_2 \in \bar{S}_{M_{1,2}}$ that satisfies $\lim_{t_1 \to \widetilde{t}_1, t_2 \to \widetilde{t}_2} M_2(t_1, t_2) / M_1(t_1, t_2) = 0$ (or ∞), where $\widetilde{t}_1$ and $\widetilde{t}_2$ are independent of $M_{2,1}$ and $M_{2,2}$. Then the class $\mathscr{H}'$ of all finite mixtures of $\mathscr{F}$ is identifiable.*

Proof. See Appendix. □

Using Corollary 1, we obtain a second corollary.

Corollary 2. *The class of all finite mixtures of a bivariate probit model is not identifiable.*

Proof. See Appendix. □

Next, we consider a probit model with a linear single equation. Let y_1 represent a binary variable that is assumed to be generated by the process $y_1 = 1$ if $y_1^* = \mathbf{x}_1'\boldsymbol{\beta}_1 + \varepsilon_1 \geq 0$ and $y_1 = 0$ otherwise. Moreover, y_2 is an observed continuous outcome that has $y_2 = \mathbf{x}_2'\boldsymbol{\beta}_2 + \varepsilon_2$, where y_1^* is a latent variable, ε_1 and ε_2 are unobserved heterogeneity, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ denote vectors of parameters, and $\mathbf{x}_1$ and $\mathbf{x}_2$ are vectors of covariates. The unobserved heterogeneity $(\varepsilon_1, \varepsilon_2)$ follows bivariate normal distribution with mean zero and covariance matrix $\left(1, \rho\sigma_2, \sigma_2^2\right)$.

The PDF of the probit model with a linear single equation takes the form

$$f(y_1, y_2) = \left[1 - \Phi(y_2)\right]^{1-y_1} \left[\Phi(y_2)\right]^{y_1} \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left[-\frac{1}{2\sigma_2^2}\left(y_2 - \mathbf{x}_2'\boldsymbol{\beta}_2\right)^2\right], \tag{3.1}$$

where $\Phi(y_2) \equiv \Phi\left[\left(\mathbf{x}_1'\boldsymbol{\beta}_1 + \rho\sigma_2^{-1}\left(y_2 - \mathbf{x}_2'\boldsymbol{\beta}_2\right)\right) / \sqrt{1-\rho^2}\right]$ and $\Phi(\cdot)$ is a CDF of a standard normal distribution.

Using Corollary 1, we obtain the following:

Corollary 3. *The class of all finite mixtures of a probit model with a linear single equation is identifiable.*

Proof. The MGF $M(t_1, t_2)$ of a probit model with a linear single equation is obtained by

$$\begin{aligned} M(t_1, t_2) = &\exp\left(t_2\mathbf{x}_2'\boldsymbol{\beta}_2 + \frac{\sigma_2^2}{2}t_2^2\right) \int_{-\infty}^{\infty} \left[1 - \Phi(y_2) + \Phi(y_2)e^{t_1}\right] \\ &\times \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left[-\frac{1}{2\sigma_2^2}\left\{y_2 - \left(\mathbf{x}_2'\boldsymbol{\beta}_2 + \sigma_2^2 t_2\right)\right\}^2\right] \mathrm{d}\, y_2. \end{aligned} \tag{3.2}$$

Since (3.2) has no explicit solution, we approximate this MGF using a numerical integration method, Gauss-Hermite quadrature. When $u_{2q} = \left\{y_2 - \left(\mathbf{x}_2'\boldsymbol{\beta}_2 + \sigma_2^2 t_2\right)\right\} / \left(\sqrt{2}\sigma_2\right)$, the transformation of the Jacobian is $\mathrm{d}\, y_2 = \sqrt{2}\sigma_2 \mathrm{d}\, u_2$. The MGF takes the following form:

$$M(t_1, t_2) = \exp\left(t_2\mathbf{x}_2'\boldsymbol{\beta}_2 + \frac{\sigma_2^2}{2}t_2^2\right) \sum_{q=1}^{Q} \frac{1}{\pi}\omega_{2q}e^{t_1}\left[\frac{\left\{1 - \Phi(u_{2q})\right\}}{e^{t_1}} + \Phi(u_{2q})\right], \tag{3.3}$$

where ω_{2q} is a weight, u_{2q} is the qth evaluation point over $[-\infty, \infty]$, Q is the number of weights, and $\Phi(u_{2q}) \equiv \Phi\left[\left(\mathbf{x}_1'\boldsymbol{\beta}_1 + \rho\sqrt{2}u_{2q} + \rho\sigma_2 t_2\right) / \sqrt{1-\rho^2}\right]$. When $t_1 \to \infty$, the square bracket of (3.3) converges to $\Phi(u_{2q})$. Moreover, when $0 < \rho < 1$, $\lim_{t_2 \to \infty} \Phi(\cdot) = 1$; when $-1 < \rho < 0$, $\lim_{t_2 \to -\infty} \Phi(\cdot) = 1$.

Let $F = F\left(y_1, y_2; \mathbf{x}_1'\boldsymbol{\beta}_1, \mathbf{x}_2'\boldsymbol{\beta}_2, \rho, \sigma_2^2\right)$ denote the CDF of a probit model with a single linear equation. Without loss of generality, order the family lexicographically by:

$$F_1 = F\left(y_1, y_2; \mathbf{x}_1'\boldsymbol{\beta}_{1,1}, \mathbf{x}_2'\boldsymbol{\beta}_{1,2}, \rho_1, \sigma_{1,2}^2\right) \prec F\left(y_1, y_2; \mathbf{x}_1'\boldsymbol{\beta}_{2,1}, \mathbf{x}_2'\boldsymbol{\beta}_{2,2}, \rho_2, \sigma_{2,2}^2\right) = F_2,$$

when (i) $\sigma_{1,2} > \sigma_{2,2}$, (ii) $\sigma_{1,2} = \sigma_{2,2}$ and $\mathbf{x}_2'\boldsymbol{\beta}_{1,2} > \mathbf{x}_2'\boldsymbol{\beta}_{2,2}$, (iii) $\sigma_{1,2} = \sigma_{2,2}$, $\mathbf{x}_2'\boldsymbol{\beta}_{1,2} = \mathbf{x}_2'\boldsymbol{\beta}_{2,2}$, and $\mathbf{x}_1'\boldsymbol{\beta}_{1,1} > \mathbf{x}_1'\boldsymbol{\beta}_{2,1}$, or (iv) $\sigma_{1,2} = \sigma_{2,2}$, $\mathbf{x}_2'\boldsymbol{\beta}_{1,2} = \mathbf{x}_2'\boldsymbol{\beta}_{2,2}$, $\mathbf{x}_1'\boldsymbol{\beta}_{1,1} = \mathbf{x}_1'\boldsymbol{\beta}_{2,1}$, and $\rho_1 > \rho_2$. Then,

$$\lim_{t_1 \to \infty} \frac{M_2(t_1, t_2)}{M_1(t_1, t_2)} = \frac{\exp\left(t_2 \mathbf{x}_2'\boldsymbol{\beta}_{2,2} + \frac{\sigma_{2,2}^2}{2} t_2^2\right)}{\exp\left(t_2 \mathbf{x}_2'\boldsymbol{\beta}_{1,2} + \frac{\sigma_{1,2}^2}{2} t_2^2\right)}.$$

Thus, (i) when $\rho_1 > 0$, $\lim_{t_1 \to \infty, t_2 \to \infty} M_2(t_1, t_2)/M_1(t_1, t_2) = 0$ since $\lim_{t_2 \to \infty} \Phi(\cdot) = 1$ if $\rho_1 > 0$. (ii) when $\rho_1 \leq 0$, $\lim_{t_1 \to \infty, t_2 \to -\infty} M_2(t_1, t_2)/M_1(t_1, t_2) = 0$ since $\lim_{t_2 \to -\infty} \Phi(\cdot) = 1$ if $\rho_1 \leq 0$. Therefore, from Corollary 1, the class of all finite mixtures of a probit model with a linear single equation is identifiable. □

Although the results of previous work require $T \geq 3$ Bernoulli trials to identify at most two components in finite binomial mixtures, our result identifies all finite mixtures if another continuous endogenous variable exists. An interesting feature of this corollary is that a finite mixture probit model with one linear equation is identifiable *even without a correlation between binary and continuous endogenous variables.* That is, if another endogenous continuous variable exists in probit estimation, we can estimate a finite mixture cross-sectional probit model.

3.3 Monte Carlo Experiments

This section presents results of Monte Carlo experiments of a finite mixture probit model with a linear single estimation. The number of simulations in all experiments is set to 100, and the sample size N is 1,000 and 2,000 observations per Monte Carlo iteration.

We generate two unobserved heterogeneity terms $\varepsilon_{1,j}$ and $\varepsilon_{2,j}$, $j = 1, 2$, normally distributed as $\mathcal{N}\left((0,0), \left(1, \rho_j \sigma_{2,j}, \sigma_{2,j}^2\right)\right)$. The probit component is obtained by $y_1^* = \mathbf{x}_1'\boldsymbol{\beta}_{1,j} + \varepsilon_{1,j}$, and a linear component is given by $y_2 = \mathbf{x}_2'\boldsymbol{\beta}_{2,j} + \varepsilon_{2,j}$, where $\mathbf{x}_1 = [1, \mathbf{x}_{12}]'$, $\mathbf{x}_{12} \sim \mathcal{N}(0, 1)$, and $\mathbf{x}_2 = [1]'$. The true values of the parameters of the first component are $\boldsymbol{\beta}_{1,1} = [-0.3, 0.3]'$, $\boldsymbol{\beta}_{2,1} = 1$, and $\sigma_{2,1} = 0.5$ or 1. True values of the second are $\boldsymbol{\beta}_{1,2} = [0.3, -0.3]'$, $\boldsymbol{\beta}_{2,2} = 1$, and $\sigma_{2,2} = 2$. That is, two components are the same except for the probit part and variance parameter $\sigma_{2,j}$. This is a severe condition for estimating finite mixture models. The correlation parameter ρ_j varies from -0.9 to 0.9. The mixing probability of the first component p_1 is 0.3.

Tables 3.1 and 3.2 present Monte Carlo results of the finite mixture probit model with a single linear estimation for $N =$ 1,000 and 2,000. Results for $(\boldsymbol{\beta}_{1,j}, \boldsymbol{\beta}_{2,j})$ are given in Tables 3.1 and 3.2 and show that the parameter estimates are unbiased for each estimation and for the value of the correlation parameter. Although the bias does not always decrease, root mean

Table 3.1: Monte Carlo Results ($\sigma_{2,1} = 0.5$ and $\sigma_{2,2} = 2$)

		$\rho = 0.9$		$\rho = 0$		$\rho = -0.9$	
	Truth	$N = 1{,}000$	$N = 2{,}000$	$N = 1{,}000$	$N = 2{,}000$	$N = 1{,}000$	$N = 2{,}000$
$\beta_{11,1}$	-0.3	0.0016	-0.0005	-0.0199	0.0029	-0.0092	-0.0001
		(0.1329)	(0.0931)	(0.1345)	(0.1015)	(0.1342)	(0.0856)
$\beta_{12,1}$	0.3	0.0047	0.0097	0.0238	0.0125	-0.0022	0.0038
		(0.0814)	(0.0550)	(0.1554)	(0.0976)	(0.0825)	(0.0591)
$\beta_{21,1}$	0	-0.0009	-0.0023	0.0017	0.0001	0.0050	0.0007
		(0.0462)	(0.0319)	(0.0491)	(0.0336)	(0.0472)	(0.0336)
ρ_1		-0.0034	-0.0031	-0.0085	-0.0019	0.0056	0.0043
		(0.0445)	(0.0312)	(0.1503)	(0.1009)	(0.0442)	(0.0340)
$\sigma_{2,1}$	0.5	-0.0077	-0.0088	-0.0063	-0.006	-0.0016	-0.0066
		(0.0438)	(0.0257)	(0.0485)	(0.0312)	(0.0456)	(0.0318)
$\beta_{11,2}$	0.3	0.0007	-0.0034	0.0016	-0.0002	0.0006	0.0023
		(0.0518)	(0.0423)	(0.0721)	(0.0464)	(0.0599)	(0.0405)
$\beta_{12,2}$	-0.3	-0.0039	-0.0072	-0.0006	-0.0045	-0.0005	-0.0037
		(0.0409)	(0.0299)	(0.0780)	(0.0504)	(0.0423)	(0.0319)
$\beta_{21,2}$	0	0.0068	-0.0014	0.0067	-0.0022	0.0057	-0.0022
		(0.0756)	(0.0602)	(0.0760)	(0.0608)	(0.0748)	(0.0610)
ρ_2		0.0000	0.0020	0.0028	-0.005	-0.0049	-0.003
		(0.0164)	(0.0109)	(0.0551)	(0.0382)	(0.0138)	(0.0100)
$\sigma_{2,2}$	2	0.0032	-0.0014	0.0032	0.0000	0.0089	-0.0001
		(0.0637)	(0.0434)	(0.0636)	(0.0480)	(0.0596)	(0.0429)
p_1	0.3	-0.0008	-0.0005	-0.0008	0.0006	0.0041	0.0007
		(0.0297)	(0.0195)	(0.0344)	(0.0238)	(0.0292)	(0.0199)

Notes: The mean bias reports appear without parentheses. Root mean squared errors are in parentheses.

squared errors (RMSE) decrease when the number of samples increases for each experiment. The bias and RMSE of ρ_j do not depend on the value of ρ_j, and the performance of these parameters is good. An interesting feature of these results is that the finite mixture probit model is identifiable if we find an endogenous (correlated or uncorrelated) continuous variable other than a binary endogenous variable.

3.4 Conclusion

Teicher (1960, 1963) and Blischke (1964) indicated that a finite mixture cross-sectional probit model is not identifiable. Nonetheless, this chapter demonstrated the identifiability of a cross-sectional probit model with one linear estimation based on Teicher's (1963) result.

Table 3.2: Monte Carlo Results ($\sigma_{2,1} = 1$ and $\sigma_{2,2} = 2$)

		$\rho = 0.9$		$\rho = 0$		$\rho = -0.9$	
	Truth	$N = 1{,}000$	$N = 2{,}000$	$N = 1{,}000$	$N = 2{,}000$	$N = 1{,}000$	$N = 2{,}000$
$\beta_{11,1}$	-0.3	-0.0632	-0.0158	-0.1416	-0.036	-0.0139	-0.0138
		(0.3132)	(0.1640)	(0.8206)	(0.2580)	(0.2255)	(0.1838)
$\beta_{12,1}$	0.3	0.0505	0.0363	0.2499	0.0560	0.0056	0.0094
		(0.2254)	(0.1128)	(1.5717)	(0.2839)	(0.1662)	(0.1351)
$\beta_{21,1}$	0	-0.0123	0.0014	-0.0079	0.0070	-0.0003	0.0022
		(0.1209)	(0.0953)	(0.1237)	(0.0924)	(0.1262)	(0.0965)
ρ_1		0.0064	0.0009	-0.0247	-0.0044	0.0017	0.0000
		(0.0518)	(0.0387)	(0.3119)	(0.1483)	(0.0611)	(0.0353)
$\sigma_{2,1}$	1	-0.0153	-0.0246	-0.0205	-0.0046	0.0040	-0.0065
		(0.1852)	(0.1180)	(0.2394)	(0.1127)	(0.1788)	(0.1218)
$\beta_{11,2}$	0.3	0.0055	-0.0053	0.0052	0.0074	0.0119	0.0058
		(0.0817)	(0.0643)	(0.1266)	(0.0775)	(0.0951)	(0.0595)
$\beta_{12,2}$	-0.3	0.0004	-0.008	0.0152	-0.0118	-0.0014	-0.005
		(0.0744)	(0.0384)	(0.1801)	(0.0801)	(0.0621)	(0.0454)
$\beta_{21,2}$	0	0.0163	-0.0026	0.0166	-0.0049	0.0114	-0.0006
		(0.0904)	(0.0702)	(0.0949)	(0.0705)	(0.0957)	(0.0710)
ρ_2		0.0002	0.0019	0.0032	-0.0031	-0.0074	-0.0037
		(0.0217)	(0.0142)	(0.0780)	(0.0480)	(0.0203)	(0.0124)
$\sigma_{2,2}$	2	-0.0165	-0.0059	-0.019	0.0045	0.0141	0.0049
		(0.1358)	(0.0555)	(0.1623)	(0.0663)	(0.0870)	(0.0551)
p_1	0.3	-0.0034	-0.0048	-0.0001	0.0071	0.0196	0.0093
		(0.0784)	(0.0602)	(0.1143)	(0.0698)	(0.0880)	(0.0630)

Notes: The mean bias reports appear without parentheses. The root mean squared errors are in parentheses.

The identifiability comes from the linear estimation. Monte Carlo experiments supported our demonstration and showed good performance. These results suggest that the finite mixture probit model is identifiable if we find a continuous correlated or uncorrelated endogenous variable that does not follow a single distribution. That is, there is little possibility of estimating the finite mixture cross-sectional probit model.

Appendix

A. Proof of Corollary 1

Proof. Suppose there are two finite sets of elements of $\mathscr{F}$, say $\mathscr{F}_1 = \{F_j, 1 \leq j \leq J\}$ and $\mathscr{F}_2 = \left\{F_{\widehat{\jmath}}, 1 \leq \widehat{\jmath} \leq \widehat{J}\right\}$ such that

$$\sum_{j=1}^{J} p_j F_j\left(y_1, y_2\right) \equiv_{y_1,y_2} \sum_{\widehat{\jmath}=1}^{\widehat{J}} p_{\widehat{\jmath}} F_{\widehat{\jmath}}(y_1, y_2), \quad 0 < p_j, p_{\widehat{\jmath}} \leq 1, \quad \sum_{j=1}^{J} p_j = \sum_{\widehat{\jmath}=1}^{\widehat{J}} p_{\widehat{\jmath}} = 1. \tag{A3.1}$$

Without loss of generality, index the CDF's so that $F_j \prec F_{\widehat{\jmath}}$, $\widehat{F}_j \prec \widehat{F}_{\widehat{\jmath}}$ for $j < \widehat{\jmath}$. If $F_1 \neq \widehat{F}_1$, suppose also without loss of generality that $F_1 \prec \widehat{F}_1$. Then, $F_1 \prec \widehat{F}_{\widehat{\jmath}}$, $1 \leq \widehat{\jmath} \leq \widehat{J}$ and from the transformed version of (A3.1), it follows that for $t_1, t_2 \in S_{M_{1,.}}$ $[t_1, t_2 : M_{1,.}\left(t_1, t_2\right) \neq 0]$,

$$p_1 + \sum_{j=2}^{J} p_j \left[\frac{M_j\left(t_1, t_2\right)}{M_1\left(t_1, t_2\right)}\right] \equiv_t \sum_{\widehat{\jmath}=1}^{\widehat{J}} \widehat{p}_{\widehat{\jmath}} \left[\frac{\widehat{M_{\widehat{\jmath}}}(t_1, t_2)}{M_1\left(t_1, t_2\right)}\right]. \tag{A3.2}$$

Letting $t_1 \to \widetilde{t}_1$ and $t_2 \to \widetilde{t}_2$ through values in T_1 and T_2 (this is possible), $p_1 = 0$ contradicting the supposition of (A3.1) that $p_1 > 0$. Then, $F_1 = \widehat{F}_1$ and for any $t_1 \in T_1, t_2 \in T_2$,

$$\left(p_1 - \widehat{p}_1\right) + \sum_{j=2}^{J} p_j \left[\frac{M_j\left(t_1, t_2\right)}{M_1\left(t_1, t_2\right)}\right] \equiv_t \sum_{\widehat{\jmath}=2}^{\widehat{J}} \widehat{p}_{\widehat{\jmath}} \left[\frac{\widehat{M_{\widehat{\jmath}}}(t_1, t_2)}{M_1\left(t_1, t_2\right)}\right]. \tag{A3.3}$$

Again letting $t_1 \to \widetilde{t}_1$ and $t_2 \to \widetilde{t}_2$ through values in T_1 and T_2, $p_1 = \widehat{p}_1$ whence

$$\sum_{j=2}^{J} p_j F_j\left(y\right) \equiv_y \sum_{\widehat{\jmath}=2}^{\widehat{J}} \widehat{p}_{\widehat{\jmath}} \widehat{F}_{\widehat{\jmath}}(y). \tag{A3.4}$$

Repeating the prior argument a finite number of times, we conclude that $F_j = \widehat{F}_j$ and $p_j = \widehat{p}_j$ for $j = 1, 2, \ldots, \min\left(j, \widehat{\jmath}\right)$. Further, if $j \neq \widehat{\jmath}$ say $j > \widehat{\jmath}$, then

$$\sum_{j=\widehat{J}+1}^{J} p_j F_j\left(y\right) = 0 \tag{A3.5}$$

implying that $p_j = 0$, $\widehat{J} + 1 \leq j \leq J$ in contradiction to (A3.1). Thus $J = \widehat{J}$, $p_j = \widehat{p}_j$ and $F_j = \widehat{F}_j$, $1 \leq j \leq J$, implying $\mathscr{F}_1 = \mathscr{F}_2$ and identifiability of $\mathscr{H}'$. □

B. Proof of Corollary 2

Proof. The bivariate probit model summarizes as follows: Let y_1 and y_2 represent binary variables assumed to be generated by the process $y_1 = 1$ if $y_1^* = \mathbf{x}_1'\boldsymbol{\beta}_1 + \varepsilon_1 \geq 0$ and $y_1 = 0$ otherwise; $y_2 = 1$ if $y_2^* = \mathbf{x}_2'\boldsymbol{\beta}_2 + \varepsilon_2 \geq 0$ and $y_2 = 0$ otherwise, where y_1^* and y_2^* are latent variables; ε_1 and ε_2 are unobserved heterogeneity; $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ denote vectors of parameters; and

$\mathbf{x}_1$ and $\mathbf{x}_2$ are vectors of covariates. The unobserved heterogeneity $(\varepsilon_1, \varepsilon_2)$ follows bivariate normal distribution with mean zero and covariance matrix $(1, \rho, 1)$.

The log-likelihood function of the bivariate probit model takes the following form:

$$\begin{aligned}
\ln f(y_1, y_2) =& (1-y_1) \times (1-y_2) \times \ln \left[\int_{-\infty}^{-\mathbf{x}_1'\boldsymbol{\beta}_1} \int_{-\infty}^{-\mathbf{x}_2'\boldsymbol{\beta}_2} \phi(\varepsilon_1, \varepsilon_2) \,\mathrm{d}\,\varepsilon_1 \mathrm{d}\,\varepsilon_2 \right] \\
&+ y_1 \times (1-y_2) \times \ln \left[\int_{-\mathbf{x}_1'\boldsymbol{\beta}_1}^{\infty} \int_{-\infty}^{-\mathbf{x}_2'\boldsymbol{\beta}_2} \phi(\varepsilon_1, \varepsilon_2) \,\mathrm{d}\,\varepsilon_1 \mathrm{d}\,\varepsilon_2 \right] \\
&+ (1-y_1) \times y_2 \times \ln \left[\int_{-\infty}^{-\mathbf{x}_1'\boldsymbol{\beta}_1} \int_{-\mathbf{x}_2'\boldsymbol{\beta}_2}^{\infty} \phi(\varepsilon_1, \varepsilon_2) \,\mathrm{d}\,\varepsilon_1 \mathrm{d}\,\varepsilon_2 \right] \\
&+ y_1 \times y_2 \times \ln \left[\int_{-\infty}^{-\mathbf{x}_1'\boldsymbol{\beta}_1} \int_{-\infty}^{-\mathbf{x}_2'\boldsymbol{\beta}_2} \phi(\varepsilon_1, \varepsilon_2) \,\mathrm{d}\,\varepsilon_1 \mathrm{d}\,\varepsilon_2 \right] \\
\equiv& \ln p_{00}^{(1-y_1)(1-y_2)} + \ln p_{10}^{y_1(1-y_2)} + \ln p_{01}^{(1-y_1)y_2} + \ln p_{11}^{y_1 y_2},
\end{aligned} \tag{B3.1}$$

where $\phi(\cdot)$ is a PDF of a bivariate standard normal distribution. Thus, the MGF $M(t_1, t_2)$ is obtained by

$$\begin{aligned}
M(t_1, t_2) =& \sum_{y_1=0}^{1} \sum_{y_2=0}^{1} e^{t_1 y_1 + t_2 y_2} p_{00}^{(1-y_1)(1-y_2)} p_{10}^{y_1(1-y_2)} p_{01}^{(1-y_1)y_2} p_{11}^{y_1 y_2} \\
=& p_{00} + p_{10} e^{t_1} + p_{01} e^{t_2} + p_{11} e^{t_1+t_2}.
\end{aligned} \tag{B3.2}$$

Note that

$$\begin{aligned}
\frac{M_2(t_1, t_2)}{M_1(t_1, t_2)} =& \frac{p_{2,00} + p_{2,10} e^{t_1} + p_{2,01} e^{t_2} + p_{2,11} e^{t_1+t_2}}{p_{1,00} + p_{1,10} e^{t_1} + p_{1,01} e^{t_2} + p_{1,11} e^{t_1+t_2}} \\
=& \frac{p_{2,00}/e^{t_1+t_2} + p_{2,10}/e^{t_2} + p_{2,01}/e^{t_1} + p_{2,11}}{p_{1,00}/e^{t_1+t_2} + p_{1,10}/e^{t_2} + p_{1,01}/e^{t_1} + p_{1,11}}.
\end{aligned}$$

Then,

$$\lim_{t_1 \to \infty, t_2 \to \infty} \frac{M_2(t_1, t_2)}{M_1(t_1, t_2)} = \frac{p_{2,11}}{p_{1,11}} > 0. \tag{B3.3}$$

Therefore, from Corollary 1, the class of all finite mixtures of the bivariate probit model is not identifiable. □

Part II

Endogeneity in Health Econometrics

Chapter 4

Endogenous Variables in Health Econometrics

4.1 Introduction

In microeconometrics, especially in health econometrics, it is widely known that endogenous regressors cause the possibility of inconsistent parameter estimation. Here endogeneity is defined as a regressor that is correlated with the error term. For example, when we analyze the influence of a physician's advice to reduce alcohol consumption, the error term contains all factors other than the advice concerning alcohol, such as whether the patient has private medical insurance (Kenkel and Terza, 2001). If privately insured patients are more likely to receive lifestyle advice, the error term and the advice are correlated, and endogeneity occurs. Endogeneity is a problem because OLS estimates of all regression parameters are generally inconsistent if any regressor is endogenous (unless the exogenous regressor is uncorrelated with the endogenous regressor).

