Semiparametric duration analysis with an endogenous binary variable: An application to hospital stays

Hiroaki Masuhara[†*]

Department of Health Services Management, Hiroshima International University,
1-5 Nobori-cho Naka-ku, Hiroshima-shi, Hiroshima 730-0016, Japan

Abstract

Background
In duration analysis, we find situations where covariates are simultaneously determined along with the duration variable. Moreover, although the models based on a hazard rate do not explicitly assume heterogeneity, in applied econometrics, the possibility of omitted variables is inevitable and controlling population heterogeneity alone is inadequate. It is important to consider both heterogeneity and endogeneity in duration analysis.

Objectives and methods
Explicitly assuming semiparametric correlated heterogeneity, this paper proposes an alternative robust duration model with an endogenous binary variable that generalizes the heterogeneity of both duration and endogeneity using Hermite polynomials. 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.

Results
The parameter values of the endogenous binary variable (insurance choice) are statistically significant at the 1% level; however, the values differ among the parametric and semiparametric models and the any type of insurance choice increases the length of hospital stays by 104.010% in the censored parametric model, and 182.074% in the censored semiparametric model. Compared with the parametric model, the increase of hospital stays in the semiparametric model is large. Moreover, we find that the semiparametric model a twin-peak distribution and that the contour lines differ from the usual ellipsoids of the bivariate normal density.

Conclusions
When applied to the duration of hospital stays of the MEPS data, the estimated results of the semiparametric model shows a good performance. The absolute values of the endogenous binary regressor coefficients of the semiparametric models are larger than that of the parametric model. The parametric model underestimates the effect of the individual's insurance choice in our example. Moreover, the estimated densities of the semiparametric models have twin peak distribution.

JEL classification: C14; C31; C34
Key words: Endogenous switching; duration analysis; probit; semi-nonparametric model; heterogeneity

---

[†] Tel.: +81 82 554 2059; Fax: +81 82 211 5166; E-mail address: h-masuha@hw.hirokoku-u.ac.jp
[*] I would like to thank Kei Hosoya for his valuable comments and suggestions. This study is supported by the Grant-in-Aid for Young Scientists (B) 21730215 from the Ministry of Education, Culture, Sports, Science and Technology of the Japanese Government. Moreover, this paper is a part of the academic project on Economic Analysis of Intergenerational Issues: Searching for Further Development, funded by the Grand-in-Aid for Specially Promoted Research from Japan's Ministry of Education, Culture, Sports, Science and Technology (grant number 22000001). All remaining errors are mine.

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 ordinary least squares (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.

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, a 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 1 explains this discussion. In Rows 2 and 3 in Table 1, the two-stage method has consistency in linear models regardless irrespective of any endogenous regressors. In Rows 4 and 5, in nonlinear models, the two-stage method is consistent when endogenous variables are continuous, but the full information maximum likelihood (FIML) has consistency when endogenous variables are discrete, censored, or truncated.

<<Insert Table 1>>

In duration (survival) analysis, 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 non-linearity, 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 controlling population heterogeneity alone is inadequate. Therefore, in duration analysis, it is important to consider both heterogeneity and endogeneity.

This paper 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 lognormal 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).

Using examples of probit models, Section 2 explains the two-stage method used 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 inadequate and the FIML is consistent. Moreover, this section provides simple Monte Carlo simulations and analyzes the consistency of proposed models. Section 3 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.

2. Endogenous Regressors in Nonlinear Econometric Models

Before discussing duration analysis with an endogenous binary variable, this subsection considers estimation for nonlinear models with an endogenous binary or continuous variable, using a probit model, because the properties of censored data are essentially same as those of binary data. 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:

$$y_1^* = x'\beta_1 + \alpha_1 y_2 + \varepsilon_1, \quad (1)$$

$$y_2 = z'\beta_2 + \varepsilon_2, \quad (2)$$

where $(\varepsilon_1, \varepsilon_2)$ has a zero mean, a bivariate normal distribution, and is independent of 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 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

$$\varepsilon_1 = \theta_1 \varepsilon_2 + u_1, \quad (3)$$

where $\theta_1 = \rho/\sigma_2$, $\sigma_2^2 = \mathrm{Var}(\varepsilon_2)$, and u_1 is independent of 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 $\mathrm{E}[u_1]=0$ and $\mathrm{Var}[u_1] = \mathrm{Var}[\varepsilon_1] - \rho^2$. We can now write

$$y_1^* = x'\beta_1 + \alpha_1 y_2 + \theta_1 \varepsilon_2 + u_1,$$
$$u_1 | z, y_2, \varepsilon_2 \sim N(0, 1-\rho^2).$$

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

$$y_1^* = x'\beta_1 + \alpha_1 y_2 + \sqrt{1-\rho^2}u + \theta_1 \varepsilon_2.$$

