Essays on Unobserved Heterogeneity and Endogeneity in Health Econometrics

Hiroaki Masuhara

Abstract

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."

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. 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 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.

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

$\mathrm{E}(y \mid \mathbf{x}) = 0 \times P(y = 0 \mid \mathbf{x}) + 1 \times P(y = 1 \mid \mathbf{x})$. Therefore, we obtain

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

If we calculate the prediction using this model, it is required that $0 \leq P(\widehat{y} = 1 \mid \mathbf{x}_0) \leq 1$. 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 $\mathrm{Var}(y = 1 \mid \mathbf{x}) = P(\widehat{y} = 1 \mid \mathbf{x})[1 - P(\widehat{y} = 1 \mid \mathbf{x})]$, 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 pa-

rameters 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.