Endogeneity does not arise in health econometrics if data are randomly assigned or regressors are not the results of incentives. However, these conditions are seldom fulfilled in social sciences research; endogeneity is inevitable, and a method to treat it correctly is required. This chapter analyzes the problem of endogeneity and explains the estimation of regression models with an endogenous regressor, focusing on health econometrics per Wooldridge (2002), Cameron and Trivedi (1998, 2005), Winkelmann (2004), and Winkelmann and Boes (2006).[1] Section 2 introduces examples of famous microdata, the British Health and Lifestyle Survey (HALS), the British Household Panel Survey (BHPS), and the Medical Expenditures Panel

[1] See also Davidson and Mackinnon (1993, 2003) and Greene (2007).

Table 4.1: Consistency in Linear and Nonlinear Regression with Endogenous Variables

	dependent variable	endogenous variable	Two-stage
(i)	continuous	continuous	consistent
(ii)	continuous	discrete, censored, or truncated	consistent
(iii)	discrete, censored, or truncated	continuous	consistent
(iv)	discrete, censored, or truncated	discrete, censored, or truncated	inconsistent

Survey (MEPS), that are often used in health economics. The examples show that there are many discrete, censored, or truncated variables and few continuous variable and that the data contain potential problems of endogeneity.

Section 3 explains the problem of endogeneity using a simple linear regression model, explains the instrumental variable method that obtains the consistent estimator even if there are endogenous variables, and describes the 2SLS method used in applied fields. Although the discussion of instrumental variable estimators is based on continuous endogenous regressors, we extend it to a binary endogenous variable, referred as treatment effects. The main result of treatment effects is that the instrumental variable or two-stage estimator has a consistent estimator when an endogenous regressor is binary.

In nonlinear (discrete, censored, or truncated) regression used in health econometrics, such as binary variable and count data models, correlation between a regressor and error term (endogeneity) leads to inconsistently estimated regression parameters. Even so, the two-stage method used in many linear models sometimes works poorly in nonlinear regression with endogeneity. More concretely, if the two stage method is applied in estimating nonlinear models, such as probit and count data models, with endogenous discrete, censored, or truncated regressors, the estimated parameters have no consistency.

Table 4.1 explains this chapter's discussion. Rows 2 and 3 in Table 4.1 present the discussion in Section 2, and the two-stage method has consistency in linear models regardless irrespective of any endogenous regressors. Rows 4 and 5 discuss Section 3. In nonlinear models, the two-stage method is consistent when endogenous variables are continuous, but the FIML has consistency when endogenous variables are discrete, censored, or truncated.

Using examples of probit and count data models, Section 4 explains the two-stage method used in nonlinear models with endogenous continuous regressors. Moreover, we demonstrate that, in nonlinear regression with endogenous discrete, censored, or truncated regressors, the two-stage method is inadequate and the FIML is consistent.

Section 5 provides simple Monte Carlo simulations of the four cases in Table 4.1 and analyzes the consistency of proposed models. We show the consistency of linear models with an endogenous continuous, discrete, censored, or truncated regressor and the inconsistency of probit models with an endogenous binary variable. Section 5 shows the limits of nonlinear econometric regression with endogenous variables and proposes more desirable analysis.

4.2 Examples of Microdata in Health Economics

4.2.1 The Health and Lifestyle Survey

The British Health and Lifestyle Survey (HALS) is a health interview survey and is designed as a representative survey of adults (age 18 and over) in Great Britain. The HALS was designed originally as a cross-section survey in 1985 and was collected in three stages: a one-hour face-to-face interview, that includes experience and attitudes towards to health and lifestyle along with general socioeconomic information; a nurse visit to collect physiological measures and indicators of cognitive function, such as memory and reasoning; a self-completion postal questionnaire to measure psychiatric health and personality. The HALS is an example of a clustered random sample; addresses were randomly selected from electoral registers using a three-stage design. Then, individuals were randomly selected from households. This selection procedure obtained a target of 12,672 interviews.

Table 4.2 shows the example of the HALS data. The main variables from the HALS sample are indicators of health (sah) and lifestyle (nsmoker and alqprud) indicators; socioeconomic indicatros (scl2, sc3, sc45, lhqdg, lhqO, lhqnone, lhqoth, full, part, unemp, sick, retd, and keephse); geographical and area indicators (omitted in Table 4.2); marital status (married, widow, divorce, seprd, and single); ethnicity (ethwheur); demographic characteristics (male, height, age, housown, and hou) and parental smoking and drinking behaviors (smother, mothsmo, fathsmo, bothsmo, alpa, and alma). The data also show that there are many binary and discrete variables and *few continuous variable.*

4.2.2 The British Household Panel Survey

While the HALS has only two waves of panel data, the British Household Panel Survey (BHPS) is a longitudinal survey from 1991 to the present in Great Britain. The BHPS is an annual survey of each adult (age 16 and over) member of a nationally representative sample of more than 5,000 households, with a total of approximately 10,000 individual interviews. The initial selection of households is a two-stage clustered systematic sampling and the same individuals are re-interviewed in successive waves. If they move overseas or geographic areas, households are also re-interviewed along with all adult members of their new households.

Table 4.2: An Example of the British Health and Lifestyle Survey

Variable	Definition	Type
alpa	Father, non to heavy drinker (0-4)	Categorical
alma	Mother, non to heavy drinker (0-4)	Categorical
male	1 = male	Binary
ethwheur	White European	Binary
smother	1 = anyone else in house smoked	Binary
housown	1 = own or rent house	Binary
hou	Number of other people in the house	Count
height	Height in inches	Continuous
widow	Widow	Binary
divorce	Divorced	Binary
seprd	Separated	Binary
married	Married	Binary
single	Single	Binary
full	Full time worker or student	Binary
part	Part time worker	Binary
sick	Absent from work due to sickness	Binary
retd	Retired	Binary
keephse	Housekeeper	Binary
unemp	Unemployed	Binary
lhqnone	No qualification	Binary
lhqO	O level/Certificate of Secondary Education (CSE)	Binary
lhqdg	University degree	Binary
lhqoth	Other vocational/professional qualifications	Binary
age	Age in years	Continuous
sah	Self-assessed health is excellent or good	Binary
nsmoker	= 1 if does not smoke, 0 if current smoker	Binary
breakfast	Does a healthy breakfast	Binary
sleepgd	Sleeps between 7 and 9 hours	Binary
alqprud	Consume alcohol prudently	Binary
nobese	Not obese	Binary
exercise	Did physical exercise in the last fortnight	Binary
wkshft1	Shift worker	Binary
suburb	Lives in the suburbs of the city	Binary
rural	Lives in the countryside	Binary
sc12	Professional/student or managerial/intermediate	Binary
sc45	Partly skilled, unskilled, unclass. or never occupied	Binary
sc3	Skilled or armed service	Binary
scgr	Social class	Categorical
lhqhndA	Higher vocational qualifications or A level or equivalent	Binary
mothsmo	Only mother smoked	Binary
fathsmo	Only father smoked	Binary
bothsmo	Both parents smoked	Binary

Note: Jones *et al.*, 2007, Chapter 5.

The BHPS contains one measure of health outcomes of self-assessed health (sah) defined by a response to: ‘Please think back over the last 12 months about how your health has been. Compared to people of your own age, would you say that your health has on the

Table 4.3: An Example of British Household Panel Survey

Variable	Definition	Type
retired	Respondent states they are retired	Binary
hlltyes	Health limits daily activities	Binary
sah	Self-assessed health; very poor or poor, fair, good or very good, and excellent	Categorical
m2lnhinc	Individual-specific mean of log equivalized real household labor and non-labor income	Continuous
HseMort	House has outstanding mortgage	Binary
HseRent	House is rented	Binary
HseAuthAss	House is owned by housing authority/association	Binary
martcoup	Married or living as a couple	Binary
deghdeg	Highest educational attainment is degree or higher degree	Binary
hndalev	Highest educational attainment is HND or A level	Binary
ocse	Highest educational attainment is O level or CSE	Binary
everppenr	Has made contributions to private pension plan	Binary
everemppr	A member of an occupational pension plan	Binary
job	Respondent's spouse/partner has a job	Binary
age5054	Aged 50 to 54	Binary
age5559	Aged 55 to 59	Binary
age6064	Aged 60 to 64	Binary
age6569	Aged 65 to 69	Binary
hlthprb	Self-reported health problems	Binary

Note: Jones *et al.*, 2007, Chapter 7.

whole been excellent/good/fair/poor/very poor?' Therefore, the self-assessed health variable (sah) is regarded as a relative health perceived status based on the individual's 'norm' for their age group. This variable is also a simple subjective measure of health that supplies an ordinal raking of perceived health status and is used widely in previous empirical studies of the relationship between health and socioeconomic.

Moreover, there are different indicators of morbidity in the BHPS. The variable health limitation (hlltyes) measures self-reported functional limitations. Respondents determine their own concepts of health and their daily activities. In contrast, for the variable measuring specified health problems (hlthprb), respondents are presented with a prompt card. The list contains arms, legs, hands, etc.; sight; hearing; skin conditions/allergies; chest/breathing; heart/blood pressure; stomach/digestion; diabetes; anxiety/depression; alcohol/drug-related; epilepsy, migraine and other.

Health economic analysis often focuses on two main measures of socioeconomic status: income and education. Income (m2lnhinc) is measured as equivalent and retail price index

(RPI) defined annual household income. Education is measured by the highest educational qualification attained by the end of the sample period. Moreover, variables that reflect individuals' demographic characteristics and stage of life: age, ethnic group, marital status and family composition are included. Table 4.3 obtains the example of the BHPS and we also find that there are many binary and discrete variables and *few continuous variable* like the HALS.

In both the HALS and BHPS obtained in Table 4.2 and 4.3, we find a potential problem of endogeneity. For example, when we identify a causal effect of health on the retirement decision of older people, it is difficult to assume self-assessed measures of health status as an exogenous variable. First, self-assessed measures are based on subjective judgements and these judgements are different among individuals. Second, self-assessed health may not be independent of working behavior. Third, respondents may answer health problems to rationalize behavior since ill health represents a reason for retirement. Therefore, the role of health on retirement contains the potential problem of endogeneity. Moreover, since the variable of self-assessed health is also binary and the retirement decision takes the form of a discrete (or censored) variable, it is inevitable to estimate a discrete (or censored) variable with an endogenous binary covariate.

4.2.3 The Medical Expenditures Panel Survey

The Medical Expenditures Panel Survey (MEPS) is a panel survey in the United States since 1996. The MEPS covers families and individuals, employers, and information on the use of medical services (doctors, hospitals, pharmacies, etc.). Moreover, the data include health services assessed, frequency of contact, the cost of services used, and information on health insurance status. The MEPS consists of two basic survey components: a Households Component and Insurance Component. The Household Component collects data on families and individuals in selected areas across the United States drawn from a nationally representative subsample of households. The Insurance Component contains information on the health insurance plans offered to their employees, such as the number and type of private insurance plans offered, premiums, contributions, eligibility criteria, and the benefits associated with the plans.

In the MEPS, obtained in Table 4.4, health care expenditures are defined as the sum of direct payments for care providers during the year, which include out-of-pocket payments and payments by private insurance, Medicaid, Medicare, and other sources. Annual direct payments consist of the use of and associated expenditures for office and hospital-based care, home health care, dental services, prescribed medicines, vision aids, and other medical supplies

Table 4.4: An Example of Medical Expenditure Panel Survey

Variable	Definition	Type
age	Age	Continuous
famsze	Size of the family	Count
educyr	Years of education	Continuous
totexp	Total medical expenditure	Continuous
private	Private supplementary insurance	Binary
retire	Retired	Binary
female	Female	Binary
white	White	Binary
hisp	Hispanic	Binary
marry	Married	Binary
northe	Northeast area	Binary
mwest	Midwest area	Binary
south	South area (West is excluded)	Binary
phylim	Has functional limitation	Binary
actlim	Has activity limitation	Binary
msa	Metropolitan statistical area	Binary
income	Annual household income/1000	Continuous
injury	Condition is caused by an accident/injury	Binary
priolist	Has medical conditions that are on the priority list	Binary
totchr	Number of chronic problems	Count
omc	Other managed care	Binary
hmo	Private insurance is Health Maintenance Organization (HMO)	Binary
suppins	Has supplementary private insurance	Binary
hvgg	Health status is excellent, good or very good	Binary
hfp	Health status is fair or poor	Binary
hins	Excellent health indicator	Categorical
hdem	Demographic group indicator	Categorical

Note: Jones *et al.*, 2012, Chapter 3.

and equipment. However, payments for over-the-consumer drugs, alternative care services, and phone contacts with medical providers are excluded.

Moreover, the MEPS collects data on respondent age, race (American Indian, Alaska Native, Asian or Pacific Islander, black, white, or other), household income, household poverty status (income relative to poverty thresholds measured as poor, near poor, low income, middle income, or high income), region and place of residence (Northeast, Midwest, South, or West), and employment status (employed if age 16 or over, and had a job for pay, owned a business, or worked without pay in a family business). Health status is recorded by asking respondents to rate the health of each person in the family using the following categories: excellent, very good, good, fair, and poor. Health insurance details for individuals under age 65 is categorized as: private health insurance (individuals who had insurance that provides cover for hospital and physician care, other than Medicare, Medicaid, or other public cover); public cover only (individuals not covered by private insurance and we not covered by Medicare, or other public cover); and uninsured. The older individuals (age 65 and over) are classified under Medicare only; Medicare and private insurance; Medicare and other public insurance. In Table 4.4, we find that there are many binary and discrete variables and a few continuous variable.

Using the MEPS in Table 4.4, to analyze health care expenditures contains the potential problem of endogeneity. When we identify how the individual's insurance choice affects the health care expenditures, it is difficult to assume that an individual's insurance choice is exogenous since healthy people may not select any insurance and their expenditures may be quite low. The insurance choice has the potential problem of endogeneity. Moreover, health status in the MEPS data may be another candidate for endogeneity.

Large-scale microdata (or survey datasets), discussed in this section, have many variables and provide a rich source of information. As in Table 4.2, 4.3, and 4.4, many variables are binary or categorical data and there are a few continuous variables. When variables increase and information is rich, the risk of the potential problem of endogeneity is high. Therefore, it is important to treat an endogenous variable correctly in health economics.

4.3 Endogenous Regressors in Linear Health Econometrics

4.3.1 Inconsistency of OLS

We consider a simple regression model with a dependent variable y_i and one explanatory variable x_i, $i = 1, \dots, N$, where N is the number of observations.[2] For simplicity, we omit the intercept and subscript i. A linear conditional mean model specifies $\mathrm{E}[y \mid x] = x\beta$, where β is an estimated parameter. The OLS model specifies

$$y = x\beta + \varepsilon, \tag{4.1}$$

where ε is an error term. Regression of y on x yields an OLS estimate $\widehat{\beta}$ of β.

Standard regression results assume the explanatory variable x is uncorrelated with the error term ε in the model (4.1). This means that parameter β is the direct effect of x on y

[2] The following discussion is based on Cameron and Trivedi (2005).

and obtains the following path analysis diagram:

$$\begin{tikzcd} x \arrow[r] & y \\ \varepsilon \arrow[ur] & \end{tikzcd}$$

where there is no association between x and ε.

However, an explanatory variable and the error term may be associated. For example, when we analyze the influence of healthy activities x, (e.g., monthly exercise) on y (the inverse of health expenditure), the error term ε embodies all other factors that influence the inverse of health expenditure, such as distance to the fitness club. Moreover, we assume people have a high level of ε if they participate extensively in health maintenance activities. Since $y = x\beta + \varepsilon$, this increases the inverse of health expenditure. However, it may also lead to higher values of x since health maintenance activities were higher when the fitness club was nearby. Therefore, the following path diagram is more appropriate:

$$\begin{tikzcd} x \arrow[r] & y \\ \varepsilon \arrow[u] \arrow[ur] & \end{tikzcd}$$

where x and ε are associated.

When a correlation exists between x and ε, higher levels of x have two effects on y: a direct effect via $x\beta$ and an indirect effect via ε affecting x which in turn affects y. The regression estimates only the first effect β, but the OLS estimate combines the two effects and obtains $\widehat{\beta} > \beta$ in this example where both effects are positive. Now we write the OLS equation as $y = x\beta + \varepsilon\,(x)$ and take the total derivative:

$$\frac{\mathrm{d}\,y}{\mathrm{d}\,x} = \beta + \frac{\mathrm{d}\,\varepsilon}{\mathrm{d}\,x}.$$

The OLS estimator is biased and inconsistent for β, unless x and ε have no association.[3]

The phenomena by which changes in x are linked to changes in y and the error term ε is endogeneity and causes inconsistency in OLS estimates. Randomized experimental data are required to avoid endogeneity, but experiments are expensive and infeasible.

Definition of Instruments

If we find an appropriate instrument z which has the property that changes in z are linked to changes in x but do not lead to changes in y, it is possible to estimate β using observational

[3] The linear regression model with K explanatory variables leads to the same conclusion.

data. This leads to the following path diagram:

$$\begin{tikzcd} z \arrow[r] & x \arrow[r] & y \\ & \varepsilon \arrow[u] \arrow[ur] & \end{tikzcd}$$

which introduces a variable z that is associated with x but is independent of ε.

Note that z does not determine y directly. It determines y indirectly through x and, as a consequence, z is correlated with y. Correctly, there are two conditions under which z is an instrument or instrumental variable for explanatory variable x: z is uncorrelated with ε and is correlated with x. The first condition requires that, if z is an explanatory variable and y is regressed on x alone, z contains the error and is correlated with error term ε. The second condition requires there to be an association between the instrumental variable and the explanatory variable that is instrumented.

4.3.2 Instrumental Variables Estimation

Consider the regression model with K explanatory variables,

$$y = \mathbf{x}'\boldsymbol{\beta} + \varepsilon,$$

where $\mathbf{x}$ and $\boldsymbol{\beta}$ are $K \times 1$ vectors. Assume there is an $r \times 1$ vector of instruments $\mathbf{z}$, with $r \geq K$, which satisfies three conditions: (i) $\mathbf{z}$ is uncorrelated with the error term ε, (ii) $\mathbf{z}$ is correlated with the explanatory variable vector $\mathbf{x}$, and (iii) $\mathbf{z}$ is strongly, not weakly, correlated with the explanatory variable vector $\mathbf{x}$. The first two conditions are necessary for consistency. The third assures good finite-sample performance of the instrumental variables (IV) estimator of the second condition.

Moreover, the number of instruments must at least equal the number of independent endogenous variables, $r \geq K$. This is called the *order condition.* The model is just-identified when $r = K$ and overidentified when $r > K$. An instrument fails the first condition is an *invalid instrument.* An instrument that fails the second condition is an *irrelevant instrument*, and the model may be unidentified. The third condition fails when correlation between the instrument and the endogenous variable is very low. The model is *weakly identified*, and the instrument is a *weak instrument.*

When the model is just-identified ($r = K$), the IV takes the form:

$$\widehat{\boldsymbol{\beta}}_{\mathrm{IV}} = \left(\mathbf{Z}'\mathbf{X}\right)^{-1}\mathbf{Z}'\mathbf{y}, \tag{4.2}$$

where $\mathbf{y}$ is an $N \times 1$ vector, $\mathbf{X}$ and $\mathbf{Z}$ are $N \times K$ matrices with ith rows $\mathbf{x}_i'$ and $\mathbf{z}_i'$. Substituting

$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ into (4.2) yields

$$\widehat{\boldsymbol{\beta}}_{\text{IV}} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'[\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}] = \boldsymbol{\beta} + \left(N^{-1}\mathbf{Z}'\mathbf{X}\right)^{-1} N^{-1}\mathbf{Z}'\boldsymbol{\varepsilon}. \tag{4.3}$$

Therefore, the IV estimator is consistent if and only if

$$\text{plim}N^{-1}\mathbf{Z}'\boldsymbol{\varepsilon} = \mathbf{0} \text{ and } \text{plim}N^{-1}\mathbf{Z}'\mathbf{X} \neq \mathbf{0}. \tag{4.4}$$

These are essentially the same as conditions (i) and (ii): $\mathbf{z}$ is uncorrelated with $\boldsymbol{\varepsilon}$ and correlated with $\mathbf{x}$. To assure that the inverse of $N^{-1}\mathbf{Z}'\mathbf{X}$ exists, it is assumed $\mathbf{Z}'\mathbf{X}$ is of full rank K. This is a stronger assumption than the order condition $r = K$. With heteroskedastic errors, the IV estimator is asymptotically normal with mean $\boldsymbol{\beta}$ and the variance matrix is consistently estimated by

$$\widehat{\text{Var}}\left[\widehat{\boldsymbol{\beta}}_{\text{IV}}\right] = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\widehat{\boldsymbol{\Omega}}\mathbf{Z}(\mathbf{X}'\mathbf{Z})^{-1},$$

where $\widehat{\boldsymbol{\Omega}} = \text{Diag}\left[\widehat{\varepsilon_i^2}\right]$.[4]

4.3.3 Two-Stage Least Squares

The IV estimator in (4.2) requires that the number of instruments equals the number of explanatory variables. When $r > K$, some instrument variables are dropped, and the model becomes just-identified. However, discarding these instruments diminishes asymptotic efficiency. Then the following 2SLS estimator is applied:

$$\widehat{\boldsymbol{\beta}}_{\text{2SLS}} = \left[\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}\right]^{-1}\left[\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}\right].$$

The 2SLS estimator is the generalization of an IV estimator, and it is often calculated in two stages. First, $\mathbf{X}$ is regressed on instrumental variable $\mathbf{Z}$ to obtain a predicted value of $\mathbf{X}$. That is, $\widehat{\mathbf{X}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$. Second, $\mathbf{y}$ is regressed on $\widehat{\mathbf{X}}$, which obtains $\widehat{\boldsymbol{\beta}}_{\text{2SLS}}$. If the model is just-identified ($r = K$), first-stage regression is unnecessary, and $\mathbf{X}$ serves as its own instrument. If the model is overidentified, the 2SLS estimator equals the IV estimator in (4.2) and the instruments are $\widehat{\mathbf{X}}$.

Moreover, the 2SLS estimator is often expressed more compactly as

$$\widehat{\boldsymbol{\beta}}_{\text{2SLS}} = [\mathbf{X}'\mathbf{P_Z}\mathbf{X}]^{-1}[\mathbf{X}'\mathbf{P_Z}\mathbf{y}],$$

where $\mathbf{P_Z} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$ is an idempotent projection matrix that satisfies $\mathbf{P_Z} = \mathbf{P_Z}'$, $\mathbf{P_Z}\mathbf{P_Z}' = \mathbf{P_Z}$, and $\mathbf{P_Z}\mathbf{Z} = \mathbf{Z}$. The 2SLS estimator can be shown to be asymptotically normally distributed with estimated asymptotic variance

$$\widehat{\text{Var}}\left[\widehat{\boldsymbol{\beta}}_{\text{2SLS}}\right] = N[\mathbf{X}'\mathbf{P}'_{\mathbf{Z}}\mathbf{X}]^{-1}\left[\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\widehat{\mathbf{S}}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}\right][\mathbf{X}'\mathbf{P}'_{\mathbf{Z}}\mathbf{X}]^{-1},$$

[4] For more details, see also White (2001).

where in the usual case of heteroskedastic errors $\widehat{\mathbf{S}} = N^{-1}\sum_i \widehat{\varepsilon}_i^2 \mathbf{z}_i\mathbf{z}_i'$ and $\widehat{\varepsilon}_i = y_i - \mathbf{x}_i'\widehat{\boldsymbol{\beta}}_{2\mathrm{SLS}}$. A commonly used small-sample adjustment is to divide by $N-K$ rather than N in the formula for $\widehat{\mathbf{S}}$. In the special case that errors are homoskedastic, simplification occurs and $\widehat{\mathrm{Var}}\left[\widehat{\boldsymbol{\beta}}_{2\mathrm{SLS}}\right] = s^2\left[\mathbf{X}'\mathbf{P}_{\mathbf{Z}}'\mathbf{X}\right]^{-1}$.