A standard calculation shows that

$$\mathrm{P}(y_1 = 1 | z, y_2, \varepsilon_2) = \Phi\left[(x'\beta_1 + \alpha_1 y_2 + \theta_1 \varepsilon_2)/(1-\rho^2)^{1/2}\right]. \quad (4)$$

Assuming for the moment that we observe ε_2, then probit of y_1 on z, y_2, and ε_2 consistently estimates $\beta_{\rho 1} \equiv \beta_1/(1-\rho^2)^{1/2}$, $\alpha_{\rho 1} \equiv \alpha_1/(1-\rho^2)^{1/2}$, and $\theta_{\rho 1} \equiv \theta_1/(1-\rho^2)^{1/2}$. 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 z and save the residuals $\hat{\varepsilon}_2$. Second, the probit y_1 on x, y_2, and $\hat{\varepsilon}_2$ obtains consistent estimators of the scaled coefficients $\beta_{\rho 1}$, $\alpha_{\rho 1}$, and $\theta_{\rho 1}$. The probit parameters are estimated only up to scale, with factor $(1-\rho^2)^{-1/2}$. An estimate for ρ is $\hat{\rho}^2 = \hat{\theta}_\rho^2 \hat{\sigma}_2^2/\left(1 + \hat{\theta}_\rho^2 \hat{\sigma}_2^2\right)$, where $\hat{\sigma}_2$ 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 $\mathrm{H}_0: \theta_1 = 0$ 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 $\mathrm{H}_0: \theta_1 = 0$, 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.

A Binary Endogenous Regressor in Binary Models

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

$$y_1 = 1[x'\beta_1 + \alpha_1 y_2 + \varepsilon_1 > 0], \quad (5)$$

$$y_2 = 1[z'\beta_2 + \varepsilon_2 > 0], \quad (6)$$

where $1[\cdot]$ is an indicator function, $(\varepsilon_1, \varepsilon_2)$ is independent of 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 β_1 and α_1. In this model, the effect of y_2 is often 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 x) is calculated by $\Phi(x'\beta_1 + \alpha_1) - \Phi(x'\beta_1)$.

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

$$\mathrm{P}(y_1 = 1|y_2, z) = \Phi\left[\frac{x'\beta_1 + \alpha_1 y_2 + \rho\varepsilon_2}{(1-\rho^2)^{1/2}}\right]. \quad (7)$$

Moreover, the truncated density of ε_2 given $\varepsilon_2 > -z'\beta_2$ obtains

$$\frac{\phi(\varepsilon_2)}{\Phi(\varepsilon_2 > -z'\beta_2)} = \frac{\phi(\varepsilon_2)}{\Phi(z'\beta_2)}. \quad (8)$$

Therefore, the density $\mathrm{P}(y_1 = 1|y_2 = 1, z)$ takes

$$\mathrm{P}(y_1 = 1|y_2 = 1, z) = \frac{1}{\Phi(z'\beta_2)} \int_{-z'\beta_2}^{\infty} \Phi\left[\frac{x'\beta_1 + \alpha_1 y_2 + \rho\varepsilon_2}{(1-\rho^2)^{1/2}}\right] \mathrm{d}\varepsilon_2. \quad (9)$$

Similarly, $\mathrm{P}(y_1 = 1|y_2 = 0, z)$ is

$$\mathrm{P}(y_1 = 1|y_2 = 0, z) = \frac{1}{1 - \Phi(z'\beta_2)} \int_{-\infty}^{-z'\beta_2} \Phi\left[\frac{x'\beta_1 + \rho\varepsilon_2}{(1-\rho^2)^{1/2}}\right] \mathrm{d}\varepsilon_2. \quad (10)$$

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 $\varepsilon_q > -z'\beta_2$, where ε_q is the evaluation point of the Gauss-Hermite quadrature. Since $-z'\beta_2$ is not constant under the maximization process, a small change in the value of β_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. 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 $\mathrm{E}(y_2|z) = \Phi(z'\beta_2)$ and β_2 is consistently estimated by the probit of y_2 on z, it is tempting to estimate β_1 and α_1 from the probit of y_1 on x and $\widehat{\Phi}_2$, where $\widehat{\Phi}_2 \equiv \Phi(z'\beta_2)$. However, the two-stage method is inappropriate because the estimated coefficients are inconsistent. Although the two-stage method requires $\mathrm{P}(y_1 = 1|z) = \Phi[x'\beta_1 + \alpha_1\Phi(z'\beta_2)]$, we can compute only the expected value $\mathrm{P}(y_1 = 1|z) = \mathrm{E}(y_1 = 1|z) = \mathrm{E}(1[x'\beta_1 + \alpha_1 y_2 + \varepsilon_1 > 0])$. Since the indicator function $1[\cdot]$ is nonlinear, we cannot correctly specify the expected value. If, substituting $\hat{\beta}_2$, we can compute the correct and complicated formula for $\mathrm{P}(y_1 = 1|z)$, the two-stage approach produces consistent estimators, but the FIML is easier and more efficient.