Although the IV (or 2SLS) estimator is consistent, it leads to a loss of efficiency in practice. Intuitively, the IV (or 2SLS) estimator does not work well if instrument $\mathbf{z}$ has low correlation with the regressor $\mathbf{x}$. To simplify discussion, we consider the just-identified case, $K=1$, and no intercept. The IV estimator takes the form

$$\begin{aligned}
\mathrm{plim}\widehat{\beta}_{\mathrm{IV}} &= \beta + \left[\frac{\sum_{i=1}^{N} z_i\varepsilon_i}{\sqrt{\sum_{i=1}^{N} z_i^2}\sqrt{\sum_{i=1}^{N}\varepsilon_i^2}}\frac{\sqrt{\sum_{i=1}^{N}\varepsilon_i^2}}{\sqrt{\sum_{i=1}^{N} z_i^2}}\right] \Big/ \left[\frac{\sum_{i=1}^{N} z_i x_i}{\sqrt{\sum_{i=1}^{N} z_i^2}\sqrt{\sum_{i=1}^{N} x_i^2}}\frac{\sqrt{\sum_{i=1}^{N} x_i^2}}{\sqrt{\sum_{i=1}^{N} z_i^2}}\right] \\
&= \beta + \frac{\mathrm{Corr}\left(z,\varepsilon\right)}{\mathrm{Corr}\left(z,x\right)}\frac{\sigma_\varepsilon}{\sigma_x}.
\end{aligned} \tag{4.5}$$

Therefore, if z is weakly correlated with explanatory variable x and $\mathrm{Corr}\left(z,x\right)$ is relatively smaller than $\mathrm{Corr}\left(z,\varepsilon\right)$, the bias of the IV estimator $\widehat{\beta}_{\mathrm{IV}}$ is large.

4.3.4 A Binary Endogenous Regressor in Linear Models

Next, we consider linear estimation with a binary endogenous variable. For simplicity, consider the case of a single binary endogenous variable. A binary endogenous regressor in linear models takes this form:

$$y_1 = \mathbf{x}'\boldsymbol{\beta}_{\mathbf{1}} + \alpha_1 y_2 + \varepsilon_1, \tag{4.6}$$

$$y_2^* = \mathbf{z}'\boldsymbol{\beta}_{\mathbf{2}} + \varepsilon_2, \tag{4.7}$$

$$y_2 = \begin{cases} 1 & \text{if and only if } \; y_2^* > 0 \\ 0 & \text{if and only if } \; y_2^* \leq 0 \end{cases}, \tag{4.8}$$

where $\boldsymbol{\beta}_1 \sim K_1 \times 1$ and $\boldsymbol{\beta}_2 \sim K_2 \times 1$ are vectors of unknown parameters and y_2^* is a latent variable that depends on $\mathbf{z}$.

Linear models with a binary endogenous regressor are essentially the same as the well-known treatment effect model discussed in Rubin (1974), Heckman and Rob (1985), Holland (1986), Moffitt (1991), Mroz (1999), Wooldridge (2002, Chapter 18), and Lee (2005b). In treatment literature, y_1 is the outcome variable of interest, and y_2 denotes an indicator for treatment. For example, outcome y_1 is the inverese of health expenditure and indicator y_2 is treatment, such as participation in a health maintenance program. If $y_2 = 1$, the individual participated and received the treatment, and if $y_2 = 0$, the individual is part of the untreated control group. Moreover, let y_{11} be the outcome for someone who participates in the health maintenance program, and let y_{10} be the outcome of someone who does not. The causal effect of treatment, the average treatment effect (ATE), and the average treatment effect on the treated (ATT) are defined as $y_{11} - y_{10}$, ATE $= \mathrm{E}\left(y_{11} - y_{10}\right)$, and ATT $= \mathrm{E}\left(y_{11} - y_{10} \mid y_2 = 1\right)$, respectively.

The fundamental problem of treatment evaluation is to observe either y_{11} or y_{10} but never both. We can directly identify $\mathrm{E}\left(y_{11} \mid y_2 = 1\right)$ and $\mathrm{E}\left(y_{10} \mid y_2 = 0\right)$ from the data since $y_1 = y_{10} + y_2\left(y_{11} - y_{10}\right)$. However, this is not in itself sufficient for estimating ATE or ATT. Since $\mathrm{E}\left(y_{11} - y_{10} \mid y_2 = 1\right) = \mathrm{E}\left(y_{11} \mid y_2 = 1\right) - \mathrm{E}\left(y_{10} \mid y_2 = 1\right)$ by definition and the second term is unobserved, ATT is not estimated. If we assume the expected outcome without treatment is the same for the treated and control groups (independence between y_{10} and y_2), we can replace the unobserved term $\mathrm{E}\left(y_{10} \mid y_2 = 1\right)$ with the observed term $\mathrm{E}\left(y_{10} \mid y_2 = 0\right)$ and estimate ATT, but doing so is arbitrary. In ATE, the assumption is somewhat stronger and requires independence between not only y_{10} and y_2 but also y_{11} and y_2, because

$$\mathrm{ATE} = \mathrm{P}\left(y_2 = 1\right)\mathrm{E}\left(y_{11} - y_{10} \mid y_2 = 1\right) + \mathrm{P}\left(y_2 = 0\right)\mathrm{E}\left(y_{11} - y_{10} \mid y_2 = 0\right). \tag{4.9}$$

Using randomized experiment data is the ideal way to avoid this problem and to estimate ATE and ATT correctly. In randomized experiments, treatment and outcome are independent by definition. ATE $=$ ATT $= \mathrm{E}\left(y_1 \mid y_2 = 1\right) - \mathrm{E}\left(y_1 \mid y_2 = 0\right)$, and simple regression of y_1 on y_2 estimates ATE. However, the assumption of independence is implausible with observational data, and a weaker assumption is required. First, we assume that selection into treatment and outcome are independent, conditional on regressor $\mathbf{x}$, such that $\mathrm{E}\left(y_{10} \mid y_2 = 1, \mathbf{x}\right) = \mathrm{E}\left(y_{10} \mid y_2 = 0, \mathbf{x}\right)$ and $\mathrm{E}\left(y_{11} \mid y_2 = 1, \mathbf{x}\right) = \mathrm{E}\left(y_{11} \mid y_2 = 0, \mathbf{x}\right)$.

If we further assume that the joint distribution of ε_1 and ε_2 is a bivariate normal distribution with variance matrix $\Sigma = \left[\sigma_1^2, \rho\sigma_1, 1\right]$, we can immediately derive the expectation of the dependent variable conditional on treatment, since

$$\begin{aligned}\mathrm{E}\left(y_1 \mid y_2 = 1\right) &= \mathbf{x}'\boldsymbol{\beta}_1 + \alpha_1 + \mathrm{E}\left(\varepsilon_1 \mid y_2 = 1\right)\\ &= \mathbf{x}'\boldsymbol{\beta}_1 + \alpha_1 + \mathrm{E}\left(\varepsilon_1 \mid \varepsilon_2 > -\mathbf{z}'\boldsymbol{\beta}_2\right)\\ &= \mathbf{x}'\boldsymbol{\beta}_1 + \alpha_1 + \rho\sigma_1\frac{\phi\left(\mathbf{z}'\boldsymbol{\beta}_2\right)}{\Phi\left(\mathbf{z}'\boldsymbol{\beta}_2\right)}.\end{aligned} \tag{4.10}$$

For non-participants, the counterpart is

$$\mathrm{E}\left(y_1 \mid y_2 = 0\right) = \mathbf{x}'\boldsymbol{\beta}_1 - \rho\sigma_1\frac{\phi\left(\mathbf{z}'\boldsymbol{\beta}_2\right)}{1 - \Phi\left(\mathbf{z}'\boldsymbol{\beta}_2\right)}. \tag{4.11}$$

The difference between participants and non-participants is

$$\mathrm{E}\left(y_1 \mid y_2 = 1\right) - \mathrm{E}\left(y_1 \mid y_2 = 0\right) = \alpha_1 + \rho\sigma_1\frac{\phi\left(\mathbf{z}'\boldsymbol{\beta}_2\right)}{\Phi\left(\mathbf{z}'\boldsymbol{\beta}_2\right)\left[1 - \Phi\left(\mathbf{z}'\boldsymbol{\beta}_2\right)\right]}. \tag{4.12}$$

Without correction for endogeneity, the OLS coefficient on the treatment dummy estimates this difference. If $\rho > 0$, it follows that $\mathrm{E}\left(y_1 \mid y_2 = 1\right) - \mathrm{E}\left(y_1 \mid y_2 = 0\right) > \alpha_1$. OLS overestimates the treatment effect; there is an upward bias. A consistent estimator of the treatment effect can be obtained from a Heckman-type two-stage estimator or by full MLE. However, both procedures rely on joint normality of ε_1 and ε_2, and the assumption of bivariate normality of ε_1 and ε_2 is too restrictive. The model with a binary endogenous variable also can be estimated under weaker assumptions: $\mathrm{E}\left(\varepsilon_1 \mid \mathbf{x}, \mathbf{z}\right) = 0$; $\mathrm{E}\left(\varepsilon_2 \mid \mathbf{z}\right) = 0$; $\mathrm{E}\left(\varepsilon_1 y_2 \mid \mathbf{x}, \mathbf{z}\right) \neq 0$; and y_2 is linear on both $\mathbf{x}$ and $\mathbf{z}$. Under these assumptions, α_1 and the other parameters in (4.6) are identified, and 2SLS on (4.6) is consistent and asymptotically normal. That is, $\widehat{\boldsymbol{\beta}} = \left[\widetilde{\mathbf{X}}'\mathbf{P}_{\mathbf{Z}}\widetilde{\mathbf{X}}\right]^{-1}\left[\widetilde{\mathbf{X}}'\mathbf{P}_{\mathbf{Z}}\mathbf{y}_1\right]$, where $\widehat{\boldsymbol{\beta}} = \left[\widehat{\boldsymbol{\beta}}_1', \widehat{\alpha}_1\right]'$, $\widetilde{\mathbf{X}} = [\mathbf{X}, \mathbf{y}_2]$, and $\mathbf{P}_{\mathbf{Z}} = \mathbf{Z}\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'$. Since the only endogenous explanatory variable in equation (4.6) is binary, this equation is called a dummy endogenous variable model (Heckman, 1978). As discussed above, there is no special consideration in estimating equation (4.6) by 2SLS when the endogenous explanatory variable is binary.

We can find a more efficient IV estimator by making stronger assumptions as follows: (i) $\mathrm{P}\left(y_2 = 1 \mid \mathbf{x}, \mathbf{z}\right) \neq \mathrm{P}\left(y_2 = 1 \mid \mathbf{x}\right)$ and $\mathrm{P}\left(y_2 = 1 \mid \mathbf{x}, \mathbf{z}\right) = F\left(\mathbf{x}, \mathbf{z}; \boldsymbol{\beta}_2\right)$ is a known parametric form (usually probit or logit); and (ii) $\mathrm{Var}\left(\varepsilon_1 \mid \mathbf{x}, \mathbf{z}\right)$ is constant. Therefore, we can use a two-stage IV method: In stage 1, estimate the binary response model $\mathrm{P}\left(y_2 = 1 \mid \mathbf{x}, \mathbf{z}\right) = F\left(\mathbf{x}, \mathbf{z}; \boldsymbol{\beta}_2\right)$ by MLE. Obtain the fitted probabilities, $\widehat{F}_i$. In stage 2, estimate (4.6) by IV using instruments $\mathbf{x}_i$ and $\widehat{F}_i$. That is, $\widehat{\boldsymbol{\beta}} = \left(\widetilde{\mathbf{Z}}'\widetilde{\mathbf{X}}\right)^{-1}\widetilde{\mathbf{Z}}'\mathbf{y}_1$, where $\widehat{\boldsymbol{\beta}} = \left[\widehat{\boldsymbol{\beta}}_1', \widehat{\alpha}_1\right]'$, $\widetilde{\mathbf{X}} = [\mathbf{X}, \mathbf{y}_2]$, and $\widetilde{\mathbf{Z}} = \left[\mathbf{X}, \widehat{\mathbf{F}}\right]$.

This procedure has an important robustness property. Even if we use $\widehat{F}_i$ as an instrument for y_{2i}, the model for $\mathrm{P}\left(y_2 = 1 \mid \mathbf{x}, \mathbf{z}\right)$ need not be correctly specified. For example, if we specify a probit model for $\mathrm{P}\left(y_2 = 1 \mid \mathbf{x}, \mathbf{z}\right)$, we do not need the probit model to be correct. This requirement is weak when $\mathbf{z}$ is partially correlated with y_2.

4.4 Endogenous Regressors in Nonlinear Health Econometrics

Endogeneity leads to inconsistency of the estimated regression parameters in nonlinear health econometric models. Correlation occurs when there is simultaneous determination of the regressor through a related model, especially in estimation of treatment effects. Since persons participating in clinical trials are randomly assigned to a treatment group or a control group, differences in outcomes are a good estimator of the treatment effect. It is observed, for example, in a drug's effect on the number of epileptic seizures and on the duration of lung cancer.

However, non-experimental observational data are not always drawn from randomly assigned treatment. For instance, the decision to seek care may depend on a patient's private health insurance status (the treatment variable), but private insurance coverage is a choice variable that might depend on subjective health status, age, and income. To obtain consistent estimators of an endogenous regressor, we disregard the standard two-stage method in nonlinear regression models with endogenous variables. This section accounts for endogenous regressors in health econometrics based on Wooldridge (2002), Cameron and Trivedi (2005), and Winkelmann (2004).

4.4.1 Endogenous Regressors in Binary Response Models

The assumption of exogenous regressors might not hold for binary response models such as probit and logit applications. Evans and Schwab (1995) demonstrated this problem in analyzing the probability of graduating high school using the binary regressor of attending a Catholic school. They suggested that the Catholic school dummy is endogenous because parents who care about child's welfare stress good grades and willingly pay for a private school. Costa (1995) analyzed this example of a continuous endogenous variable. This paper estimated a binary (yes/no) retirement decision as a function of a pension benefit. In fact, the estimation may have been plagued with endogeneity because the decision depends on earlier decisions and the decision setting is continuous. This subsection considers estimation for probit models with an endogenous binary or continuous variable. For simplicity, we discuss one endogenous regressor.

A Continuous Endogenous Regressor in Binary Models

A probit model with a continuous endogenous explanatory variable takes the following form:

$$y_1^* = \mathbf{x}'\boldsymbol{\beta}_1 + \alpha_1 y_2 + \varepsilon_1, \tag{4.13}$$

$$y_2 = \mathbf{z}'\boldsymbol{\beta}_2 + \varepsilon_2, \tag{4.14}$$

where $(\varepsilon_1, \varepsilon_2)$ has a zero mean, a bivariate normal distribution, and is independent of $\mathbf{z}$. The observed binary outcome is $y_1 = 1$ if $y_1^* > 0$ and $y_1 = 0$ otherwise. If ε_1 and ε_2 are independent, there is no endogeneity. Since ε_2 is normally distributed, we assume y_2 is normal given $\mathbf{z}$. Therefore, y_2 is a normal random variable.

Rivers and Vuong (1988) proposed a method for estimating a probit model with a continuous endogenous explanatory variable. Their method is a useful two-stage approach leading to a simple test for endogeneity of y_2. See also Wooldridge (2002) and Wikelmann and Boes (2006) for a discussion of the procedure. Assume that ε_1 and ε_2 are bivariate normal distributed with zero mean, correlation ρ, and variance 1 and σ_2^2, respectively. We can write

$$\varepsilon_1 = \theta_1 \varepsilon_2 + u_1, \tag{4.15}$$

where $\theta_1 = \rho/\sigma_2$, $\sigma_2^2 = \text{Var}(\varepsilon_2)$, and u_1 is independent of $\mathbf{z}$ and ε_2 (and therefore of y_2). Because of joint normality of $(\varepsilon_1, \varepsilon_2)$, u_1 is also normally distributed with $\text{E}(u_1) = 0$ and $\text{Var}(u_1) = \text{Var}(\varepsilon_1) - \rho^2$. We can now write

$$\begin{aligned}
&y_1^* = \mathbf{x}'\boldsymbol{\beta}_1 + \alpha_1 y_2 + \theta_1 \varepsilon_2 + u_1, \\
&u_1 \mid \mathbf{z}, y_2, \varepsilon_2 \sim \mathcal{N}\left(0, 1 - \rho^2\right).
\end{aligned}$$

Thus, $\varepsilon_1 = \sqrt{1-\rho^2}u + \rho\varepsilon_2/\sigma_2$, where $u \sim \mathcal{N}(0,1)$. We can write the first equation *conditional* on ε_2 as

$$y_1^* = \alpha_1 y_2 + \mathbf{x}'\boldsymbol{\beta}_1 + \sqrt{1-\rho^2}u + \theta_1\varepsilon_2.$$

A standard calculation shows that

$$\text{P}(y_1 = 1 \mid \mathbf{z}, y_2, \varepsilon_2) = \Phi\left[\left(\mathbf{x}'\boldsymbol{\beta}_1 + \alpha_1 y_2 + \theta_1\varepsilon_2\right) / \left(1-\rho^2\right)^{1/2}\right]. \tag{4.16}$$

Assuming for the moment that we observe ε_2, then probit of y_1 on $\mathbf{z}$, y_2, and ε_2 consistently estimates $\boldsymbol{\beta}_{\rho 1} \equiv \boldsymbol{\beta}_1 / \left(1-\rho^2\right)^{1/2}$, $\alpha_{\rho 1} \equiv \alpha_1 / \left(1-\rho^2\right)^{1/2}$, and $\theta_\rho \equiv \theta_1 / \left(1-\rho^2\right)^{1/2}$. Note that because $\rho^2 < 1$, each scaled coefficient is greater than its unscaled counterpart unless y_2 is exogenous ($\rho = 0$).

The Rivers and Vuong (1988) approach takes the following two stages. First, run the OLS regression y_2 on $\mathbf{z}$ and save the residuals $\widehat{\varepsilon}_2$. Second, the probit y_1 on $\mathbf{x}$, y_2, and $\widehat{\varepsilon}_2$ obtains consistent estimators of the scaled coefficients $\boldsymbol{\beta}_{\rho 1}$, $\alpha_{\rho 1}$, and θ_ρ. The probit parameters are estimated only up to scale, with factor $\left(1-\rho^2\right)^{-1/2}$. An estimate for ρ is $\widehat{\rho}^2 = \widehat{\theta}_\rho^2 \widehat{\sigma}_2^2 / \left(1 + \widehat{\theta}_\rho^2 \widehat{\sigma}_2^2\right)$, where $\widehat{\sigma}_2$ is the square root of the usual error variance estimator from the first stage regression.

The Rivers and Vuong approach simplifies testing the exogeneity of y_2. A z-test of the null hypothesis H_0: $\theta_1 = 0$ tests whether y_2 is exogenous. If there is evidence of endogeneity ($\theta_1 \neq 0$) and we apply a two-stage procedure to find consistent estimators, the usual probit parameters must be adjusted to account for the first stage estimation. Under H_0: $\theta_1 = 0$, we find $u_1 = \varepsilon_1$, and the distribution of ε_2 plays no role under the null. Therefore, the test of exogeneity is effective without assuming normality or homoskedasticity of ε_2. Unfortunately, if y_2 and ε_1 are correlated, normality of ε_2 is crucial.[5]

An alternative approach is to estimate a probit model with a continuous endogenous explanatory variable using conditional maximum likelihood. The joint distribution of (y_1, y_2) conditional on $\mathbf{z}$ takes the form

$$f(y_1, y_2 \mid \mathbf{z}) = f(y_1 \mid y_2, \mathbf{z})\, f(y_2 \mid \mathbf{z}). \tag{4.17}$$

[5] The Rivers and Vuong two-stage approach discusses average partial effects. For more detail, see Wooldridge (2002).

Since $y_2 \mid \mathbf{z} \sim \mathcal{N}\left(\mathbf{z}'\boldsymbol{\beta}_2, \sigma_2^2\right)$, the PDF $f(y_2 \mid \mathbf{z})$ is easily evaluated. Moreover, since $\varepsilon_2 = y_2 - \mathbf{z}'\boldsymbol{\beta}_2$ and $y_1 = 1[y_1^* > 0]$, we also can derive the conditional density of y_1 given $(y_2, \mathbf{z})$:

$$\mathrm{P}(y_1 = 1 \mid y_2, \mathbf{z}) = \Phi\left[\frac{\mathbf{x}'\boldsymbol{\beta}_1 + \alpha_1 y_2 + (\rho/\sigma_2)(y_2 - \mathbf{z}'\boldsymbol{\beta}_2)}{(1-\rho^2)^{1/2}}\right]. \tag{4.18}$$

Then we have derived

$$f(y_1, y_2 \mid \mathbf{z}) = \{\Phi(\cdot)\}^{y_1}\{1 - \Phi(\cdot)\}^{1-y_1}(1/\sigma_2)\,\phi\left[(y_1 - \mathbf{z}'\boldsymbol{\beta}_2)/\sigma_2\right] \tag{4.19}$$

and the log likelihood for observation i is

$$y_1 \ln\Phi(\cdot) + (1-y_1)\ln[1-\Phi(\cdot)] - \frac{1}{2}\ln(2\pi) - \ln\sigma_2 - \frac{1}{2\sigma_2^2}(y_1 - \mathbf{z}'\boldsymbol{\beta}_2)^2. \tag{4.20}$$

Summing (4.20) across all i and maximizing with respect to all parameters obtains the estimators of $(\boldsymbol{\beta}_1, \alpha_1, \rho, \boldsymbol{\beta}_2, \sigma_2)$. Standard errors can be calculated using the estimated Hessian, the estimated expected Hessian, or the outer product of the score.

Conditional maximum likelihood estimation offers advantages over two-stage procedures. First, it is more efficient than any two-stage procedure. Second, we acquire direct estimators of $\boldsymbol{\beta}_1$ and α_1 and the parameters of interest for computing partial effects. Third, it is easy to examine exogeneity of y_2 using the asymptotic t test under H_0: $\rho = 0$ or a likelihood ratio test.

The Rivers and Vuong approach is a limited information procedure and focuses on the conditional density $f(y_1 \mid y_2, \mathbf{z})$, where they replace the unknown $\boldsymbol{\beta}_2$ with the OLS estimator $\widehat{\boldsymbol{\beta}}_2$. However, the conditional maximum likelihood method estimates the parameters using the information in $f(y_1 \mid y_2, \mathbf{z})$ and $f(y_2 \mid \mathbf{z})$ simultaneously. Therefore, Rivers and Vuong has significant computational advantages in testing whether y_2 is exogenous, and using the conditional maximum likelihood method is worthwhile if exogeneity is rejected.

A Binary Endogenous Regressor in Binary Models

We now consider the case where the probit model contains an endogenous binary explanatory variable.[6] The model describes as follows:

$$y_1 = 1\left[\mathbf{x}'\boldsymbol{\beta}_1 + \alpha_1 y_2 + \varepsilon_1 > 0\right], \tag{4.21}$$

$$y_2 = 1\left[\mathbf{z}'\boldsymbol{\beta}_2 + \varepsilon_2 > 0\right], \tag{4.22}$$

where $1[\cdot]$ is an indicator function, $(\varepsilon_1, \varepsilon_2)$ is independent of $\mathbf{z}$ and distributed as bivariate normal with mean zero and covariance matrix $(1, \rho, 1)$. If $\rho \neq 0$, then ε_1 and y_2 are correlated, and the probit estimation is inconsistent for $\boldsymbol{\beta}_1$ and α_1. In this model, the effect of y_2 is often

[6] The following discussion is based on van de Ven *et al.* (1981) and Winkelmann and Boes (2006).

of primary interest, especially when y_2 indicates participation in some program, such as health maintenance, and the binary outcome y_1 might denote a subjective health index. Then the average treatment effect (for a given value of $\mathbf{x}$) is calculated by $\Phi\left(\mathbf{x}'\boldsymbol{\beta}_1 + \alpha_1\right) - \Phi\left(\mathbf{x}'\boldsymbol{\beta}_1\right)$.

The likelihood function is easily calculated using the conditional density and truncated normal distributions. The conditional density of y_1 given $(y_2, \mathbf{z})$ takes the following form:

$$\mathrm{P}\left(y_1 = 1 \mid y_2, \mathbf{z}\right) = \Phi\left[\frac{\mathbf{x}'\boldsymbol{\beta}_1 + \alpha_1 y_2 + \rho\varepsilon_2}{\left(1-\rho^2\right)^{1/2}}\right]. \tag{4.23}$$

Moreover, the truncated density of ε_2 given $\varepsilon_2 > -\mathbf{z}'\boldsymbol{\beta}_2$ obtains

$$\frac{\phi\left(\varepsilon_2\right)}{\mathrm{P}\left(\varepsilon_2 > -\mathbf{z}'\boldsymbol{\beta}_2\right)} = \frac{\phi\left(\varepsilon_2\right)}{\Phi\left(\mathbf{z}'\boldsymbol{\beta}_2\right)}. \tag{4.24}$$

Therefore, the density $\mathrm{P}\left(y_1 = 1 \mid y_2 = 1, \mathbf{z}\right)$ takes

$$\begin{aligned}\mathrm{P}\left(y_1 = 1 \mid y_2 = 1, \mathbf{z}\right) &= \mathrm{E}\left\{\Phi\left[\frac{\mathbf{x}'\boldsymbol{\beta}_1 + \alpha_1 y_2 + \rho\varepsilon_2}{\left(1-\rho^2\right)^{1/2}}\right] \mid y_2 = 1, \mathbf{z}\right\} \\ &= \frac{1}{\Phi\left(\mathbf{z}'\boldsymbol{\beta}_2\right)} \int_{-\mathbf{z}'\boldsymbol{\beta}_2}^{\infty} \Phi\left[\frac{\mathbf{x}'\boldsymbol{\beta}_1 + \alpha_1 + \rho\varepsilon_2}{\left(1-\rho^2\right)^{1/2}}\right] \mathrm{d}\,\varepsilon_2. \end{aligned} \tag{4.25}$$

Similarly, $\mathrm{P}\left(y_1 = 1 \mid y_2 = 0, \mathbf{z}\right)$ is

$$\mathrm{P}\left(y_1 = 1 \mid y_2 = 0, \mathbf{z}\right) = \frac{1}{1 - \Phi\left(\mathbf{z}'\boldsymbol{\beta}_2\right)} \int_{-\infty}^{-\mathbf{z}'\boldsymbol{\beta}_2} \Phi\left[\frac{\mathbf{x}'\boldsymbol{\beta}_1 + \rho\varepsilon_2}{\left(1-\rho^2\right)^{1/2}}\right] \mathrm{d}\,\varepsilon_2. \tag{4.26}$$

Combining the four possible outcomes of (y_1, y_2), we obtain the log-likelihood function of the probit model with a binary endogenous explanatory variable.

Since the log-likelihood function includes a single integral but has no analytical solution, we evaluate the likelihood using a numerical integral. If the integral is distributed over $[-\infty, \infty]$, the log-likelihood is easily evaluated by applying Gauss-Hermite quadrature. In this model, we calculate the log-likelihood function using $\varepsilon_q > -\mathbf{z}'\boldsymbol{\beta}_2$, where ε_q is the evaluation point of the Gauss-Hermite quadrature. Since $-\mathbf{z}'\boldsymbol{\beta}_2$ is not constant under the maximization process, a small change in the value of $\boldsymbol{\beta}_2$ does not alter the likelihood, and thus the performance of the Gauss-Hermite quadrature is low. The simulated maximum likelihood method avoids this problem but needs many evaluation points to approximate the integral accurately.[7] Moreover, calculating the accurate likelihood is time consuming.

Therefore, it is possible to apply the Rivers and Vuong two-stage approach for estimating the probit model with an endogenous binary explanatory variable: since $\mathrm{E}\left(y_2 \mid \mathbf{z}\right) = \Phi\left(\mathbf{z}'\boldsymbol{\beta}_2\right)$ and $\boldsymbol{\beta}_2$ is consistently estimated by the probit of y_2 on $\mathbf{z}$, it is tempting to estimate $\boldsymbol{\beta}_1$ and α_1 from the probit of y_1 on $\mathbf{x}$ and $\widehat{\Phi}_2$, where $\widehat{\Phi}_2 \equiv \Phi\left(\mathbf{z}'\widehat{\boldsymbol{\beta}}_2\right)$. However, the two-stage method

[7] Judd (1998, Chapter 7) accounts for numerical integrals like the Gauss-Hermite quadrature. For details of the simulated likelihood, see Gouriéroux and Monfort (1996) and Gouriéroux (2000).

is inappropriate because the estimated coefficients are inconsistent. Although the two-stage method requires $\mathrm{P}(y_1 = 1 \mid \mathbf{z}) = \Phi[\mathbf{x}'\boldsymbol{\beta}_1 + \alpha_1 \Phi(\mathbf{z}'\boldsymbol{\beta}_2)]$, we can compute only the expected value $\mathrm{P}(y_1 = 1 \mid \mathbf{z}) = \mathrm{E}(y_1 = 1 \mid \mathbf{z}) = \mathrm{E}(1[\mathbf{x}'\boldsymbol{\beta}_1 + \alpha_1 y_2 + \varepsilon_1 > 0])$. Since the indicator function $1[\cdot]$ is nonlinear, we cannot correctly specify the expected value. If, substituting $\widehat{\boldsymbol{\beta}}_2$, we can compute the correct and complicated formula for $\mathrm{P}(y_1 = 1 \mid \mathbf{z})$, the two-stage approach produces consistent estimators, but the full maximum likelihood method is easier and more efficient.[8]

4.4.2 Endogenous Regressors in Count Data Models

It is easy to expand the Poisson regression model into one with endogenous variables and, as with a probit model containing endogenous variables, count data models present the same problem.[9] Fortunately, the two-stage approach discussed concerning binary models can be applied to the Poisson regression model with a continuous endogenous variable, and the estimated coefficients are consistent. The following analysis discusses count data models with endogenous binary variables. For simplicity, we discuss a single endogenous regressor only.

A Binary Endogenous Regressors in Count Data Models

A Poisson model with a binary endogenous variable is essentially the same as a binary response model with a binary endogenous variable. We consider the triangular model

$$y_1 = \exp(\alpha_1 y_2 + \mathbf{x}'\boldsymbol{\beta}_1 + \varepsilon_1), \tag{4.27}$$

$$y_2^* = \mathbf{z}'\boldsymbol{\beta}_2 + \varepsilon_2, \tag{4.28}$$

where ε_1 and ε_1 are unobserved heterogeneity that satisfy $\mathrm{E}(\varepsilon_1 \mid \mathbf{x}, \mathbf{z}) = \mathrm{E}(\varepsilon_2 \mid \mathbf{x}, \mathbf{z}) = 0$ and $\mathrm{Cov}(\varepsilon_1, \varepsilon_2) \neq 0$. Moreover, y_2^* is a latent variable, and a binary variable y_2 is observed when

$$y_2 = \begin{cases} 1 & \text{if} \quad y_2^* > 0 \\ 0 & \text{otherwise} \end{cases}.$$

[8] If a sample is censoring or truncated and might have endogenous explanatory variables, microeconometrics utilizes Tobit models. In this case, the same discussion of binary models is applied because the properties of censored or truncated data are essentially same as those of binary data. See Angrist (2001) and Lee and Vella (2006).

[9] The following discussion derives from Greene (1997), Windmeijer and Santos Silva (1997), Terza (1998), Schellhorn (2001), and Winkelmann (2003). Miranda and Rabe-Hesketh (2006) provide a useful stata ado file to estimate binary, count, and ordinal variables with endogenous variables.

Next, we consider the consequences of ignoring the endogeneity of y_2. From (4.27),

$$\begin{aligned}\frac{\mathrm{E}\left(y_1 \mid \mathbf{x}, y_2=1\right)}{\mathrm{E}\left(y_1 \mid \mathbf{x}, y_2=0\right)} &= \frac{\mathrm{E}_{\varepsilon_1}\mathrm{E}\left(y_1 \mid \mathbf{x}, \varepsilon_1, y_2=1\right)}{\mathrm{E}_{\varepsilon_1}\mathrm{E}\left(y_1 \mid \mathbf{x}, \varepsilon_1, y_2=0\right)} \\ &= \frac{\exp\left(\alpha_1+\mathbf{x}'\boldsymbol{\beta}_1\right)\mathrm{E}\left(\varepsilon_1 \mid y_2=1\right)}{\exp\left(\mathbf{x}'\boldsymbol{\beta}_1\right)\mathrm{E}\left(\varepsilon_1 \mid y_2=0\right)}. \qquad (4.29)\end{aligned}$$

To simplify the discussion and to evaluate expectations easily, we specify that the vector $(\varepsilon_1, \varepsilon_2)$ follows a bivariate normal distribution with zero mean and covariance matrix $\left(\sigma_1^2, \rho\sigma_1, 1\right)$. Moreover, using the results on truncation in the log-normal distribution, we obtain

$$\mathrm{E}\left(\exp\left(\varepsilon_1\right) \mid y_2=1\right) = \exp\left(1/2\sigma_1^2\right)\frac{\Phi\left(\mathbf{z}'\boldsymbol{\beta}_2+\rho\sigma_1\right)}{\Phi\left(\mathbf{z}'\boldsymbol{\beta}_2\right)} \qquad (4.30)$$

and

$$\mathrm{E}\left(\exp\left(\varepsilon_1\right) \mid y_2=0\right) = \exp\left(1/2\sigma_1^2\right)\frac{\Phi\left(-\mathbf{z}'\boldsymbol{\beta}_2-\rho\sigma_1\right)}{\Phi\left(-\mathbf{z}'\boldsymbol{\beta}_2\right)}. \qquad (4.31)$$

Therefore, under the assumption of this model,

$$\frac{\mathrm{E}\left(y_1 \mid \mathbf{x}, y_2=1\right)}{\mathrm{E}\left(y_1 \mid \mathbf{x}, y_2=0\right)} = \exp\left(\alpha_1\right)\frac{\Phi\left(\mathbf{z}'\boldsymbol{\beta}_2+\rho\sigma_1\right)}{\Phi\left(\mathbf{z}'\boldsymbol{\beta}_2\right)}\frac{\Phi\left(-\mathbf{z}'\boldsymbol{\beta}_2\right)}{\Phi\left(-\mathbf{z}'\boldsymbol{\beta}_2-\rho\sigma_1\right)}. \qquad (4.32)$$

If $\rho > 0$, the factor following $\exp\left(\alpha_1\right)$ is greater than 1. That is, the overall relative difference between the two expected counts exceeds $\exp\left(\alpha_1\right)-1$. This leads to an upward bias in the estimated effect, making it important to treat the endogeneity of y_2 carefully.

In estimation, we can, as before, replace y_2 in the first equation with its probability $F\left(\mathbf{z}'\boldsymbol{\beta}_2\right)$, say, $\Phi\left(\mathbf{z}'\boldsymbol{\beta}_2\right)$. However, as in binary models, this procedure will not work in this nonlinear model. In this case,

$$\mathrm{E}\left(y_1 \mid \mathbf{z}'\boldsymbol{\beta}_2, \mathbf{x}, y_2\right) = \exp\left(\alpha_1 F\left(\mathbf{z}'\boldsymbol{\beta}_2\right)+\mathbf{x}'\boldsymbol{\beta}_1\right)\exp\left(\alpha_1 u_2\right),$$

where $u_2 = y_2 - F\left(\mathbf{z}'\boldsymbol{\beta}_2\right)$. Under this assumption, since the moments of u_2 depend on $\mathbf{z}$, e.g., $\mathrm{E}\left(u_2^2 \mid \mathbf{z}\right) = F\left(\mathbf{z}'\boldsymbol{\beta}_2\right)\left[1-F\left(\mathbf{z}'\boldsymbol{\beta}_2\right)\right]$, $\exp\left(u_2\right)$ and $\mathbf{z}$ are not independent However, $\mathrm{E}\left(\exp\left(u_2\right)\right)$ is an increasing function of its variance because of the convexity of the exponential transformation and depends on both parameters and regressors. Therefore, as with a binary model containing a binary endogenous variable, the two-stage approach is inappropriate for count data models with a binary endogenous variable.

Another explanation is as follows. As before, we assume for simplicity that the vector $(\varepsilon_1, \varepsilon_2)$ follows a bivariate normal distribution with zero mean and covariance matrix. Using the results for truncation in the log-normal distribution,

$$\mathrm{E}\left(y_1 \mid \mathbf{x}, y_2=1\right) = \exp\left(\alpha_1+\mathbf{x}'\boldsymbol{\beta}_1+\frac{\sigma_1^2}{2}\right)\left(\frac{\Phi\left(\mathbf{z}'\boldsymbol{\beta}_2+\rho\sigma_1\right)}{\Phi\left(\mathbf{z}'\boldsymbol{\beta}_2\right)}\right), \qquad (4.33)$$

This regression may be estimated by nonlinear least squares after substituting in the first stage estimators of $\boldsymbol{\beta}_2$ denoted $\widehat{\boldsymbol{\beta}}_2$. This stage introduces heteroskedastic errors into the regression and complicates the estimation of asymptotic covariance matrix. The effect of sample selection on the exponential conditional mean of the count data models is multiplicative, not additive as in the normal linear case. This means that *ad hoc* adjustment based on adding Mill's ratio to the conditional mean does not work well. Windmeijer and Santos Silva (1997) propose an instrumental variables estimator, instrumenting y_2 by $F\left(\mathbf{z}'\widehat{\boldsymbol{\beta}}_2\right)$, where $\widehat{\boldsymbol{\beta}}_2$ is obtained from estimating a probit or logit model first. Terza (1998) derives a two-stage moment estimator that does not require the specification of the full distribution of y_1, and Kozumi (1999) provides a Bayesian analysis of this model. However, the full maximum likelihood method is easier and more efficient. We describe it below.[10]

We consider a fully specified count data model with endogenous binary regressors. As before, error terms ε_1 and ε_2 follow a bivariate normal distribution with correlation parameter ρ. The joint PDF takes the form

$$\begin{aligned} f(y_1, y_2) &= y_2 f(y_1, y_2 = 1) + (1 - y_2) f(y_1, y_2 = 0) \\ &= y_2 \left[f(y_1 \mid y_2 = 1) \mathrm{P}(y_2 = 1) \right] + (1 - y_2) \left[f(y_1 \mid y_2 = 0) \mathrm{P}(y_2 = 0) \right]. \end{aligned}$$

Under the independence assumption, expectations $f(y_1 \mid y_2 = 1) \mathrm{P}(y_2 = 1)$ are easily calculated by multiplying standard distributions. We must consider these expressions conditional on ε_1 and obtain the desired quantities by integrating $f(y_1, y_2, \varepsilon_1)$ over ε_1 in a second stage:

$$\begin{aligned} f(y_1, y_2) &= \int_{-\infty}^{\infty} f(y_1, y_2, \varepsilon_1) \, \mathrm{d}\, \varepsilon_1 \\ &= \int_{-\infty}^{\infty} f(y_1 \mid y_2, \varepsilon_1) f(y_2 \mid \varepsilon_1) g(\varepsilon_1) \, \mathrm{d}\, \varepsilon_1. \end{aligned} \tag{4.34}$$

The first distribution under the integral is simply the specified count data PDF with mean function $\exp(\alpha_1 y_2 + \mathbf{x}'\boldsymbol{\beta}_1)$. The second distribution under the integral is a Bernoulli distribution

$$f(y_2 \mid \varepsilon_1) = \mathrm{P}(y_2 = 1 \mid \varepsilon_1)^{y_2} \left[1 - \mathrm{P}(y_2 = 0 \mid \varepsilon_1)^{1 - y_2} \right], \tag{4.35}$$

where

$$\mathrm{P}(y_2 = 1 \mid \varepsilon_1) = \mathrm{P}(\varepsilon_2 > -\mathbf{z}'\boldsymbol{\beta}_2 \mid \varepsilon_1) \equiv \Phi^*(\varepsilon_1) \tag{4.36}$$

[10] For other methods to estimate count data or nonlinear models with endogenous regressors, see Freund *et al.* (1999), Windmeijer (2000), Romeu and Vera-Hernández (2000, 2005), van Ophem (2000), Munkin (2003), Munkin and Trivedi (2003), Lee (2004), Deb and Trivedi (2006), Deb *et al.* (2006a, 2006b), Zimmer and Trivedi (2006), and Firpo (2007).

and $\Phi^*(\varepsilon_1)$ is the cumulative distribution function of a standard normal distribution. Finally, $g(\varepsilon_1)$ is a normal distribution with mean 0 and variance σ_1^2.

Therefore, the log-likelihood function of count data models with an endogenous binary variable takes the form

$$f(y_1, y_2) = \int_{-\infty}^{\infty} f(y_1 \mid y_2, \varepsilon_1) \, \Phi^*(\varepsilon_1)^{y_2} \left[1 - \Phi^*(\varepsilon_1)\right]^{1-y_2} g(\varepsilon_1) \, \mathrm{d}\,\varepsilon_1. \tag{4.37}$$

The integral is easily evaluated using quadrature or other simulation methods. The parameter can be estimated by maximizing the log-likelihood function of the sample with respect to α_1, $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, σ_1, and ρ.

4.5 Monte Carlo Results

This section is devoted to a Monte Carlo study to evaluate the finite sample performance of the various analyzed models with endogenous variables. We show the inconsistency of the two-stage method in estimating nonlinear models, such as probit and Poisson models, with an endogenous binary variable.

First, we summarize results of the Monte Carlo experiments of linear estimation with endogenous continuous and binary variables. The Monte Carlo simulations are designed as follows. We generate one explanatory variable, z_1, drawn independently from $\mathcal{N}(0, 1/4)$, and two unobserved heterogeneity terms, ε_1 and ε_2, normally distributed as $\mathcal{N}\left((0,0), \left(\sigma_1^2, \rho\sigma_1\sigma_2, \sigma_2^2\right)\right)$. The variable y_2 represents an endogenous continuous variable assumed to be generated by the process $y_2 = \mathbf{z}'\boldsymbol{\beta}_2 + \varepsilon_2$; d represents an endogenous binary variable and is assumed to be generated by the process $d = 1$ if $d^* = \mathbf{z}'\boldsymbol{\beta}_2 + \varepsilon_2 > 0$ and $d = 0$; otherwise, where $\mathbf{z} = [1, z_1]'$ and $\boldsymbol{\beta}_2 = [\beta_{21}, \beta_{22}]'$. Moreover, y_1 represents a continuous variable assumed to be generated by the process $y_1 = \beta_{11} + \beta_{12}y_2 + \varepsilon_1$ or $y_1 = \beta_{11} + \beta_{12}d + \varepsilon_1$.

All true values for parameters $\beta_{11} = \beta_{12} = \beta_{21} = \beta_{22} = 0.5$ and $\sigma_1 = \sigma_2 = 1$ are the same for each experiment. Correlation parameter ρ takes values of 0.3, 0.6, and 0.9. The number of simulations used in all experiments is set to 100, and the sample sizes are 1,000 and 2,000 observations per Monte Carlo iteration. Simulations are performed on Intel Core 2 Duo workstations using GAUSS.

Tables 4.5 and 4.6 show the results of Monte Carlo experiments on linear models with endogenous continuous and binary variables. We estimate both experiments by 2SLS using instrumental variable $\mathbf{z}$. Results for parameters (β_{11}, β_{12}), given in Tables 4.5 and 4.6, show the estimates are consistent for each estimation and the value of the correlation parameter. Actually, the test statistics of H_0: $\beta_{11} = 0.5$ or $\beta_{12} = 0.5$ are not rejected at the 50% level in all experiments. Moreover, the bias decreases when the sample size increases for each iteration.

Table 4.5: Monte Carlo Results of Linear Models with an Endogenous Continuous Variable

	Truth	$N = 1{,}000$	$N = 2{,}000$	$N = 1{,}000$	$N = 2{,}000$	$N = 1{,}000$	$N = 2{,}000$
Two-stage							
β_{11}	0.5	0.012	−0.002	0.014	0.002	0.016	0.006
		(0.077)	(0.053)	(0.079)	(0.055)	(0.082)	(0.058)
β_{12}	0.5	−0.018	−0.001	−0.020	−0.008	−0.019	−0.017
		(0.141)	(0.091)	(0.146)	(0.096)	(0.150)	(0.101)
σ_1	1	0.009	0.004	0.014	0.009	0.016	0.016
		(0.050)	(0.029)	(0.096)	(0.059)	(0.142)	(0.093)
ρ		0.008	0.000	0.002	0.001	−0.001	0.001
		(0.129)	(0.088)	(0.092)	(0.065)	(0.027)	(0.020)

Notes: Figures without parentheses are mean bias (BIAS) values. Root mean squared errors (RMSE) appear in parentheses. Two-stage and FIML are the two-stage estimation and full information MLE, respectively.

Table 4.6: Monte Carlo Results of Linear Models with an Endogenous Binary Variable

		$\rho = 0.3$		$\rho = 0.6$		$\rho = 0.9$	
	Truth	$N = 1{,}000$	$N = 2{,}000$	$N = 1{,}000$	$N = 2{,}000$	$N = 1{,}000$	$N = 2{,}000$
Two-stage							
β_{11}	0.5	0.041	−0.003	0.048	0.012	0.047	0.031
		(0.287)	(0.187)	(0.295)	(0.195)	(0.305)	(0.207)
β_{12}	0.5	−0.056	0.000	−0.062	−0.022	−0.059	−0.048
		(0.418)	(0.268)	(0.430)	(0.281)	(0.442)	(0.297)
σ_1	1	0.017	0.007	0.023	0.012	0.026	0.020
		(0.055)	(0.030)	(0.102)	(0.060)	(0.150)	(0.096)
ρ		−0.057	−0.075	−0.141	−0.146	−0.224	−0.218
		(0.173)	(0.119)	(0.145)	(0.104)	(0.096)	(0.073)

Note: See Table 4.5 for notes.

In particular, the bias of endogenous variable parameter β_{12} shows good performance in equations featuring an endogenous continuous or a binary variable. RMSE decreases as the sample size increases for each experiment. In the case of binary variables, the bias of ρ does not necessarily decrease when the number of observations is large, but it increases when ρ increases. This may be the limitation of 2SLS.

Second, we present results of the Monte Carlo experiments of probit estimations in models containing an endogenous continuous or a binary variable. As demonstrated, the two-stage procedure does not yield consistent estimation in the case of a binary endogenous variable.

Table 4.7: Monte Carlo Results of Probit Models with an Endogenous Continuous Variable

		$\rho = 0.3$		$\rho = 0.6$		$\rho = 0.9$	
	Truth	$N = 1{,}000$	$N = 2{,}000$	$N = 1{,}000$	$N = 2{,}000$	$N = 1{,}000$	$N = 2{,}000$
Two-stage							
β_{11}	0.5	−0.003	−0.011	0.006	−0.008	0.007	−0.001
		(0.086)	(0.064)	(0.057)	(0.051)	(0.041)	(0.033)
β_{12}	0.5	−0.005	0.001	−0.010	0.019	−0.016	0.014
		(0.237)	(0.154)	(0.245)	(0.180)	(0.242)	(0.194)

Note: See Table 4.5 for notes.

However, we attempt this method and compare results with the FIML method. The data-generating processes of endogenous variables are the same as those of linear models. Variable y_1 represents a binary dependent variable assumed to be generated by the process $y_1 = 1$ if $y_1^* = \beta_{11} + \beta_{12} d + \varepsilon_1 > 0$ (for an endogenous binary variable) or $y_1^* = \beta_{11} + \beta_{12} y_2 + \varepsilon_1 > 0$ (for an endogenous continuous variable) and $y_1 = 0$; otherwise. All true values for the parameters $\beta_{11} = \beta_{12} = \beta_{21} = \beta_{22} = 0.5$ and $\sigma_1 = \sigma_2 = 1$ are the same for each experiment. Correlation parameter ρ takes values of 0.3, 0.6, and 0.9.

In estimating the probit model with an endogenous binary variable using FIML, the log-likelihood function includes a single integral but has no analytical solution. Moreover, it is obvious in (4.25) that the integral is not distributed over $[-\infty, \infty]$ and contains a censored part. In this case, it is difficult to apply Gauss-Hermite (GH) quadrature, which is often used in a numerical integral and is easily evaluated. Instead, we use the simulated maximum likelihood (SML) method, Halton (1960) sequences, and the GHK simulator,[11] due to Geweke (1992), Hajivassiliou and McFadden (1994), and Keane (1994) to evaluate log-likelihood in the censored part. Chapter 6 discusses details.

Tables 4.7 and 4.8 show the results of Monte Carlo experiments on probit models with endogenous continuous and binary variables. The experiment is estimated using the two-stage method in Table 4.7 and both 2SLS and FIML in Table 4.8. Although, as previously analyzed, the two-stage method is inconsistent in estimating the probit model with an endogenous binary variable, we confirm the extent of this problem. From Table 4.7, the mean bias (BIAS) and RMSE of an endogenous continuous variable decrease when the number of observations

[11] Train (2003) explains the simplified version of the GHK simulator and applies it to mixed logit models. See also McFadden and Train (2000), Train (2000), and Bhat (2001). For the discussion of simulated maximum likelihood or numerical integration, see Gouriéroux and Monfort (1991), Geweke (1995), and Geweke and Keane (1999).

Table 4.8: Monte Carlo Results of Probit Models with an Endogenous Binary Variable

		$\rho = 0.3$		$\rho = 0.6$		$\rho = 0.9$	
	Truth	$N = 1{,}000$	$N = 2{,}000$	$N = 1{,}000$	$N = 2{,}000$	$N = 1{,}000$	$N = 2{,}000$
FIML							
β_{11}	0.5	0.012	−0.014	0.009	−0.001	0.010	0.005
		(0.307)	(0.216)	(0.231)	(0.170)	(0.123)	(0.094)
β_{12}	0.5	−0.030	0.008	−0.007	−0.002	0.013	0.005
		(0.494)	(0.331)	(0.436)	(0.304)	(0.316)	(0.226)
		$N = 1{,}000$	$N = 2{,}000$	$N = 1{,}000$	$N = 2{,}000$	$N = 1{,}000$	$N = 2{,}000$
Two-stage							
β_{11}	0.5	0.056	0.011	0.133	0.114	0.362	0.351
		(0.348)	(0.240)	(0.292)	(0.220)	(0.178)	(0.142)
β_{12}	0.5	−0.123	−0.049	−0.274	−0.250	−0.792	−0.787
		(0.581)	(0.385)	(0.574)	(0.416)	(0.444)	(0.343)

Note: See Table 4.5 for notes.

is large. Since the test statistics that an endogenous variable equals the true value are not rejected at 50%, this experiment shows consistency of the parameter β_{12}.

The results of probit models with an endogenous binary variable appear in Table 4.8. FIML results show the phenomena of consistency, although BIAS and RMSE do not always decrease. The test statistics of H_0: $\beta_{11} = 0.5$ or $\beta_{12} = 0.5$ are not rejected at the 50% level. However, in the two-stage method, the values of BIAS and RMSE are larger than those of FIML. The test statistics of H_0: $\beta_{11} = 0.5$ or $\beta_{12} = 0.5$ are rejected at the 5% level in the case of $N = 2{,}000$ and $\rho = 0.9$. That is, the estimated estimators of the two-stage method are statistically different from the true values. Although the test statistics are not rejected if ρ is small, the inconsistency is apparent if ρ is large. These results suggest it is necessary to use the FIML estimator and not the two-stage method when estimating a probit model with an endogenous binary variable.

Finally, we examine results of Monte Carlo experiments on Poisson log-normal estimation with an endogenous continuous and binary variable. The data-generating processes of endogenous variables are the same as those of linear models, and that of the Poisson component is obtained by $y_1 \sim \text{Poisson}(\lambda)$ with $\lambda = \exp(\beta_{11} + \beta_{12} y_2 + \varepsilon_1)$, where ε_1 is an unobserved error term. All true values for parameters $\beta_{11} = \beta_{12} = \beta_{21} = \beta_{21} = \beta_{22} = 0.5$ and $\sigma_1 = \sigma_2 = 1$ are the same for each experiment, and correlation parameter ρ takes values of 0.3, 0.6, and 0.9.

Table 4.9: Monte Carlo Results of Poisson Models with an Endogenous Continuous Variable

		$\rho = 0.3$		$\rho = 0.6$		$\rho = 0.9$	
	Truth	$N = 1{,}000$	$N = 2{,}000$	$N = 1{,}000$	$N = 2{,}000$	$N = 1{,}000$	$N = 2{,}000$
Two-stage							
β_{11}	0.5	0.010	0.007	0.005	0.004	0.005	0.008
		(0.090)	(0.082)	(0.094)	(0.074)	(0.098)	(0.065)
β_{12}	0.5	-0.008	-0.007	0.004	0.003	-0.002	-0.017
		(0.164)	(0.146)	(0.168)	(0.125)	(0.180)	(0.118)

Note: See Table 4.5 for notes.

Table 4.10: Monte Carlo Results of Poisson Models with an Endogenous Binary Variable

		$\rho = 0.3$		$\rho = 0.6$		$\rho = 0.9$	
	Truth	$N = 1{,}000$	$N = 2{,}000$	$N = 1{,}000$	$N = 2{,}000$	$N = 1{,}000$	$N = 2{,}000$
FIML							
β_{11}	0.5	-0.028	-0.009	-0.017	-0.042	-0.026	-0.031
		(0.283)	(0.210)	(0.219)	(0.189)	(0.159)	(0.116)
β_{12}	0.5	0.037	0.013	0.016	0.059	0.030	0.038
		(0.404)	(0.296)	(0.302)	(0.266)	(0.205)	(0.144)
		$N = 1{,}000$	$N = 2{,}000$	$N = 1{,}000$	$N = 2{,}000$	$N = 1{,}000$	$N = 2{,}000$
Two-stage							
β_{11}	0.5	0.034	-0.029	-0.045	-0.069	-0.141	-0.167
		(0.370)	(0.235)	(0.356)	(0.219)	(0.332)	(0.239)
β_{12}	0.5	-0.044	0.039	0.057	0.085	0.155	0.184
		(0.534)	(0.336)	(0.505)	(0.314)	(0.481)	(0.346)

Note: See Table 4.5 for notes.

Tables 4.9 and 4.10 present the results of Monte Carlo experiments on Poisson models with endogenous continuous and binary variables. Table 4.9 presents using the two-stage method and Table 4.10 the results of both 2SLS and FIML although the two-stage method is theoretically inconsistent in the probit model with an endogenous binary variable. Table 4.9 shows that the BIAS and RMSE of an endogenous continuous variable decrease when the number of observations is large. Moreover, this experiment shows consistency of parameters (β_{11}, β_{12}) because the test statistics of H_0: $\beta_{11} = 0.5$ or $\beta_{12} = 0.5$ are not rejected at the 50% level.

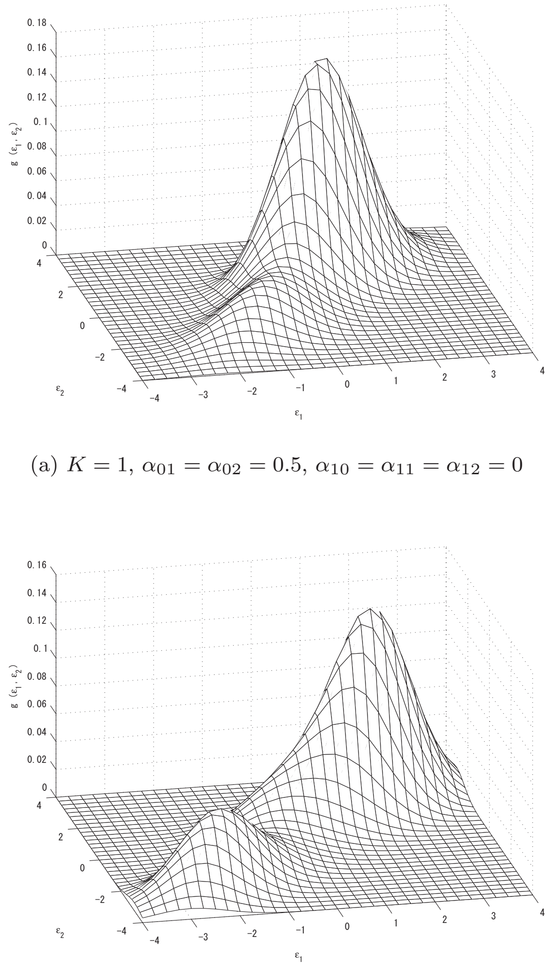
In Table 4.10, Monte Carlo results of Poisson models with an endogenous binary variable show that the FIML results show the phenomena of consistency, but the BIAS and RMSE of the two-stage method are larger than those of the FIML. Test statistics that an endogenous variable equals the true value are not rejected at 50% confidence in the FIML; this experiment shows the parameter β_{12} is consistent. In the case of the two-stage method, the test statistics for $\beta_{12} = 0.5$ are rejected at the 50% level for $N = 2{,}000$ and $\rho = 0.9$. However, no clear evidence of inconsistency appears in our setup. When $\rho = 0.9$ and $N = 5{,}000$, the test statistics are rejected at the 25% level. There are two reasons why we find no clear inconsistency with the two-stage method in the Poisson estimation with an endogenous binary variable: the count variables take the non-negative integer, and the Poisson estimation is easily approximated by linear models. That is, the nonlinearity of count data models is weaker than that of binary variable models. Nonetheless, when estimating the Poisson model with an endogenous binary variable, the two-stage method is not recommended because it is theoretically inconsistent and we document inconsistency when ρ is large.

4.6 Limitations and Extensions of Endogenous Regressors in Health Econometrics

As explained, regardless of whether endogenous regressors are continuous, linear models with those variables display consistency using 2SLS. This characteristic is convenient because OLS achieves a range of consistency. Therefore, although it is difficult to obtain instruments that are uncorrelated with the error term and correlated with regressors, the two-stage method in linear models with endogenous variables presents no serious theoretical problem.

If the two-stage method is applied in estimating nonlinear models with endogenous discrete, censored, or truncated regressors, the estimated parameter has no consistency. Hence, estimating such models requires FIML. However, this method always contains the problem of specifications of distribution. That is, if the distributions of both nonlinear dependent and endogenous variables are not specified correctly, estimated coefficients fail to attain consistency. Since the true distribution remains unknown, this problem is always discussed. One way to avoid a specification problem is to generalize distributions of dependent and endogenous variables. That is, introduce semi- or non-parametric distributions.

The following chapters discuss semiparametric nonlinear models with an endogenous binary variable. Chapter 5 generalizes a count data model with an endogenous binary variable. Chapter 6 analyzes duration analysis with treatment effects. Both chapters introduce semiparametric (semi-nonparametric) binary distribution with Hermite polynomials, proposed by van der Klaauw and Koning (2003). This distribution is natural expansion of bivariate normal



(a) $K = 1$, $\alpha_{01} = \alpha_{02} = 0.5$, $\alpha_{10} = \alpha_{11} = \alpha_{12} = 0$

(b) $K = 2$, $\alpha_{01} = \alpha_{02} = 0.5$, $\alpha_{10} = \alpha_{11} = \alpha_{12} = -0.1$, $\alpha_{20} = \alpha_{21} = \alpha_{22} = 0.2$

Figure 4.1: Semi-nonparametric Bivariate Normal Distributions

distribution and takes the following form:

$$\begin{aligned} g\left(\varepsilon_1, \varepsilon_2\right) &= \frac{1}{P}\left(\sum_{j=0}^{K}\sum_{k=0}^{K}\alpha_{jk}\varepsilon_1^j\varepsilon_2^k\right)^2 \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \\ &\times \exp\left[-\frac{1}{2\left(1-\rho^2\right)}\left\{\left(\frac{\varepsilon_1}{\sigma_1}\right)^2 - 2\rho\frac{\varepsilon_1}{\sigma_1}\frac{\varepsilon_2}{\sigma_2} + \left(\frac{\varepsilon_2}{\sigma_2}\right)^2\right\}\right] \equiv \frac{g^*}{P}, \end{aligned} \tag{4.38}$$

where ε_1 and ε_2 are random variables, $P = \iint_{-\infty}^{\infty} g^* \mathrm{d}\,\varepsilon_1\,\mathrm{d}\,\varepsilon_2$ ensures integration to 1 by scaling the density, σ_1 and ρ are standard deviation and correlation parameters, respectively, and α_{jk} is the parameter to be estimated. When $\alpha_{jk} = 0$ ($\forall j \geq 1$ and $\forall k \geq 1$), this density results in a bivariate normal distribution. Figure 4.1 shows some examples of semiparametric (semi-nonparametric) bivariate normal distributions and this distribution is sometimes fat-tailed and has twin peaks.

This distribution is simple but has difficulty to built in nonlinear regression models with an endogenous binary variable because there are computation problems in estimating these models. In Chapters 5 and 6, we avoid computational difficulties and propose generalized models with endogeneity.

Chapter 5

Semiparametric Count Data Estimation with an Endogenous Binary Variable

5.1 Introduction[1]

Count data models explain the behavior of discrete and non-negative dependent random variables and are used in applied econometrics such as industrial organization, health economics, and population economics. A Poisson model is one of the methods to estimate count data. Moreover, many recent studies use a negative binomial 2 (NB2) model that assumes additive separable log-gamma distributed heterogeneity.

In microeconometric applications, we often come across situations where explanatory variables (in particular, an endogenous binary variable) are simultaneously determined with the dependent variable. In this case, the Poisson and NB2 models yield biased estimates of parameters of interest because these models assume perfect exogeneity of explanatory variables. Therefore, count data models with an endogenous binary variable are required, and many studies have been conducted to analyze this problem. For example, Terza (1998) proposes a nonlinear weighted least squares (NWLS) estimator; Mullahy (1997a) and Windmeijer and Santos-Silva (1997) use Generalized Method of Moments (GMM) to estimate such a model; and Kenkel and Terza (2001) analyze the endogeneity bias using Box-Cox transformation.[2]

[1] This chapter is the modified version of Masuhara (2008).

[2] From a Bayesian point of view, Kozumi (2002), Jochmann (2003), Munkin and Trivedi (2003), and Deb *et al.* (2006b) analyze the endogeneity of count data.

Moreover, Romeu and Vera-Hernández (2005) develop another count data model with an endogenous binary variable on the basis of the polynomial Poisson model proposed by Cameron and Johannson (1997). The main feature of their model is that it comprises a semiparametric model using a polynomial expansion by a dependent variable. However, the binary endogenous variable part is parametric, and the dependent variable does not explicitly assume heterogeneity.

This chapter proposes another semiparametric model to estimate a count data variable with an endogenous binary variable. This chapter considers a simple Poisson model, which has one endogenous binary variable, and the heterogeneity of both count dependent and binary variables. In this model setup, we propose a Poisson model that comprises a semiparametric (semi-nonparametric) joint distribution using Hermite polynomials based on the discussion of Gallant and Nychka (1987), Gabler *et al.* (1993), and van der Klaauw and Koning (2003). Our model is semiparametric and includes the natural extension of a bivariate normal distribution. That is, both the count dependent and endogenous binary variables explicitly assume semiparametric heterogeneity. We investigate the difference between the endogenous binary variable's coefficients of the parametric and semiparametric models using the 1990 National Health Interview Survey (NHIS) data employed by Kenkel and Terza (2001).

The rest of the chapter is organized as follows. Section 2 proposes a semiparametric count data model with an endogenous binary variable and discusses an efficient maximization algorithm that contains a numerical integral. Section 3 depicts the application of the NHIS data, and Section 4 presents our concluding remarks.

5.2 Poisson Estimation with an Endogenous Binary Variable

We consider a count data model with an endogenous binary variable proposed by Terza (1998). Let y_i, $i = 1, \ldots, N$, denote a count dependent variable that takes a nonnegative integer value; let $\mathbf{x}_i$ and $\mathbf{z}_i$ denote explanatory variables (covariates), where $\mathbf{x}_i$ is a $k_1 \times 1$ vector and $\mathbf{z}_i$ is a $k_2 \times 1$ vector. The marginal distribution of y_i takes the following form:

$$f\left(y_i \mid d_i, \varepsilon_{1i}\right) = \frac{\exp\left(-\lambda_i\right)\left(\lambda_i\right)^{y_i}}{y_i!}, \qquad \lambda_i = \exp\left(\beta_d d_i + \mathbf{x}_i'\boldsymbol{\beta}_1 + \varepsilon_{1i}\right), \tag{5.1}$$

where $\boldsymbol{\beta}_1$ and β_d denote vectors of unknown parameters, and ε_{1i} is unobserved heterogeneity. Moreover, d_i represents an endogenous binary variable and is assumed to be generated by the process $d_i = 1$ if $d_i^* = \mathbf{z}_i'\boldsymbol{\beta}_2 + \varepsilon_{2i} \geq 0$ and $d_i = 0$; otherwise, where d_i^* is a latent variable, ε_{2i} is unobserved heterogeneity, and $\boldsymbol{\beta}_2$ denotes a vector of parameters.

Many studies assume that the vector $(\varepsilon_{1i}, \varepsilon_{2i})$ follows a bivariate normal distribution with zero mean and covariance matrix $\left(\sigma^2, \rho\sigma, 1\right)$. In this assumption, the joint density is easily evaluated using a numerical integral. However, this normally distributed assumption leads to a specification problem. Under a linear-exponential mean specification assumption and a set of instruments, Mullahy (1997a) shows that the GMM estimators have consistency. In the GMM, to improve the efficiency of the estimators, it is necessary to use higher order moment conditions. The NWLS proposed in Terza (1998), which requires some additional distributional assumptions, has the same properties. Therefore, we require an alternative robust method for this count data model with an endogenous binary regressor.

Semiparametric estimation of this model implies approximating an unknown error term using Hermite polynomials (Gallant and Nychka, 1987; Gabler *et al.*, 1993). Following van der Klaauw and Koning (2003), the joint distribution of ε_{1i} and ε_{2i} takes the following semiparametric (semi-nonparametric) bivariate normal density:

$$\begin{aligned} g\left(\varepsilon_{1i}, \varepsilon_{2i}\right) = {} & \frac{1}{P}\left(\sum_{j=0}^{K}\sum_{k=0}^{K}\alpha_{jk}\varepsilon_{1i}^{j}\varepsilon_{2i}^{k}\right)^2 \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \\ & \times \exp\left[-\frac{1}{2\left(1-\rho^2\right)}\left\{\left(\frac{\varepsilon_{1i}}{\sigma_1}\right)^2 - 2\rho\frac{\varepsilon_{1i}}{\sigma_1}\frac{\varepsilon_{2i}}{\sigma_2} + \left(\frac{\varepsilon_{2i}}{\sigma_2}\right)^2\right\}\right] \equiv \frac{g^*}{P}, \end{aligned} \tag{5.2}$$

where $P = \iint_{-\infty}^{\infty} g^* \mathrm{d}\,\varepsilon_{1i}\,\mathrm{d}\,\varepsilon_{2i}$ ensures integration to 1 by scaling the density, σ_1 and ρ are standard deviation and correlation parameters, respectively, and α_{jk} is the parameter to be estimated. To identify the parameters, we set $\alpha_{00} = 1$ and $\sigma_2 = 1$. When $\alpha_{jk} = 0$ ($\forall j \geq 1$ and $\forall k \geq 1$), this density results in a bivariate normal distribution.

Hence, the log-likelihood function of a Poisson model with a semiparametric bivariate normal density takes the following form:

$$\begin{aligned} \ln f_i = {} & (1-d_i)\ln\left[\int_{-\infty}^{-\mathbf{z}_i'\boldsymbol{\beta}_2}\int_{-\infty}^{\infty} f\left(y_i \mid d_i, \varepsilon_{1i}\right) g\left(\varepsilon_{1i}, \varepsilon_{2i}\right) \mathrm{d}\,\varepsilon_{1i}\,\mathrm{d}\,\varepsilon_{2i}\right] \\ & + d_i \ln\left[\int_{-\mathbf{z}_i'\boldsymbol{\beta}_2}^{\infty}\int_{-\infty}^{\infty} f\left(y_i \mid d_i, \varepsilon_{1i}\right) g\left(\varepsilon_{1i}, \varepsilon_{2i}\right) \mathrm{d}\,\varepsilon_{1i}\,\mathrm{d}\,\varepsilon_{2i}\right]. \end{aligned} \tag{5.3}$$

Substituting (5.1) and (5.2) into (5.3) yields the full information maximum likelihood (FIML) of the semiparametric Poisson model with an endogenous dummy variable.[3] This model generalizes heterogeneity and contains the FIML model with a bivariate normal distribution as a special case.

The model in (5.3) includes double integrals and has no analytical solution. Fortunately,

[3] This model has another restriction of $\mathrm{E}\left[\varepsilon_{1i}\right] = \mathrm{E}\left[\varepsilon_{2i}\right] = 0$ (location normalization). However, this restriction is cumbersome when $K \geq 2$. Following Melenberg and van Soest (1996), we use an alternative restriction, setting the constant terms equal to those in the parametric model.

we simplify the double integrals to the following single integral:

$$\ln f_i = \ln \left[\int_{-\infty}^{\infty} f\left(y_i \mid \varepsilon_{1i}\right) \frac{G_2\left(\varepsilon_{1i}\right)}{P} g_1\left(\varepsilon_{1i}\right) \mathrm{d}\,\varepsilon_{1i} \right], \tag{5.4}$$

where g_1 is the probability density function of a normal distribution. The term G_2 contains Hermite series and depends only on ε_1, which takes the following form:

$$G_2\left(\varepsilon_{1i}\right) = \begin{cases} \int_{-\infty}^{-\mathbf{z}_i'\boldsymbol{\beta}_2} g_2\left(\varepsilon_{2i} \mid \varepsilon_{1i}\right) \mathrm{d}\,\varepsilon_{2i} & \text{if} \quad d_i = 0 \\ \int_{-\mathbf{z}_i'\boldsymbol{\beta}_2}^{\infty} g_2\left(\varepsilon_{2i} \mid \varepsilon_{1i}\right) \mathrm{d}\,\varepsilon_{2i} & \text{if} \quad d_i = 1 \end{cases}. \tag{5.5}$$

After some algebraic computation, (5.5) has an analytical solution.[4]

Since (5.4) has a single integral over $[-\infty, \infty]$, the Gauss-Hermite quadrature method is applied to evaluate the log-likelihood. However, Rabe-Hesketh *et al.* (2002, 2005) demonstrate the results of Monte Carlo simulation and conclude that the log-likelihood function approximated by this method often has a sharp peak and is poorly approximated by a low-degree polynomial. Moreover, they propose the *adaptive* Gaussian quadrature based on importance sampling and the Bayesian Markov chain method.[5] Following Rabe-Hesketh *et al.* (2002, 2005), this chapter applies the adaptive Gaussian quadrature to estimate the proposed model.[6]

Let the parameter vector of this density and the mean and variance of the posterior density be $\boldsymbol{\theta} = [\boldsymbol{\beta}_1', \boldsymbol{\beta}_2', \sigma_1, \rho, \alpha_{jk}, \ldots]'$, μ_i, and τ_i, respectively. Recall that (5.4) can be rewritten as follows:

$$\ln f_i = \ln \left[\sum_{q=1}^{Q} \omega_q f\left(y_i \mid \boldsymbol{\theta}, u_q\right) \frac{G_2\left(u_q \mid \boldsymbol{\theta}\right)}{P} \frac{g_1\left(u_q \mid \boldsymbol{\theta}\right)}{h\left(u_q \mid \mu, \tau\right)} \frac{1}{\sqrt{\pi}} \right] \equiv \ln \left[\sum_{q=1}^{Q} \omega_q f_i\left(\boldsymbol{\theta} \mid u_q\right) \right],$$

where ω_q is the qth weight, u_q is the qth evaluation point of the Gauss-Hermite quadrature over $[-\infty, \infty]$, Q is the number of weights, and $h(\cdot)$ is the importance function of a normal distribution with mean μ_i and variance τ_i. Further, the adaptive Gaussian quadrature obtains the parameters as follows:

1) Set the initial parameters $\boldsymbol{\theta}^{(t)}$, $\mu_i^{(t)}$, $\tau_i^{(t)}$, and $t \leftarrow 0$.

[4] See the Appendix for further detail.

[5] Using the adaptive Gaussian quadrature, Miranda and Rabe-Hesketh (2006) propose the stata program (ssm.ado) of the parametric binary, ordinary, and Poisson models with an endogenous binary variable.

[6] Prior to investigating the proposed model, we estimate a parametric Poisson model with an endogenous binary variable using both the Gauss-Hermite and adaptive Gaussian quadratures. The value of the log-likelihood under the latter is higher than that under the former. Moreover, the difference between the endogenous binary variable's coefficients is not negligible (See Table 5.3 and Footnote 7).

2) Calculate the following posterior density based on $\mu_{i,T-1}^{(t)}$ and $\tau_{i,T-1}^{(t)}$ until convergence:

$$\mu_{i,T}^{(t)} = \frac{\sum_{q=1}^{Q} \left(\mu_{i,T-1}^{(t)} + \sqrt{2}\tau_{i,T-1}^{(t)} u_q\right) \omega_q f_i \left(\boldsymbol{\theta}^{(t)} \mid u_q\right)}{f_i \left(\boldsymbol{\theta}^{(t)}\right)},$$

$$\tau_{i,T}^{(t)} = \sqrt{\frac{\sum_{q=1}^{Q} \left(\mu_{i,T-1}^{(t)} + \sqrt{2}\tau_{i,T-1}^{(t)} u_q\right)^2 \omega_q f_i \left(\boldsymbol{\theta}^{(t)} \mid u_q\right)}{f_i \left(\boldsymbol{\theta}^{(t)}\right)} - \left(\mu_{i,T}^{(t)}\right)^2},$$

where $f_i\left(\boldsymbol{\theta}^{(t)}\right) = \sum_{q=1}^{Q} \omega_q f_i \left(\boldsymbol{\theta}^{(t)} \mid u_q\right)$ and T is the number of iterations in this step.

3) Maximize the log-likelihood function with respect to $\boldsymbol{\theta}^{(t)}$ given $\mu_i^{(t)}$ and $\tau_i^{(t)}$.

4) Set $t \leftarrow t + 1$. Repeat steps 2 to 3 until convergence.

5.3 An Application to Drinking Behavior

We present the results of the simplified application of the proposed model using a subsample of 2,467 observations from the 1990 National Health Interview Survey (NHIS) data, originally employed by Kenkel and Terza (2001).All observations comprise males and current drinkers with high blood pressure. The dependent variable is the number of alcoholic beverages consumed in the last two weeks (D). The mean of this variable is 14.70 (21% of the observations are zero observations), and the minimum and maximum values of this variable are 0 and 168, respectively. Moreover, 687 of the individuals have been advised by a physician to reduce drinking (ADVICE). The explanatory variables are as follows: monthly income (EDITINC), years of schooling (EDUC), a dummy for $30 <$ age ≤ 40 (AGE30), $40 <$ age ≤ 50 (AGE40), $50 <$ age ≤ 60 (AGE50), $60 <$ age ≤ 70 (AGE60), age > 70 (AGEGT70), black (BLACK), non-white and non-black (OTHER), married (MARRIED), widowed (WIDOW), divorced or separated (DIVSEP), employed (EMPLOYED), unemployed (UNEMPLOY), northeastern residents (NORTHE), midwestern residents (MIDWEST), south resident (SOUTH), medicare status (MEDICARE), public insurance status (MEDICAID), military insurance status (CHAMPUS), health insurance status (HLTHINS), regional source of care (REGMED), consulting the same doctor (DRI), limits on major daily activity (MAIORLIM), limits on some daily activity (SOMELIM), having diabetes (HVDIAB), having a heart condition (HHRTCOND), and having had stroke (HADSTROKE). The entire description of the variables and summary statistics is obtained by Table 5.1.

Table 5.2 shows the estimated result of the selection equation and Table 5.3 shows that of the drinking equation of the parametric, $K = 1$, and $K = 2$ models. Since the semiparametric models nest the parametric model as a special case, we apply the log-likelihood ratio (LR) test to select the best model. The test statistics of normality against the semiparametric models with $K = 1$ and $K = 2$ equal 9.408 and 226.890, respectively. This implies that we must

Table 5.1: Drinking Behavior: Variable Description

Variable	Definition	Mean	Std. Dev.	Min.	Max.
Dependent variable					
D	Total drinks	14.698	22.753	0	168
ADVICE	Drinking advice	0.279	0.448	0	1
Socioeconomic variables ($\mathbf{x}$ and $\mathbf{z}$)					
EDITINC	Monthly income (\$1000)	2.575	5.008	-1	101
AGE30	$30 <$ age ≤ 40	0.180	0.384	0	1
AGE40	$40 <$ age ≤ 50	0.195	0.397	0	1
AGE50	$50 <$ age ≤ 60	0.182	0.386	0	1
AGE60	$60 <$ age ≤ 70	0.199	0.399	0	1
AGEGT70	$70 <$ age	0.122	0.327	0	1
EDUC	Years of schooling	12.925	3.087	0	18
BLACK	Black d.v.	0.133	0.340	0	1
OTHER	Non-white, non-black	0.018	0.132	0	1
MARRIED	Married	0.645	0.479	0	1
WIDOW	Widowed	0.052	0.223	0	1
DIVSEP	Divorced or separated	0.160	0.367	0	1
EMPLOYED	Employed	0.666	0.472	0	1
UNEMPLOY	Unemployed	0.029	0.168	0	1
NORTHE	Northeast	0.217	0.413	0	1
MIDWEST	Midwest	0.275	0.447	0	1
SOUTH	South	0.295	0.456	0	1
MEDICARE	Insurance through Medicare	0.252	0.434	0	1
MEDICAID	Insurance through Medicaid	0.031	0.174	0	1
CHAMPUS	Military insurance	0.059	0.236	0	1
HLTHINS	Health insurance	0.815	0.389	0	1
REGMED	Reg. source of care	0.821	0.384	0	1
DRI	See same doctor	0.721	0.449	0	1
MAIORLIM	Limits on major daily activ.	0.086	0.280	0	1
SOMELIM	Limits on some daily activ.	0.077	0.266	0	1
HVDIAB	Have diabetes	0.061	0.239	0	1
HHRTCOND	Have heart condition	0.146	0.353	0	1
HADSTROKE	Had stroke	0.036	0.186	0	1

Data: NHIS 1990. The data are downloadable from the Journal of Applied Econometrics Data Archive (http://econ.queensu.ca/jae/).

reject the hypothesis that heterogeneity follows a bivariate normal distribution. Moreover, the test statistic of $K = 1$ against the semiparametric model with $K = 2$ equals 217.482. Hence, the semiparametric model with $K = 2$ is the best of the three models.

Table 5.2: Estimates of the Selection Equation

	parametric		semiparametric			
			$K=1$		$K=2$	
EDITINC	−0.001	(0.005)	0.000	(0.005)	0.001	(0.007)
AGE30	0.206	(0.107)	0.141	(0.124)	0.237	(0.121)
AGE40	0.104	(0.109)	0.050	(0.118)	0.087	(0.119)
AGE50	0.061	(0.112)	0.003	(0.119)	0.044	(0.122)
AGE60	0.051	(0.123)	−0.068	(0.133)	0.015	(0.135)
AGEGT70	0.115	(0.151)	−0.088	(0.164)	0.019	(0.168)
EDUC	−0.028	(0.010)	−0.065	(0.026)	−0.032	(0.012)
BLACK	0.299	(0.080)	0.253	(0.126)	0.350	(0.106)
OTHER	0.262	(0.215)	0.174	(0.235)	0.327	(0.230)
MARRIED	0.147	(0.089)	0.028	(0.096)	0.122	(0.097)
WIDOW	0.244	(0.142)	0.155	(0.163)	0.286	(0.161)
DIVSEP	0.294	(0.105)	0.166	(0.128)	0.249	(0.119)
EMPLOYED	−0.005	(0.082)	−0.146	(0.103)	−0.049	(0.089)
UNEMPLOY	0.220	(0.176)	0.017	(0.187)	0.124	(0.194)
NORTHE	0.062	(0.083)	−0.033	(0.089)	0.102	(0.093)
MIDWEST	−0.052	(0.079)	−0.155	(0.100)	−0.041	(0.086)
SOUTH	−0.046	(0.079)	−0.155	(0.100)	−0.038	(0.086)
MEDICARE	−0.023	(0.081)	−0.053	(0.091)	0.001	(0.094)
MEDICAID	0.039	(0.113)	0.000	(0.125)	0.013	(0.129)
CHAMPUS	0.017	(0.082)	0.024	(0.090)	−0.034	(0.092)
HLTHINS	−0.142	(0.060)	−0.166	(0.094)	−0.179	(0.076)
REGMED	0.126	(0.090)	0.115	(0.103)	0.258	(0.107)
DRI	0.032	(0.081)	0.038	(0.088)	−0.051	(0.088)
MAJORLIM	0.148	(0.083)	0.126	(0.107)	0.072	(0.102)
SOMELIM	0.033	(0.081)	0.023	(0.087)	0.031	(0.093)
HVDIAB	0.302	(0.087)	0.337	(0.162)	0.344	(0.121)
HHRTCOND	0.183	(0.063)	0.204	(0.102)	0.181	(0.079)
HADSTROK	0.085	(0.128)	0.091	(0.145)	0.206	(0.158)
CONSTANT	−0.583	(0.182)	−0.583	−−	−0.583	−−

Note: Standard errors are in parentheses.

In Tables 5.2 and 5.3, we find certain features of the estimated parameters. First, both the estimated parameters and standard errors of the three models, except for the endogenous binary variable's coefficients, closely resemble each other. Second, the parameter values of the endogenous variable (ADVICE) are statistically significant at the 1% level; however, the values differ among the three models: −2.291 in the parametric model, −1.979 in the semiparametric

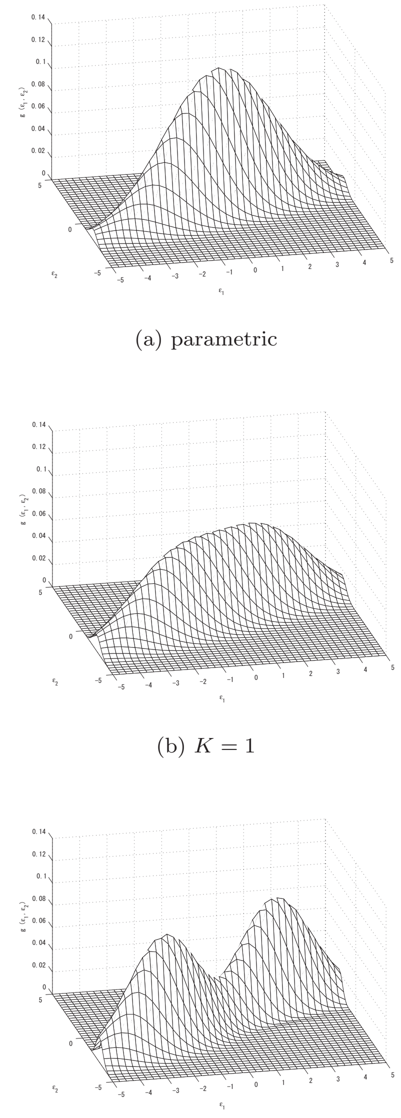
Table 5.3: Estimates of the Drinking Equation

	parametric		semiparametric			
			$K = 1$		$K = 2$	
ADVICE	−2.291	(0.248)	−1.979	(0.341)	−1.566	(0.213)
EDITINC	0.010	(0.011)	0.013	(0.013)	0.005	(0.012)
AGE30	0.153	(0.193)	0.010	(0.189)	0.142	(0.157)
AGE40	−0.075	(0.194)	−0.173	(0.188)	−0.005	(0.156)
AGE50	−0.243	(0.194)	−0.330	(0.188)	−0.101	(0.157)
AGE60	−0.201	(0.204)	−0.420	(0.197)	−0.066	(0.164)
AGEGT70	−0.285	(0.238)	−0.661	(0.227)	−0.280	(0.188)
EDUC	−0.027	(0.017)	−0.084	(0.019)	−0.041	(0.014)
BLACK	0.048	(0.146)	−0.102	(0.139)	−0.097	(0.118)
OTHER	−0.233	(0.400)	−0.441	(0.365)	−0.379	(0.296)
MARRIED	0.012	(0.154)	−0.207	(0.151)	−0.071	(0.125)
WIDOW	0.329	(0.245)	0.141	(0.238)	0.172	(0.200)
DIVSEP	0.403	(0.187)	0.135	(0.181)	0.335	(0.150)
EMPLOYED	0.084	(0.131)	−0.126	(0.128)	−0.035	(0.105)
UNEMPLOY	0.729	(0.304)	0.353	(0.288)	0.424	(0.241)
NORTHE	−0.063	(0.148)	−0.231	(0.142)	−0.077	(0.120)
MIDWEST	−0.272	(0.140)	−0.429	(0.137)	−0.212	(0.114)
SOUTH	−0.238	(0.139)	−0.403	(0.135)	−0.170	(0.113)
CONSTANT	2.584	(0.318)	2.584	–	2.584	–
σ_1	2.199	(0.094)	1.843	(0.237)	1.730	(0.398)
ρ	0.835	(0.039)	0.755	(0.101)	0.784	(0.097)
α_{01}			0.326	(2.125)	−1.206	(1.205)
α_{02}					−0.643	(0.437)
α_{10}			0.051	(0.956)	0.717	(0.735)
α_{11}			0.156	(0.090)	0.805	(0.166)
α_{12}					0.183	(0.238)
α_{20}					−0.220	(0.115)
α_{21}					−0.061	(0.130)
α_{22}					−0.018	(0.011)
log-likelihood	−10,202.043		−10,197.339		−10,088.598	

Note: Standard errors are in parentheses.

model with $K = 1$, and −1.566 in the semiparametric model with $K = 2$.[7] This means that

[7] Using the Gauss-Hermite quadrature, the log-likelihood values of the parametric, $K = 1$, and $K = 2$ models are −10,732.920, −10,460.757, and −10,271.797, respectively. The coefficient values of the endogenous binary variable (ADVICE) are −1.235, −1.038, and −0.819, respectively. Moreover, the advice effects of the three models are −70.9%, −64.6%, and −55.9%, respectively.



(a) parametric

(b) $K = 1$

(c) $K = 2$

Figure 5.1: Estimated Densities of Heterogeneity

advice appears to reduce the consumption of alcoholic beverages by $[\exp(-2.291) - 1] \times 100 = -89.9\%$ in the parametric model, $[\exp(-1.979) - 1] \times 100 = -86.2\%$ in the semiparametric model with $K = 1$, and $[\exp(-1.566) - 1] \times 100 = -79.1\%$ in the semiparametric model with $K = 2$. Compared to the result of the parametric model, the reduction of alcoholic beverages in the two semiparametric models is small. Based on the LR test and estimated results, the influence of the doctor's advice has a negative effect on drinking behavior; however, it can be overestimated by the parametric model.

Figure 5.1 graphs the estimated densities of the three models using the 10% significant coefficients. We find that the semiparametric (semi-nonparametric) model with $K = 1$ has a fatter tail than the normal density (parametric) model; moreover, the semiparametric model with $K = 2$ is a twin-peak distribution.

5.4 Conclusion

This chapter proposes a new semiparametric count data estimation with an endogenous binary variable that generalizes bivariate correlated unobserved heterogeneity using Hermite polynomials. In an example using the 1990 NHIS data, the semiparametric model with $K = 2$ overcomes the other models in terms of the LR test. The absolute values of the endogenous binary regressor coefficients of the semiparametric models are smaller than that of the parametric model, and that of the semiparametric model with $K = 2$ is the smallest of the three. This introduces the interpretation of the binary endogenous variable, that is, the effect of the advice variable. The parametric model overestimates the effect of doctor's advice in our example. Moreover, the estimated densities of the semiparametric models have fatter tail than that of the parametric model.

One major advantage of the semiparametric model is the flexibility of bivariate distributed heterogeneity. The difference between the endogenous binary variable's coefficients of the parametric and semiparametric models is not negligible in our example. Therefore, it is useful to generalize bivariate heterogeneity using Hermite polynomials.

Appendix

Following van der Klaauw and Koning (2003), we specify the bivariate semiparametric (semi-nonparametric) normal density as follows:

$$\begin{aligned} g\left(\varepsilon_{1i}, \varepsilon_{2i}\right) =& \frac{1}{P}\left(\sum_{j=0}^{K}\sum_{k=0}^{K}\alpha_{jk}\varepsilon_{1i}^{j}\varepsilon_{2i}^{k}\right)^{2} \times \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \\ & \times \exp\left[-\frac{1}{2\left(1-\rho^2\right)}\left\{\left(\frac{\varepsilon_{1i}}{\sigma_1}\right)^2 - 2\rho\frac{\varepsilon_{1i}}{\sigma_1}\frac{\varepsilon_{2i}}{\sigma_2} + \left(\frac{\varepsilon_{2i}}{\sigma_2}\right)^2\right\}\right], \end{aligned} \tag{A5.1}$$

where

$$\begin{aligned} P =& \iint_{-\infty}^{\infty}\left(\sum_{j=0}^{K}\sum_{k=0}^{K}\alpha_{jk}\varepsilon_{1i}^{j}\varepsilon_{2i}^{k}\right)^{2} \times \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \\ & \times \exp\left[-\frac{1}{2\left(1-\rho^2\right)}\left\{\left(\frac{\varepsilon_{1i}}{\sigma_1}\right)^2 - 2\rho\frac{\varepsilon_{1i}}{\sigma_1}\frac{\varepsilon_{2i}}{\sigma_2} + \left(\frac{\varepsilon_{2i}}{\sigma_2}\right)^2\right\}\right] \mathrm{d}\,\varepsilon_{1i}\,\mathrm{d}\,\varepsilon_{2i}. \end{aligned}$$

We now consider the case of $K = 2$. The round bracket of (A5.1) can be rearranged as

$$\left(\sum_{j=0}^{2}\sum_{k=0}^{2}\alpha_{jk}\varepsilon_{1i}^{j}\varepsilon_{2i}^{k}\right)^2 = \gamma_0 + \gamma_1\varepsilon_{2i} + \gamma_2\varepsilon_{2i}^2 + \gamma_3\varepsilon_{2i}^3 + \gamma_4\varepsilon_{2i}^4,$$

where

$$\begin{aligned}
\gamma_0 &= \left(\alpha_{00} + \alpha_{10}\varepsilon_{1i} + \alpha_{20}\varepsilon_{1i}^2\right)^2, \\
\gamma_1 &= 2\left(\alpha_{00} + \alpha_{10}\varepsilon_{1i} + \alpha_{20}\varepsilon_{1i}^2\right)\left(\alpha_{01} + \alpha_{11}\varepsilon_{1i} + \alpha_{21}\varepsilon_{1i}^2\right), \\
\gamma_2 &= 2\left(\alpha_{00} + \alpha_{10}\varepsilon_{1i} + \alpha_{20}\varepsilon_{1i}^2\right)\left(\alpha_{02} + \alpha_{12}\varepsilon_{1i} + \alpha_{22}\varepsilon_{1i}^2\right) + \left(\alpha_{01} + \alpha_{11}\varepsilon_{1i} + \alpha_{21}\varepsilon_{1i}^2\right)^2, \\
\gamma_3 &= 2\left(\alpha_{01} + \alpha_{11}\varepsilon_{1i} + \alpha_{21}\varepsilon_{1i}^2\right)\left(\alpha_{02} + \alpha_{12}\varepsilon_{1i} + \alpha_{22}\varepsilon_{1i}^2\right), \\
\gamma_4 &= \left(\alpha_{02} + \alpha_{12}\varepsilon_{1i} + \alpha_{22}\varepsilon_{1i}^2\right)^2.
\end{aligned}$$

We require the following algebraic computation to obtain (5.4) or P:

$$\begin{aligned}
&\int_{-\infty}^{\infty}\int_{\hat{a}}^{\hat{b}} f\left(y_i \mid d_i, \varepsilon_{1i}\right) g\left(\varepsilon_{1i}, \varepsilon_{2i}\right) \mathrm{d}\,\varepsilon_{2i}\,\mathrm{d}\,\varepsilon_{1i} \\
&\qquad = \int_{-\infty}^{\infty} f\left(y_i \mid d_i, \varepsilon_{1i}\right) \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left(-\frac{1}{2}\left(\frac{\varepsilon_{1i}}{\sigma_1}\right)^2\right) \\
&\qquad\quad \times \frac{1}{P}\int_{\hat{a}}^{\hat{b}} \sum_{j=0}^{4}\gamma_j\varepsilon_2^j \frac{1}{\sqrt{2\pi}\sigma_2\sqrt{1-\rho^2}} \\
&\qquad\quad \times \exp\left[-\frac{1}{\left(\sigma_2\sqrt{2\left(1-\rho^2\right)}\right)^2}\left(\varepsilon_{2i} - \rho\frac{\sigma_2}{\sigma_1}\varepsilon_{1i}\right)^2\right] \mathrm{d}\,\varepsilon_{2i}\,\mathrm{d}\,\varepsilon_{1i} \\
&\qquad \equiv \int_{-\infty}^{\infty} f\left(y_i \mid d_i, \varepsilon_{1i}\right) g_1\left(\varepsilon_{1i}\right) \int_{\hat{a}}^{\hat{b}} \frac{g_2\left(\varepsilon_{2i} \mid \varepsilon_{1i}\right)}{P} \mathrm{d}\,\varepsilon_{2i}\,\mathrm{d}\,\varepsilon_{1i} \\
&\qquad \equiv \int_{-\infty}^{\infty} f\left(y_i \mid d_i, \varepsilon_{1i}\right) g_1\left(\varepsilon_{1i}\right) \frac{G_2\left(\varepsilon_{1i}\right)}{P} \mathrm{d}\,\varepsilon_{1i}, \qquad \text{(A5.2)}
\end{aligned}$$

where $g_1(\cdot)$ is the probability density function of a normal distribution. When $\hat{a} = -\infty$ and $\hat{b} = \infty$, (A5.2) results in P. Substituting $\xi = \rho\sigma_2\varepsilon_{1i}/\sigma_1$, $u = \varepsilon_{2i} - \xi$ and $\delta = \sigma_2\sqrt{2\left(1-\rho^2\right)}$ into (A5.2) yields

$$\begin{aligned}
\frac{G_2}{P} &= \frac{1}{\sqrt{\pi}\delta}\frac{1}{P}\int_{\hat{a}-\xi}^{\hat{b}-\xi}\sum_{j=0}^{4}\gamma_j\left(u+\xi\right)^j \exp\left[-\left(\frac{u}{\delta}\right)^2\right]\mathrm{d}\,u \\
&\equiv \frac{1}{\sqrt{\pi}\delta}\frac{1}{P}\int_{a}^{b}\sum_{j=0}^{4}\eta_j u^j \exp\left[-\left(\frac{u}{\delta}\right)^2\right]\mathrm{d}\,u,
\end{aligned}$$

where $a = \hat{a} - \xi$, $b = \hat{b} - \xi$, and

$$\begin{aligned}
\eta_0 &= \gamma_4\xi^4 + \gamma_3\xi^3 + \gamma_2\xi^2 + \gamma_1\xi + \gamma_0, & \eta_1 &= 4\gamma_4\xi^3 + 3\gamma_3\xi^2 + 2\gamma_2\xi + \gamma_1, \\
\eta_2 &= 6\gamma_4\xi^2 + 3\gamma_3\xi + \gamma_2, & \eta_3 &= 4\gamma_4\xi + \gamma_3, \\
\eta_4 &= \gamma_4.
\end{aligned}$$

Therefore, using the following recursion formula (van der Klaauw and Koning, 2003):

$$\begin{aligned}I_k\left(a,b\right) &= \int_a^b u^k \exp\left(-\frac{u^2}{\delta^2}\right) \mathrm{d}\, u \\ &= \frac{\delta^2}{2}\left[a^{k-1}\exp\left(-\frac{a^2}{\delta^2}\right) - b^{k-1}\exp\left(-\frac{b^2}{\delta^2}\right)\right] + \frac{\left(k-1\right)\delta^2}{2} I_{k-2}\left(a,b\right)\end{aligned}$$

and substituting $b = -z_i'\boldsymbol{\beta}_2 - \xi$, we obtain the following relation:

$$\begin{aligned}\frac{G_2\left(\varepsilon_{1i} \mid d_i = 0\right)}{P} =& \frac{1}{\sqrt{\pi}\delta}\frac{1}{P}\left[\eta_0 I_0\left(-\infty,b\right) + \eta_1 I_1\left(-\infty,b\right) + \eta_2 I_2\left(-\infty,b\right)\right. \\ & \left. + \eta_3 I_3\left(-\infty,b\right) + \eta_4 I_4\left(-\infty,b\right)\right].\end{aligned}$$

Using the same procedure, we can calculate the term P.

Chapter 6

Semiparametric Duration Analysis with an Endogenous Binary Variable: An Application to Hospital Stays

6.1 Introduction

Recently, duration (survival) analysis has been widely used in applied econometrics. Moreover, we find situations where covariates (especially an endogenous binary variable) are simultaneously determined along with the duration variable. As is the case with many nonlinear models, the endogeneity problem in duration analysis is cumbersome because the existence of censored duration data leads to nonlinearity, leading the two-stage method to become inconsistent (Wooldridge, 2002, p.478). Some studies have been conducted to analyze the endogeneity problem in duration analysis. Bijwaard and Ridder (2005) propose a two-stage instrumental variable estimator for duration data based on the generalized accelerated failure model that contains the proportional hazard model as a special case. However, the models based on a hazard rate do not explicitly assume heterogeneity. In applied econometrics, the possibility of omitted variables is inevitable and controling population heterogeneity alone is inadequate. Therefore, in duration analysis, it is important to consider both heterogeneity and endogeneity.

This chapter proposes an alternative semiparametric duration model with an endogenous binary variable that generalizes the heterogeneity of both duration and endogeneity. The generalization of heterogeneity is done as follows: first, we consider a simple log-normal duration model with an endogenous binary variable; next, we assume heterogeneity that follows a semiparametric bivariate distribution using Hermite polynomials based on van der Klaauw and Koning (2003). Under these setups, we investigate the difference between the endogenous binary variable's coefficients of the parametric and semiparametric models using the Medical Expenditure Panel Survey (MEPS) data employed by Prieger (2002).

Section 2 proposes a semiparametric duration model with an endogenous binary variable and censored data. Section 3 depicts the application of the length of hospitalizations, and Section 4 presents our concluding remarks.

6.2 Semiparametric Duration Analysis with an Endogenous Binary Variable

We consider a log-normal model in duration analysis based on Masuhara (2007): $\ln t_i = \beta_d d_i + \mathbf{x}_i' \boldsymbol{\beta}_1 + \varepsilon_{1i}$, where t_i, $i = 1, \ldots, N$, is an observed duration outcome that has a continuous probability density $f(t_i)$; $\mathbf{x}_i \sim k_1 \times 1$ and $\mathbf{z}_i \sim k_2 \times 1$ denote explanatory variables (covariates); $\boldsymbol{\beta}_1$ and β_d denote vectors of unknown parameters; ε_{1i} is unobserved heterogeneity. Moreover, d_i represents a binary endogenous variable and is assumed to be generated by the process $d_i = 1$ if $d_i^* = \mathbf{z}_i' \boldsymbol{\beta}_2 + \varepsilon_{2i} \geq 0$ and $d_i = 0$ otherwise, where d_i^* is a latent variable; ε_{2i} is unobserved heterogeneity; $\boldsymbol{\beta}_2$ denotes a vector of parameters. In this model, a random variable t_i is a linear function of ε_{1i}. Therefore, we concentrate on the joint distribution of $(\varepsilon_{1i}, \varepsilon_{2i})$. It is natural to assume that $(\varepsilon_{1i}, \varepsilon_{2i})$ follows bivariate normal distribution with mean zero and covariance matrix $\left(\sigma_1^2, \rho\sigma_1, 1\right)$, i.e., a linear model with an endogenous binary variable. However, this normally distributed assumption leads to a specification problem. Therefore, we require a more flexible and robust estimation for this duration analysis with an endogenous binary variable.

Semiparametric estimation of this model is to approximate an unknown error term using Hermite polynomials. Following van der Klaauw and Koning (2003), the joint distribution of $(\varepsilon_{1i}, \varepsilon_{2i})$ takes the following semi-nonparametric (SNP) normal density:

$$\begin{aligned} f(\varepsilon_{1i}, \varepsilon_{2i}) &= \frac{1}{P} \left(\sum_{j=0}^{K} \sum_{k=0}^{K} \alpha_{jk} \varepsilon_{1i}^{j} \varepsilon_{2i}^{k} \right)^2 \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \\ &\times \exp\left[-\frac{1}{2(1-\rho^2)} \left\{ \left(\frac{\varepsilon_{1i}}{\sigma_1}\right)^2 - 2\rho \frac{\varepsilon_{1i}}{\sigma_1} \frac{\varepsilon_{2i}}{\sigma_2} + \left(\frac{\varepsilon_{2i}}{\sigma_2}\right)^2 \right\} \right] \equiv \frac{f^*}{P}, \end{aligned} \tag{6.1}$$

where $P = \iint_{-\infty}^{\infty} f^* \mathrm{d}\,\varepsilon_{1i}\, \mathrm{d}\,\varepsilon_{2i}$ ensures integration to 1 by scaling the density, σ_1 and ρ are standard deviation and correlation parameters, and α_{jk} are parameters to be estimated. To identify the parameters, we set $\alpha_{00} = 1$ and $\sigma_2 = 1$.[1]

This model includes double integrals but has no analytical solution. Therefore, we evaluate the likelihood using a numerical integral. Fortunately, the log-likelihood results in the following single integral:

$$\ln L = \sum_{i=1}^{N} (1 - c_i) \ln \left[\frac{\Psi(\varepsilon_{1i})}{P} \frac{1}{\sigma_1} \phi\left(\frac{\varepsilon_{1i}}{\sigma_1} \right) \right] + c_i \ln \left[\int_{\underline{\varepsilon}_{1i}}^{\infty} \frac{\Psi(\varepsilon_{1i})}{P} \frac{1}{\sigma_1} \phi\left(\frac{\varepsilon_{1i}}{\sigma_1} \right) \mathrm{d}\,\varepsilon_{1i} \right],$$

where c_i is a censoring indicator ($c_i = 1$ if the observation is censored and $c_i = 0$ if the observation is uncensored) and $\underline{\varepsilon}_{1i} \equiv \ln t_i - \beta_d d_i - \mathbf{x}_i' \boldsymbol{\beta}_1$. The term $\phi(\cdot)$ is the probability density function of the standard normal distribution; $\Psi(\cdot)$ contains a Hermite series and depends only on ε_{1i},[2] which takes the following form:

$$\Psi(\varepsilon_{1i}) = \begin{cases} \int_{-\infty}^{-z_i' \boldsymbol{\beta}_2} \psi(\varepsilon_{2i} \mid \varepsilon_{1i}) \,\mathrm{d}\,\varepsilon_{2i} & \text{if} \quad d_i = 0 \\ \int_{-z_i' \boldsymbol{\beta}_2}^{\infty} \psi(\varepsilon_{2i} \mid \varepsilon_{1i}) \,\mathrm{d}\,\varepsilon_{2i} & \text{if} \quad d_i = 1 \end{cases}. \tag{6.2}$$

After some algebraic computation, Equation (6.2) has an analytical solution.

Although we avoid double integrals, $\ln L$ remains a single integral over $[\underline{\varepsilon}_{1i}, \infty]$ in a censored part. If the integral is distributed over $[-\infty, \infty]$, the log-likelihood is easily evaluated by applying the Gauss-Hermite (GH) quadrature. In this censored part, we can calculate the log-likelihood function using $\varepsilon_s > \underline{\varepsilon}_{1i}$, where ε_s is the evaluation point of the GH quadrature. Since $\underline{\varepsilon}_{1i}$ contains the vector $\boldsymbol{\beta}_1$ (or β_d), a small change in the value of $\boldsymbol{\beta}_1$ (or β_d) does not change the likelihood, and thus, the performance of the GH quadrature is low. When we use the simulated maximum likelihood (SML) method instead of the GH quadrature, this problem still holds. It is necessary to use many evaluation points to approximate the integral accurately. Hence, it takes much time to calculate the accurate likelihood. To maximize the log-likelihood, it is not realistic to use the GH quadrature or SML.

This paper applies the GHK simulator, due to Geweke (1992), Hajivassiliou and McFadden (1994), and Keane (1994) to evaluate the log-likelihood in the censored part.[3] The GHK simulator is described as follows: (i) Generate the value of ε_{1s} from a *truncated* normal distribution at $\underline{\varepsilon}_{1i}$ as follows: (a) generate a standard uniform random variable u_s; (b) calculate $\varepsilon_{1s} = \sigma_1 \Phi^{-1}\left(\Phi\left(\underline{\varepsilon}_{1i}/\sigma_1\right) + u_s \left\{1 - \Phi\left(\underline{\varepsilon}_{1i}/\sigma_1\right)\right\}\right)$, where $\Phi(\cdot)$ is a cumulative distribution of the standard normal distribution. (ii) Calculate $\left[1 - \Phi\left(\underline{\varepsilon}_{1i}/\sigma_1\right)\right] \Psi(\varepsilon_{1s})/P$. iii) Repeat the steps 1 to 2 S times, and calculate the simulated probability: $\left[1 - \Phi\left(\underline{\varepsilon}_{1i}/\sigma_1\right)\right] \sum_{s=1}^{S} \Psi(\varepsilon_{1s}) / (P \times S)$.

[1] This model has another restriction: $E[\varepsilon_{1i}] = E[\varepsilon_{2i}] = 0$. The restriction is equivalent to setting the constant term equal to that in the parametric model.

[2] For further details, see Masuhara (2008).

[3] Train (2003) explains the simplified version of the GHK simulator and applies this simulator to mixed logit models.

Although the random variable ε_{1s} should be generated from a *censored* normal distribution, the GHK simulator generates the *truncated* normal distribution. Therefore, it is necessary to use the weight $[1 - \Phi(\underline{\varepsilon}_{1i}/\sigma_1)]$ for $\Psi(\varepsilon_{1s})/P$. Unlike the GH quadrature or SML, the GHK simulator calculates the log-likelihood on *fixed* evaluation points.

Moreover, this chapter uses Halton (1960) sequences for a standard uniform random variable u_s. The SML method requires a large number of pseudo-random draws u_s to achieve a suitable level of precision. However, it is computationally expensive to increase the number of simulation draws in order to reduce the simulation error to acceptable levels. Quasi-random numbers like the Halton sequence, which use non-random points within the domain of integration, are another method to evaluate the simulated likelihood. In general, the convergence rate for the quasi-random numbers is faster than that for the pseudo-random numbers. Bhat (2001) and Train (2003) report that the Halton sequences are more uniformly distributed than pseudo-random numbers.

Halton sequences are constructed as follows. Consider the prime number 2. Take the unit interval $(0, 1)$ and divide it into two parts. The dividing point 1/2 is the first element of the Halton sequence. Next, divide each part into two parts. The dividing points, 1/4 and 3/4, are the next two elements of the sequence. Divide each of the four parts into two parts each. The dividing points are 1/8, 5/8, 3/8, and 7/8 (which are 1/8 added to zero and the previous numbers: 0, 1/2, 1/4, and 3/4). Continue this process to obtain the Halton sequences based on the prime number 2 $(1/2, 1/4, 3/4, 1/8, 5/8, 3/8, 7/8, \ldots)$. Similar sequences are defined for other prime numbers, such as 3 $(1/3, 2/3, 1/9, 4/9, 7/9, 2/9, 5/9, 8/9, \ldots)$. In order to obtain corresponding standard normal points from each Halton draw, we take the inverse standard normal distribution transformation: $\Phi^{-1}(1/2) = 0$, $\Phi^{-1}(1/4) = -0.67$, $\Phi^{-1}(3/4) = 0.67$, $\ldots$, where Φ^{-1} is an inverse of the cumulative density function of the standard normal.

6.3 Application to Hospital Stays

We present the results of the simplified application of the model, using a subsample of 1,257 observations from the 1996 Medical Expenditure Panel Survey (MEPS), originally employed by Prieger (2002).We regard the variable length of all hospitalizations (HOSPDUR) as duration and employ the data with HOSPDUR > 0 on 1,257 out of the original 14,956 observations to concentrate on the duration analysis. The average of the annual length of all hospitalizations is 7.105 although it is strange that the minimum is 0.5 and the maximum is 99. The explanatory variables are as follows: (1) health status measures — the number of self-reported medical conditions (CONDN), the number of conditions on the priority list (PROLIST), a dummy for self-perceived excellent health (EXCLHLTH), self-perceived

Table 6.1: Hospital Stays: Variable Description

Variable	Definition	Mean	Std. Dev.	Min.	Max.
HOSPDUR	Length of all hospitalizations	7.105	10.958	0.5	99
HOSPNUM	Number of hospitals stays	1.403	0.836	1	9
PRIVINS	1 = covered by private insurance of any type	0.637	0.481	0	1
MEDICARE	1 = currently covered by Medicare	0.353	0.478	0	1
MEDICAID	1 = currently covered by Medicaid	0.177	0.382	0	1
HMO	1 = enrolled in a HMO	0.369	0.483	0	1
CONDN	Number of self-reported medical conditions	2.970	2.696	0	22
PRIOLIST	Number of conditions on the priority list	1.194	1.551	0	11
EXCLHLTH	1 = individual reports health to be 'excellent'	0.164	0.370	0	1
POORHLTH	1 = individual reports health to be 'poor'	0.121	0.326	0	1
ADLHELP	1 = requires assistance with daily living tasks	0.149	0.356	0	1
MIDWEST	Regional indicator (EAST is the excluded dummy)	0.238	0.426	0	1
SOUTH	Regional indicator (EAST is the excluded dummy)	0.363	0.481	0	1
WEST	Regional indicator (EAST is the excluded dummy)	0.203	0.402	0	1
FEMALE	1 = female	0.652	0.476	0	1
AGE	Age	51.080	20.193	18	90
BLACK	1 = black (not Hispanic)	0.126	0.332	0	1
HISPANIC	1 = of Hispanic ethnicity	0.173	0.378	0	1
EDUC	Years of education	11.691	3.318	0	17
MARRIED	Marital status: 1 = currently married	0.563	0.496	0	1
EMPLOYED	Employment status: 1 = currently employed	0.425	0.495	0	1
PRIVMCAR	1 = covered by private insurance and Medicare	0.201	0.401	0	1
INSCUR	Health insurance offered from the current main job	0.284	0.451	0	1
INSPREV	Health insurance offered through a job other than the current main job	0.219	0.414	0	1
INSURED	Insured	0.908	0.290	0	1

Data: MEPS 1996. The data are downloadable from the Journal of Applied Econometrics Data Archive (http://econ.queensu.ca/jae/).

poor health (POORHLTH), and assistance for the physical limitations in daily living (ADLHELP); (2) socioeconomic variables — exact age (AGE), years of education (EDUC), a dummy for south residents (SOUTH), midwestern residents (MIDWEST), western residents (WEST), African-Americans (BLACK), Hispanic (HISPANIC), female (FEMALE), marital status (MARRIED), employment status (EMPLOYED), health insurance offered from the current main job (INSCUR), and health insurance offered through a job other than the current main job (INSPREV). The entire description of the variables and summary statistics is obtained by Table 6.1.

Table 6.2: Estimated Results of Hospital Stays (Selection Equation)

	non-censored data				artificial censored data at $t = 30$			
	parametric		SNP ($K = 2$)		parametric		SNP ($K = 2$)	
INSCUR	1.283	(0.173)	1.525	(0.198)	1.276	(0.172)	1.403	(0.176)
INSPREV	0.328	(0.183)	0.292	(0.172)	0.313	(0.182)	0.259	(0.162)
CONDN	0.017	(0.033)	0.013	(0.037)	0.017	(0.033)	0.020	(0.033)
PRIOLIST	0.091	(0.069)	0.110	(0.072)	0.089	(0.069)	0.087	(0.067)
EXCLHLTH	0.454	(0.171)	0.535	(0.182)	0.448	(0.170)	0.480	(0.166)
POORHLTH	0.052	(0.193)	0.092	(0.200)	0.056	(0.193)	0.106	(0.186)
ADLHELP	0.376	(0.227)	0.334	(0.212)	0.382	(0.226)	0.336	(0.202)
MIDWEST	−0.438	(0.204)	−0.437	(0.204)	−0.444	(0.203)	−0.436	(0.190)
SOUTH	−0.741	(0.181)	−0.846	(0.193)	−0.743	(0.180)	−0.807	(0.176)
WEST	−0.307	(0.202)	−0.256	(0.206)	−0.311	(0.201)	−0.252	(0.191)
FEMALE	0.130	(0.131)	0.096	(0.132)	0.128	(0.130)	0.078	(0.123)
AGE	0.021	(0.004)	0.024	(0.004)	0.021	(0.004)	0.021	(0.004)
BLACK	0.050	(0.178)	0.083	(0.197)	0.048	(0.178)	0.059	(0.180)
HISPANIC	−0.188	(0.141)	−0.183	(0.158)	−0.190	(0.140)	−0.199	(0.144)
EDUC	0.040	(0.018)	0.047	(0.020)	0.040	(0.018)	0.041	(0.018)
MARRIED	−0.022	(0.118)	−0.037	(0.127)	−0.022	(0.118)	−0.057	(0.117)
EMPLOYED	−0.494	(0.139)	−0.605	(0.174)	−0.495	(0.138)	−0.567	(0.152)
CONSTANT	0.037	(0.350)	0.037	–	0.045	(0.349)	0.045	–
α_{01}			−0.148	(0.059)			−0.177	(0.039)
α_{02}			2.857	(0.055)			1.963	(0.034)
α_{10}			−0.135	(0.056)			−0.418	(0.032)
α_{11}			4.361	(0.051)			3.047	(0.029)
α_{12}			−0.894	(0.018)			−0.440	(0.011)
α_{20}			1.843	(0.047)			1.267	(0.024)
α_{21}			−0.802	(0.020)			−0.384	(0.011)
α_{22}			0.004	(0.009)			−0.010	(0.005)
log-likelihood	−2,081.017		−2,068.231		−2,076.447		−2,063.234	

Notes: SNP denotes the semi-nonparametric duration model; standard errors are in parentheses.

Many empirical works demonstrate that an individual's insurance choice is endogenous when health outcomes are considered to be a dependent variable. We are interested in how the individual's insurance choice affects the duration of hospital stays (HOSPDUR). Following Prieger (2002), this chapter uses a single insurance indicator (INSURED), which includes all types of insurance such as indemnity private insurance, medicare, medicaid, and HMO; this is done so as to avoid the difficulties involved in estimating multivariate probit models of high

Table 6.3: Estimated Results of Hospital Stays (Duration Equation)

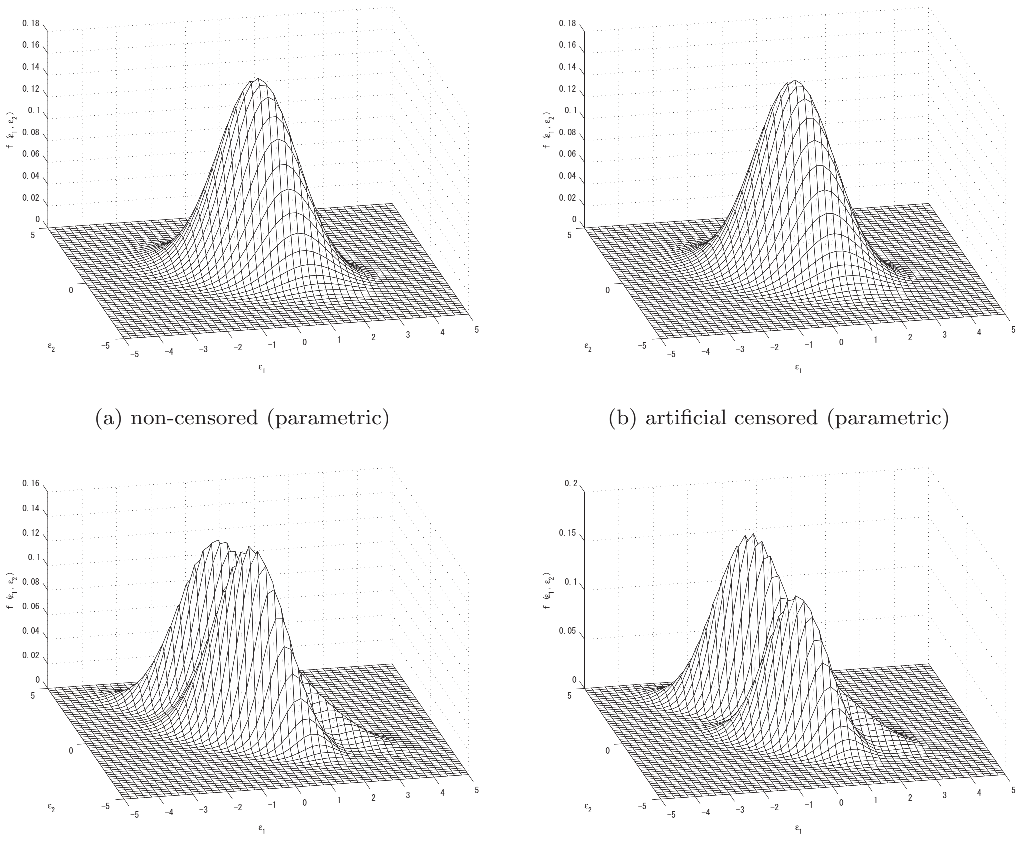
	non-censored data				artificial censored data at $t = 30$			
	parametric		SNP ($K = 2$)		parametric		SNP ($K = 2$)	
INSURED	0.676	(0.097)	0.986	(0.074)	0.713	(0.096)	1.037	(0.076)
CONDN	−0.013	(0.016)	−0.014	(0.016)	−0.015	(0.016)	−0.014	(0.015)
PRIOLIST	0.056	(0.029)	0.054	(0.028)	0.056	(0.029)	0.053	(0.028)
EXCLHLTH	−0.144	(0.081)	−0.155	(0.076)	−0.148	(0.081)	−0.162	(0.077)
POORHLTH	0.330	(0.096)	0.329	(0.093)	0.328	(0.097)	0.330	(0.093)
ADLHELP	0.315	(0.091)	0.323	(0.088)	0.335	(0.092)	0.323	(0.089)
MIDWEST	−0.041	(0.088)	−0.009	(0.084)	−0.043	(0.088)	0.005	(0.086)
SOUTH	0.068	(0.082)	0.131	(0.078)	0.068	(0.082)	0.139	(0.079)
WEST	−0.238	(0.092)	−0.185	(0.088)	−0.242	(0.092)	−0.181	(0.089)
FEMALE	−0.261	(0.063)	−0.219	(0.060)	−0.256	(0.063)	−0.204	(0.061)
AGE	0.009	(0.002)	0.008	(0.002)	0.009	(0.002)	0.008	(0.002)
BLACK	0.221	(0.091)	0.225	(0.086)	0.225	(0.092)	0.239	(0.088)
HISPANIC	0.074	(0.084)	0.099	(0.078)	0.080	(0.084)	0.121	(0.079)
EDUC	−0.018	(0.010)	−0.015	(0.009)	−0.017	(0.010)	−0.014	(0.009)
MARRIED	−0.134	(0.060)	−0.120	(0.056)	−0.133	(0.060)	−0.111	(0.057)
EMPLOYED	−0.186	(0.066)	−0.163	(0.061)	−0.191	(0.066)	−0.163	(0.062)
CONSTANT	0.657	(0.198)	0.657	–	0.617	(0.198)	0.617	–
σ_1	1.032	(0.020)	0.913	(0.013)	1.034	(0.021)	1.038	(0.016)
ρ	−0.473	(0.047)	−0.776	(0.010)	−0.491	(0.046)	−0.817	(0.010)

Notes: SNP denotes the semi-nonparametric duration model; standard errors are in parentheses.

order. Although, when analyzing duration data, censored data play an important role, our data do not have censored data. Hence, we compare the coefficients between (1) non-censored data and (2) artificial censored data at $t = 30$ (the proportion of right-censored samples is 4.14%).

Table 6.2 and 6.3 show the estimated results of parametric and SNP duration analysis with $K = 2$.[4] The parameter values of the endogenous variable (INSURED) are statistically significant at the 1% level; however, the values differ among the four models: 0.676 in the non-censored parametric model, 0.986 in the non-censored SNP model with $K = 2$, 0.713 in the censored parametric model, and 1.037 in the censored SNP model with $K = 2$. This means that the any type of insurance choice increases the length of hospital stays by 96.696% in the non-censored parametric model, 168.168% in the non-censored SNP model with $K = 2$, 104.010% in the censored parametric model, and 182.074% in the censored SNP model with

[4] Since the log-likelihood ratio tests support the SNP model with $K = 2$, we omit the results of the SNP model with $K = 1$.



(a) non-censored (parametric)

(b) artificial censored (parametric)

(c) non-censored ($K = 2$)

(d) artificial censored ($K = 2$)

Figure 6.1: Estimated Densities of Heterogeneity

$K = 2$.[5] Compared with the parametric models, the increase of hospital stays in the two SNP models is large, especially in the case of non-censored data. Although there is the difference between the INSURED of the censored and non-censored parametric models, the values of INSURED in the two SNP models resemble each other.

Figure 6.1 graphs the estimated densities of the three models using the 5% significant coefficients. We find that the semiparametric model with $K = 2$ is a twin-peak distribution and that the contour lines differ from the usual ellipsoids of the bivariate normal density.

6.4 Conclusion

This chapter proposes a new semiparametric duration model with an endogenous binary variable and censored data that generalizes bivariate correlated unobserved heterogeneity using Hermite polynomials. When applied to the duration of hospital stays of the MEPS

[5] In the case of the artificial censored data at $t = 15$, the INSURED values of the parametric and SNP models are 0.861 and 1.074, respectively.

data, the estimated results of both the non-censored and artificial censored SNP models show a good performance. The absolute values of the endogenous binary regressor coefficients of the semiparametric models are larger than that of the parametric models, if the data are censored or not. This introduces the interpretation of the binary endogenous variable, that is, the individual's insurance choice variable. The parametric model underestimates the effect of the individual's insurance choice in our example. The difference of the estimated endogenous coefficients of both the two models is smaller than those of the parametric models. This means that, if the data are censored, the parametric model have a large inconsistency. Moreover, the estimated densities of the semiparametric models have twin peak distribution.

The semiparametric model proposed in this chapter has one major advantage of the flexibility of bivariate distributed heterogeneity. When the difference between the endogenous binary variable's coefficients of the parametric and semiparametric models is not negligible, it is useful to generalize bivariate heterogeneity using Hermite polynomials.

References

Amemiya, T. (1985), *Advanced Econometrics*, Harvard University Press, Cambridge.

Alfò, M. and G. Trovato (2004), "Semiparametric Mixture Models for Multivariate Count Data, with Application," *Econometrics Journal*, 7 (2), 426–454.

Alfò, M., G. Trovato and R.J. Waldmann (2008), "Testing for Country Heterogeneity in Growth Models Using a Finite Mixture Approach," *Journal of Applied Econometrics*, 23 (4), 487–514.

Andrews, D.W.K. (1988), "Chi-Square Diagnostic Tests for Econometric Models: Introduction and Applications," *Journal of Econometrics*, 37 (1), 135–156.

Angrist, J.D. (2001), "Estimation of Limited Dependent Variable Models with Dummy Endogenous Regressors: Simple Strategies for Empirical Practice," *Journal of Business and Economic Statistics*, 19 (1), 2–16.

Arcidiacono, P. and J.B. Jones (2003), "Finite Mixture Distributions, Sequential Likelihood and the EM Algorithm," *Econometrica*, 71 (3), 933–946.

Bago d'Uva, T. (2005), "Latent Class Models for Use of Primary Care: Evidence from a British Panel," *Health Economics*, 14 (9), 873–892.

Bago d'Uva, T. (2006), "Latent Class Models for Utilisation of Health Care," *Health Economics*, 15 (4), 329–343.

Bhat, C. (2001), "Quasi-Random Maximum Simulated Likelihood Estimation of the Mixed Logit Model," *Transportation Research part B*, 35 (7), 677–693.

Bijwaard, G. and G. Ridder (2005), "Correcting for Selective Compliance in a Re-Employment Bonus Experiment," *Journal of Econometrics*, 125 (1–2), 77–111.

Blischke, W.R. (1964), "Estimating the Parameters of Mixtures of Binomial Distributions," *Journal of the American Statistical Association*, 59 (306), 510–528.

Brooks, S.P., B.J.T. Morgan, M.S. Ridout, and S.E. Pack (1997), "Finite Mixture Models for Proportions," *Biometrics*, 53 (3), 1097–1115.

Buntin, M.B. and A.M. Zaslavsk (2004), "Too Much Ado about Two-Part Models and Transformation?: Comparing Methods of Modeling Medicare Expenditures," *Journal* *of Health Economics*, 23 (3), 525–542.

Cai, T., L. Tian, and L.J. Wei (2005), "Semiparametric Box-Cox Power Transformation Models for Censored Survival Observations," *Biometrika*, 92 (3), 619–63.

Cameron, A.C. and P. Johansson (1997), "Count Data Regression Using Series Expansions: With Applications," *Journal of Applied Econometrics*, 12 (3), 203–223.

Cameron, A.C. and P. Johansson (1998), "Bivariate Count Data Regression Using Series Expansions: With Applications," Department of Economics Discussion Paper 9815, University of California Davis.

Cameron, A.C., F. Milne, and J. Piggott (1988), "A Microeconomic Model of the Demand for Health Care and Health Insurance in Australia," *Review of Economic Studies*, 55 (1), 85–106.

Cameron, A.C. and P.K. Trivedi (1998), *Regression Analysis of Count Data*, Cambridge University Press, Cambridge.

Cameron, A.C. and P.K. Trivedi (2005), *Microeconometrics: Methods and Applications*, Cambridge University Press, Cambridge.

Cameron, A.C., T. Li, P.K. Trivedi, and D.M. Zimmer (2004), "Modelling the Differences in Counted Outcomes Using Bivariate Copula Models with Application to Mismeasured Counts," *Econometrics Journal*, 7 (2), 566–584.

Cameron, A.C. and F.A.G. Windmeijer (1996), "R-Squared Measures for Count Data Regression Models with Applications to Health-Care Utilization," *Journal of Business and Economic Statistics*, 14 (2), 209–220.

Cantoni, E. and E. Ronchetti (2006), "A Robust Approach for Skewed and Heavy-Tailed Outcomes in the Analysis of Health Care Expenditures," *Journal of Health Economics*, 25 (2), 198–213.

Chen, S. and S. Khan (2003), "Semiparametric Estimation of a Heteroskedastic Sample Selection Model," *Econometric Theory*, 19 (6), 1040–1064.

Christofides, L.N., Q. Li, Z. Liu, and I. Min (2003), "Recent Two-Stage Sample Selection Procedures with an Application to the Gender Wage Gap," *Journal of Business and Economic Statistics*, 21 (3), 396–405.

Coppejans, M. (2001), "Estimation of the Binary Response Model Using a Mixture of Distributions Estimator (MOD)," *Journal of Econometrics*, 102 (2), 231–269.

Coppejans, M. and A.R. Gallant (2002), "Cross-Validated SNP Density Estimates," *Journal of Econometrics*, 110 (1), 27–65.

Costa, D.L. (1995), "Pensions and Retirement: Evidence from Union Army Veterans," *Quarterly Journal of Economics*, 110 (2), 297–319.

Cruz-Medina, I.R., T.P. Hettmansperger, and H. Thomas (2004), “Semiparametric Mixture Models and Repeated Measures: The Multinomial Cut Point Model,” *Journal of the Royal Statistical Society Series C*, 53 (3), 463–474.

Davidson, R. and J.G. Mackinnon (1993), *Estimation and Inference in Econometrics*, Oxford University Press, Oxford.

Davidson, R. and J.G. Mackinnon (2003), *Econometric Theory and Methods*, Oxford University Press, Oxford.

Deb, P. (2001), “A Discrete Random Effects Probit Model with Application to the Demand for Preventive Care,” *Health Economics*, 10 (5), 371–383.

Deb, P. and A.M. Holmes (1998), “Substitution of Physicians and Other Providers in Outpatient Mental Health Care,” *Health Economics*, 7 (4), 347–361.

Deb, P. and A.M. Holmes (2000), “Estimates of Use and Costs of Behavioural Health Care: A Comparison of Standard and Finite Mixture Models,” *Health Economics*, 9 (6), 475–489.

Deb, P., M.K. Munkin, and P.K. Trivedi (2006a), “Bayesian Analysis of the Two-Part Model with Endogeneity: Application to Health Care Expenditure,” *Journal of Applied Econometrics*, 21 (7), 1081–1099.

Deb, P., M.K. Munkin, and P.K. Trivedi (2006b), “Private Insurance, Selection, and Health Care Use: A Bayesian Analysis of A Roy-Type Model,” *Journal of Business and Economic Statistics*, 24 (4), 403–415.

Deb, P. and P.K. Trivedi (1997), “Demand for Medical Care by the Elderly: A Finite Mixture Approach,” *Journal of Applied Econometrics*, 12 (3), 313–336.

Deb, P. and P.K. Trivedi (2001), “Equity in Swedish Health Care Reconsidered: New Results Based on the Finite Mixture Model,” *Health Economics*, 10 (6), 565–572.

Deb, P. and P.K. Trivedi (2002), “The Structure of Demand for Health Care: Latent Class Versus Two-Part Models,” *Journal of Health Economics*, 21 (4), 601–625.

Deb, P. and P.K. Trivedi (2006), “Specification and Simulated Likelihood Estimation of a Non-Normal Treatment-Outcome Model with Selection: Application to Health Care Utilization,” *Econometrics Journal*, 9 (2), 307–331.

Dempster, A.P., N.M. Laird, and D.B. Rubin (1977), “Maximum Likelihood from Incomplete Data via the EM Algorithm,” *Journal of the Royal Statistical Society Series B*, 39 (1), 1–38.

DeSarbo, W.S. and J. Choi (1998), “A Latent Structure Double Hurdle Regression Model for Exploring Heterogeneity in Consumer Search Patterns,” *Journal of Econometrics*, 89 (1–2), 423–455.

Dow, W.H. and E.C. Norton (2003), "Choosing between and Interpreting the Heckit and Two-Part Models for Corner Solutions," *Health Services and Outcomes Research Methodology*, 4 (1), 5–18.

Duan, N., W.G. Manning, C.N. Morris, and J.P. Newhouse (1983), "A Comparison of Alternative Models for the Demand for Medical Care," *Journal of Business and Economic Statistics*, 1 (2), 115–126.

Duan, N., W.G. Manning, C.N. Morris, and J.P. Newhouse (1984), "Choosing between the Sample-Selection Model and the Multi-Part Model," *Journal of Business and Economic Statistics*, 2 (3), 283–289.

Elmore, R.T. and S. Wang (2003), "Identifiability and Estimation in Finite Mixture Models with Multinomial Components," Technical Report 03–04 Department of Statistics, Pennsylvania State University.

Evans, W.N. and R.M. Schwab (1995), "Finishing High School and Starting College: Do Catholic Schools Make a Difference?," *Quarterly Journal of Economics*, 110 (4), 941–974.

Ferrall, C. (2005), "Solving Finite Mixture Models: Efficient Computation in Economics under Serial and Parallel Execution," *Computational Economics*, 25 (4), 343–379.

Firpo, S. (2007), "Efficient Semiparametric Estimation of Quantile Treatment Effects," *Econometrica*, 75 (1), 259–276.

Follmanna, D.A. and D. Lambert (1989), "Generalizing Logistic Regression by Nonparametric Mixing," *Journal of the American Statistical Association*, 84 (405), 295–300.

Freund, D.A., T.J. Kniesner, and A.T. Losasso (1999), "Dealing with the Common Econometric Problems of Count Data with Excess Zeros, Endogenous Treatment Effects, and Attrition Bias," *Economics Letters*, 62 (1), 7–12.

Gabler, S., F. Laisney, and M. Lechner (1993), "Seminonparametric Estimation of Binary-Choice Models with an Application to Labor-Force Participation," *Journal of Business and Economic Statistics*, 11 (1), 61–80.

Gallant, A.R. (1981), "On the Bias in Flexible Functional Forms and an Essentially Unbiased Form : The Fourier Flexible Form," *Journal of Econometrics*, 15 (2), 211–245.

Gallant, A.R. and D.W. Nychka (1987), "Semi-Nonparametric Maximum Likelihood Estimation," *Econometrica*, 55 (2), 363–390.

Gallant, A.R. and G. Tauchen (1989), "Seminonparametric Estimation of Conditionally Constrained Heterogeneous Processes: Asset Pricing Applications," *Econometrica*, 57 (5), 1091–1120.

Gerdtham, U.G. (1997), "Equity in Health Care Utilization: Further Tests Based on Hurdle Models and Swedish Micro Data," *Health Economics*, 6 (3), 303–19.

Gerdtham, U.G. and P.K. Trivedi (2001), "Equity in Swedish Health Care Reconsidered: New Results Based on the Finite Mixture Model," *Health Economics*, 10 (6), 565–572.

Geweke, J. (1992), "Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments (with Discussion)," in *Bayesian Statistics*, J.Bernardo, J.Berger, A.P. Dawid, and A.F.M. Smith (Eds.), 4, 169–193, Oxford University Press, Oxford.

Geweke, J. (1995), "Monte Carlo Simulation and Numerical Integration," Federal Reserve Bank of Minneapolis Research Department Staff Report 192.

Geweke, J. and M.P. Keane (1999), "Mixture of Normals Probit Models," Cambridge University.

Goffe, W., G.D. Ferrier, and J. Rogers (1994), "Global Optimization of Statistical Functions with Simulated Annealing," *Journal of Econometrics*, 60 (1), 65–99.

Gouriéroux, C. (2000), *Econometrics of Qualitative Dependent Variables*, Cambridge University Press, Cambridge.

Gouriéroux, C. and A. Monfort (1991), "Simulation Based Inference in Models with Heterogeneity," *Annals Economic Statistics*, 20–21, 69–107.

Gouriéroux, C. and A. Monfort (1996), *Simulation Based Econometric Methods*, Oxford University Press, New York.

Greene, W.H. (1997), "FIML Estimation of Sample Selection Models for Count Data," Working Papers from New York University, Leonard N. Stern School of Business, Department of Economics.

Greene, W.H. (2007a), *Econometric Analysis 6th ed.*, Prentice Hall, New Jersey.

Greene, W.H. (2007b), *Functional Form and Heterogeneity in Models for Count Data*, Now Publishers, Boston.

Gurmu, S. (1997), "Semi-Parametric Estimation of Hurdle Regression Models with an Application to Medicaid Utilization," *Journal of Applied Econometrics*, 12 (3), 225–242.

Gurmu, S. (1998), "Generalized Hurdle Count Data Regression Models," *Economics Letters*, 58 (3), 263–268.

Gurmu, S. and J. Elder (2000), "Generalized Bivariate Count Data Regression Models," *Economics Letters*, 68 (1), 31–36.

Gurmu, S., P. Rilstone, and S. Stern (1999), "Semiparametric Estimation of Count Regression Models," *Journal of Econometrics*, 88 (1), 123–150.

Gurmu, S. and P.K. Trivedi (1996), "Excess Zeros in Count Models for Recreational Trips," *Journal of Business and Economic Statistics*, 14 (4), 469–477.

Hajivassiliou, V.A. and D. McFadden (1994), "A Simulation Estimation Analysis of the External Debt Crises of Developing Countries," *Journal of Applied Econometrics*, 9 (2), 109–131.

Hall, P., A. Neeman, P. Pakyari and R. Elmore (2005), "Nonparametric Inference in Multivariate Mixtures," *Biometrika*, 92 (3), 667–678.

Halton, J. (1960), "On the Efficiency of Evaluating Certain Quasi-Random Sequences of Points in Evaluating Multi-Dimensional Integrals," *Numerische Mathematik*, 2 (1), 84–90.

Heckman, J.J. (1978), "Dummy Endogenous Variables in a Simultaneous Equations System," *Econometrica*, 46 (4), 931–959.

Heckman, J.J. and B. Singer (1984a), "A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data," *Econometrica*, 52 (2), 271–320.

Heckman, J.J. and B. Singer (1984b), "The Identifiability of the Proportional Hazard Model," *The Review of Economic Studies*, 51 (2), 231–241.

Heckman, J.J. and R. Robb (1985), "Alternative Methods for Evaluating the Impact of Interventions: An Overview," *Journal of Econometrics*, (1–2), 239–267.

Hellström, J. (2006), "A Bivariate Count Data Model for Household Tourism Demand," *Journal of Applied Econometrics*, 21 (2), 213–226.

Hettmansperger, T.P. and H. Thomas (2000), "Almost Nonparametric Inference for Repeated Measures in Mixture Models," *Journal of the Royal Statistical Society Series B*, 62 (4), 811–825.

Holland, P.W. (1986), "Statistics and Causal Inference," *Journal of the American Statistical Association*, 81 (396), 945–960.

Jemernéz-Martín, S., J.M. Labeaga, and M. Matínez-Granado (2002), "Latent Class Versus Two-Part Models in the Demand for Physician Services across the European Union," *Health Economics*, 11 (4), 301–321.

Jochmann, M. (2003), "Semiparametric Bayesian Inference for Count Data Treatment Models," mimeo.

Jones, A.M. (2007), *Applied Econometrics for Health Economists: A Practical Guide*, Radcliffe Publishing, Oxford.

Jones, A.M. and O. O'Donnell (2002), *Econometric Analysis of Health Data*, Wiley, New York.

Jones, A.M., N. Rice, T. Bago d'Uva, and S. Balia (2007), *Applied Health Economics*, Routledge, London.

Jones, A.M., N. Rice, T. Bago d'Uva, and S. Balia (2012), *Applied Health Economics 2nd ed.*, Routledge, London.

Judd, K.L. (1998), *Numerical Methods in Economics*, MIT Press, Oxford.

Kasahara, H. and K. Shimotsu (2010), "Nonparametric Identification of Multivariate Mixtures," Discussion Papers 2010–09 Graduate School of Economics Hitotsubashi University.

Keane, M.P. (1994), "A Computationally Practical Simulation Estimator for Panel Data," *Econometrica*, 62 (1), 95–116.

Kenkel, D.S. and J.V. Terza (2001), "The Effect of Physician Advice on Alcohol Consumption: Count Regression with an Endogenous Treatment Effect," *Journal of Applied Econometrics*, 16 (2), 165–184.

Kozumi, H. (2002), "A Bayesian Analysis of Endogenous Switching Models for Count Data," *Journal of the Japan Statistical Society*, 32 (2), 141–154.

Lahiri, K. and G. Xing (2004), "An Econometric Analysis of Veterans' Health Care Utilization Using Two-Part Models," *Empirical Economics*, 29 (2), 431–449.

Lee, M.J. (2004), "Selection Correction and Sensitivity Analysis for Ordered Treatment Effect on Count Response," *Journal of Applied Econometrics*, 19 (3), 323–337.

Lee, M.J. (2005a), *Micro-Econometrics for Policy, Program, and Treatment Effects*, Oxford University Press, Oxford.

Lee, M.J. (2005b), "Monotonicity Conditions and Inequality Imputation for Sample-Selection and Non-Response Problems," *Econometric Reviews*, 24 (2), 175–194.

Lee, M.J. and F. Vella (2006), "A Semi-Parametric Estimator for Censored Selection Models with Endogeneity," *Journal of Econometrics*, 130 (2), 235–252.

Leung, S.F. and S. Yu (1996), "On the Choice between Sample Selection and Two-Part Models," *Journal of Econometrics*, 72 (1–2), 197–22.

Lopez-Nicolas, A. (1998), "Unobserved Heterogeneity and Censoring in the Demand for Health Care," *Health Economics*, 7 (5), 429–437.

Maddala, G.S. (1986), Limited Dependent and Qualitative Variables in Econometrics, Cambridge University Press, Cambridge.

Manning, W.G., N. Duan, and W.H. Rogers (1987), "Monte Carlo Evidence on the Choice between Sample Selection and Two-Part Models," *Journal of Econometrics*, 35 (1), 59–82.

Martínez-Espiñeira, R. (2006), "A Box-Cox Double-Hurdle Model of Wildlife Valuation: The Citizen's Perspective," *Ecological Economics*, 58 (1), 192–208.

Martins, M.F.O. (2001), "Parametric and Semiparametric Estimation of Sample Selection Models: An Empirical Application to the Female Labour Force in Portugal," *Journal of Applied Econometrics*, 16 (1), 23–39.

Masuhara, H. (2007), "Semi-Nonparametric Estimation of Regression-Based Survival Models," *Economics Bulletin*, 3 (61), 1–12.

Masuhara, H. (2008), "Semi-Nonparametric Count Data Estimation with an Endogenous Binary Variable," *Economics Bulletin*, 3 (42), 1–13.

McFadden, D. and K. Train (2000), "Mixed MNL Models for Discrete Response," *Journal of Applied Econometrics*, 15 (5), 447–470.

McLachlan, G.J. and T. Krishnan (1996), *The EM Algorithm and Extensions*, Wiley, New York.

McLachlan, G.J. and D. Peel (2000), *Finite Mixture Models*, Wiley, New York.

Melenberg, B. and A. van Soest (1993), "Semi-Parametric Estimation on the Sample Selection Model," Working Paper 9334, CentER Tilburg.

Melenberg, B. and A. van Soest (1996), "Parametric and Semi-parametric Modelling of Vacation Expenditures," *Journal of Applied Econometrics*, 11 (1), 59–76.

Miranda, A. and S. Rabe-Hesketh (2006), "Maximum Likelihood Estimation of Endogenous Switching and Sample Selection Models for Binary, Count, and Ordinal Variables," *Stata Journal*, 6 (3), 285–308.

Moffitt, R. (1991), "Program Evaluation With Nonexperimental Data," *Evaluation Review*, 15 (3), 291–314.

Mroz, T.A. (1999), "Discrete Factor Approximations in Simultaneous Equation Models: Estimating the Impact of a Dummy Endogenous Variable on a Continuous Outcome," *Journal of Econometrics*, 92 (2), 233–274.

Mullahy, J. (1986), "Specification and Testing of Some Modified Count Data Models," *Journal of Econometrics*, 33 (3), 341–365.

Mullahy, J. (1997a), "Instrumental-Variable Estimation of Count Data Models: Applications to Models of Cigarette Smoking Behavior," *Review of Economics and Statistics*, 79 (4), 586–593.

Mullahy, J. (1997b), "Heterogeneity, Excess Zeros, and the Structure of Count Data Models," *Journal of Applied Econometrics*, 12 (3), 337–350.

Mullahy, J. (1998), "Much Ado about Two: Reconsidering Retransformation and the Two-Part Model in Health Econometrics," *Journal of Health Economics*, 17 (3), 247–281.

Mullahy, J. (2001), "Estimating Log Models: To Transform or not to Transform?," *Journal of Health Economics*, 20 (2), 461–494.

Munkin, M.K. (2003), "The MCMC and SML Estimation of a Self-Selection Model with Two Outcomes," *Computational Statistics and Data Analysis*, 42 (3), 403–424.

Munkin, M.K. and P.K. Trivedi (1999), "Simulated Maximum Likelihood Estimation on Multivariate Mixed-Poisson Regression Models, with Application," *Econometrics Journal*, 2 (1), 29–48.

Munkin, M.K. and P.K. Trivedi (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 (2), 197–220.

Nielsen, S.F. (2000), "On Simulated EM Algorithms," *Journal of Econometrics*, 96 (2), 267–292.

Newey, W.K., J.L. Powell, and J.R. Walker (1990), "Semiparametric Estimation of Selection Models: Some Empirical Results," *American Economic Review*, 80 (2), 324–328.

Pohlmeier, W. and V. Ulrich (1995), "An Econometric Model of the Two-Part Decision-making Process in the Demand for Health Care," *Journal of Human Resources*, 30 (2), 339–361.

Prieger, J. (2002), "A Flexible Parametric Selection Model for Non-Normal Data with Application to Health Care Usage," *Journal of Applied Econometrics*, 17 (4), 367–392.

Rabe-Hesketh, S., A. Skrondal, and A. Pickles (2002), "Reliable Estimation of Generalized Linear Mixed Models Using Adaptive Quadrature," *The Stata Journal*, 2 (1), 1–21.

Rabe-Hesketh, S., A. Skrondal, and A. Pickle (2005), "Maximum Likelihood Estimation of Limited and Discrete Dependent Variable Models with Nested Random Effects," *Journal of Econometrics*, 128 (2), 301–323.

Raikou, M. and A. McGuire (2004), "Estimating Medical Costs under Conditions of Censoring," *Journal of Health Economics*, 23 (3), 443–470.

Riphahn, R.T., A. Wambach, and A. Million (2003), "Incentive Effects in the Demand for Health Care: A Bivariate Panel Count Data Estimation," *Journal of Applied Econometrics*, 18 (4), 387–405.

Rivers, D. and Q.H. Vuong (1988), "Limited Information Estimators and Exogeneity Tests for Simultaneous Probit Models," *Journal of Econometrics*, 39 (3), 347–366.

Romeu, A. and A.M. Vera-Hernández (2000), "A Semi-Nonparametric Estimator for Counts with an Endogenous Dummy Variable," UFAE and IAE Working Papers from Unitat de Fonaments de l'Anàlisi Econòmica (UAB) and Institut d'Anàlisi Econòmica (CSIC) No. 452.

Romeu, A. and M. Vera-Hernández (2005), "Counts with an Endogenous Binary Regressor: A Series Expansion Approach," *Econometrics Journal*, 8 (1), 1–22.

Rubin, D.B. (1974), "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66 (5), 688–701.

Ruud, P.A. (1991), "Extensions of Estimation Methods Using the EM Algorithm," *Journal of Econometrics*, 49 (3), 305–341.

Saab, Y.G. and V.B. Rao (1990), "Stochastic Evolution: A Fast Effective Heuristic for Some Genericlayout Problems," Design Automation Conference 1990 Proceedings, ACM/IEEE.

Sait, S.M. and H. Youssef (2000), *Iterative Computer Algorithms with Applications in Engineering: Solving Combinatorial Optimization Problems*, Wiley IEEE Computer Society Press, New York.

Santos Silva, J.M.C. (1997a), "Generalized Poisson Regression for Positive Count Data," *Communications in Statistics Part B: Simulation and Computation*, 26 (3), 1089–1102.

Santos Silva, J.M.C. (1997b), "Unobservables in Count Data Models for On-Site Samples," *Economics Letters*, 54 (3), 217–220.

Santos Silva, J.M.C. (2001), "A Score Test for Non-Nested Hypotheses with Applications to Discrete Data Models," *Journal of Applied Econometrics*, 16 (5), 577–597.

Santos Silva, J.M.C. and F. Covas (2000), "A Modified Hurdle Model for Completed Fertility," *Journal of Population Economics*, 13 (2), 173–188.

Santos Silva, J.M.C. and F. Windmeijer (2001), "Two-Part Multiple Spell Models for Health Care Demand," *Journal of Econometrics*, 104 (1), 67–89.

Schellhorn, M. (2001), "A Comparison of Alternative Methods to Model Endogeneity in Count Models. An Application to the Demand for Health Care and Health Insurance Choice," Social and Economic Dimensions of an Aging Population Research Papers No. 40.

Stewart, M.B. (2004), "Semi-Nonparametric Estimation of Extended Ordered Probit Models," *Stata Journal*, 4 (1), 27–39.

Stewart, M.B. (2005), "A Comparison of Semiparametric Estimators for the Ordered Response Model," *Computational Statistics and Data Analysis*, 49 (2), 555–573.

Teicher, H. (1960), "On the Mixture of Distributions," *Annals of Mathematical Statistics*, 31 (1), 55–73.

Teicher, H. (1963), "Identifiability of Finite Mixtures," *Annals of Mathematical Statistics*, 34 (4), 1265–1269.

Terza, J.V. (1998), "Estimating Count-Data Models with Endogenous Switching: Sample Selection and Endogenous Treatment Effects," *Journal of Econometrics*, 84 (1), 129–54.

Terza, J.V. and P.W. Wilson (1990), “Analyzing Frequencies of Several Types of Events: A Mixed Multinomial-Poisson Approach,” *Review of Economics and Statistics*, 72 (1), 108–115.

Train, K.E. (2000), “Halton Sequences for Mixed Logit,” Working Paper No. E00–278, Department of Economics, University of California Berkley.

Train, K.E. (2003), *Discrete Choice Methods with Simulation*, Cambridge University Press, Cambridge.

Ueda, N. and R. Nakano (1998), “Deterministic Annealing EM Algorithm,” *Neural Networks*, 11 (2), 271–282.

van de Ven, W.P.M.M. and B.M.S. van Praag (1981), “The Demand for Deuctibles in Private Health Insurance: A Probit Model with Sample Selection,” *Journal of Econometrics*, 17 (2), 229–252.

van der Klaauw, B. and R.H. Koning (2003), “Testing the Normality Assumption in the Sample Selection Model with an Application to Travel Demand,” *Journal of Business and Economic Statistics*, 21 (1), 31–42.

van Ophem, H. (2000), “Modeling Selectivity in Count-Data Models,” *Journal of Business and Economic Statistics*, 18 (4), 503–511.

van Outri, T. (2004), “Measuring Horizontal Inequity in Belgian Health Care Using a Gaussian Random Effects Two Part Count Data Model,” *Health Economics*, 3 (7), 705–724.

Vuong, Q.H. (1989), “Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses,” *Econometrica*, 57 (2), 307–333.

Wang, P. (2003), “A Bivariate Zero-Inflated Negative Binomial Regression Model for Count Data with Excess Zeros,” *Economics Letters*, 78 (3), 373–378.

Wang, P. and J.D. Alba (2006), “A Zero-Inflated Negative Binomial Regression Model with Hidden Markov Chain,” *Economics Letters*, 92 (2), 209–213.

Wedel, M., W.S. Desarbo, J.R. Bult, and V. Ramaswamy (1993), “A Latent Class Poisson Regression Model for Heterogeneous Count Data,” *Journal of Applied Econometrics*, 8 (4), 397–411.

White, H. (2001), *Asymptotic Theory for Econometricians*, Academic Press, Orlando.

Windmeijer, F.A.G. (2000), “Moment Conditions for Fixed Effects Count Data Models with Endogenous Regressors,” *Economics Letters*, 68 (1), 21–24.

Windmeijer, F.A.G. and J.M.C. Santos Silva (1997), “Endogeneity in Count Data Models: An Application to Demand for Health Care,” *Journal of Applied Econometrics*, 12 (3), 281–294.

Winkelmann, R. (1996), “A Count Data Model for Gamma Waiting Times,” *Statistical Papers*, 37 (2), 177–187.

Winkelmann, R. (2000), “Seemingly Unrelated Negative Binomial Regression,” *Oxford Bulletin of Economics and Statistics*, 62 (4), 553–560.

Winkelmann, R. (2003), *Econometric Analysis of Count Data 4th ed.*, Springer, Berlin.

Winkelmann, R. (2004a), “Co-Payments for Prescription Drugs and the Demand for Doctor Visits: Evidence from a Natural Experiment,” *Health Economics*, 13 (11), 1081–1089.

Winkelmann, R. (2004b), “Health Care Reform and the Number of Doctor Visits: An Econometric Analysis,” *Journal of Applied Econometrics*, 19 (4), 455–472.

Winkelmann, R. (2006), “Reforming Health Care: Evidence from Quantile Regressions for Counts,” *Journal of Health Economics*, 25 (1), 131–145.

Winkelmann, R. and S. Boes (2006), *Analysis of Microdata*, Springer, Berlin.

Winkelmann, R. and K.F. Zimmermann (1995), “Recent Developments in Count Data Modelling: Theory and Application,” *Journal of Economic Surveys*, 9 (1), 1–24.

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

Zimmer, D.M. and P.K. Trivedi (2006), “Using Trivariate Copulas to Model Sample Selection and Treatment Effects: Application to Family Health Care Demand,” *Journal of Business and Economic Statistics*, 24 (1), 63–76.