Monte Carlo Results

We summarize results of the Monte Carlo experiments of linear estimation with endogenous continuous and binary variables to evaluate the finite sample performance. We show the

inconsistency of the two-stage method in estimating nonlinear models, such as probit models, with an endogenous binary variable. The Monte Carlo simulations are designed as follows. We generate one explanatory variable, z_1, drawn independently from $N(0,1/4)$, and two unobserved heterogeneity terms, ε_1 and ε_2, normally distributed as $N\big((0,0),(\sigma_1^2,\rho\sigma_1\sigma_2,\sigma_2^2)\big)$. The variable y_2 represents an endogenous continuous variable assumed to be generated by the process $y_2 = z'\beta_2 + \varepsilon_2$; d represents an endogenous binary variable and is assumed to be generated by the process $d = 1$ if $d^* = z'\beta_2 + \varepsilon_2$ and $d = 0$; otherwise, where $z = [1, z_1]'$ and $\beta_2 = [\beta_{21}, \beta_{22}]'$. Variable y_1 represents a binary dependent variable assumed to be generated by the process $y_1 = 1$ if $y_1^* = x'\beta_1 + \alpha_1 d + \varepsilon_1 > 0$ (for an endogenous binary variable) or $y_1^* = x'\beta_1 + \alpha_1 y_2 + \varepsilon_1 > 0$ (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. 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 2 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 2 (a) and both the two-stage method and FIML in Table 2 (b). 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 2 (a), the mean bias (BIAS) and root mean squared error (RMSE) of an endogenous continuous variable decrease when the number of observations 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}.

<<Insert Table 2>>

The results of probit models with an endogenous binary variable appear in Table 2 (b). 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.

3. Semiparametric Duration Analysis with an Endogenous Binary Variable

Regardless of whether endogenous regressors are continuous, linear models with those variables display consistency using the two-stage method. 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 duration analysis with an endogenous binary variable requires FIML. However, this method always contains the problem of specifications of distribution. That is, if the distributions of both nonlinear duration 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 semiparametric distributions.

We consider a lognormal model in duration analysis based on Masuhara (2007): $\ln t_i = \beta_d d_i + x_i'\beta_1 + \varepsilon_{1i}$, where t_i, $i = 1, \ldots, N$, is an observed duration outcome that has a continuous

probability density $f(t_i)$; $x_i \sim k_1 \times 1$ and $z_i \sim k_2 \times 1$ denote regressors (explanatory variables or covariates); β_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 $d_i^* = z_i'\beta_2 + \varepsilon_{2i} \geq 0$ and $d_i = 0$ otherwise, where d_i^* is a latent variable; ε_{2i} is unobserved heterogeneity; β_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 $(\sigma_1^2, \rho\sigma_1, 1)$, 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:

$$f(\varepsilon_{1i}, \varepsilon_{2i}) = \frac{1}{P}\left(\sum_{j=0}^{K}\sum_{k=0}^{K}\alpha_{jk}\varepsilon_{1i}^{j}\varepsilon_{2i}^{k}\right)^2 \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \times \exp\left[-\frac{1}{2(1-\rho^2)}\left\{\left(\frac{\varepsilon_{1i}}{\sigma_1}\right)^2 - 2\rho\frac{\varepsilon_{1i}}{\sigma_1}\frac{\varepsilon_{2i}}{\sigma_2} + \left(\frac{\varepsilon_{2i}}{\sigma_2}\right)^2\right\}\right] \equiv \frac{f^*}{P}, \tag{11}$$

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:

$$\ln L = \sum_{i=1}^{N}(1-c_i)\ln\left[\frac{\Psi(\varepsilon_{1i})}{P}\frac{1}{\sigma_1}\phi\left(\frac{\varepsilon_{1i}}{\sigma_1}\right)\right] + c_i \ln\left[\int_{\underline{\varepsilon}_{1i}}^{\infty}\frac{\Psi(\varepsilon_{1i})}{P}\frac{1}{\sigma_1}\phi\left(\frac{\varepsilon_{1i}}{\sigma_1}\right)\mathrm{d}\varepsilon_{1i}\right], \tag{12}$$

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 $\underline{\varepsilon}_{1i} \equiv \ln t_i - \beta_d d_i - x_i'\beta_1$. 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:

$$\Psi(\varepsilon_{1i}) = \begin{cases} \int_{-\infty}^{-z_i'\beta_2} \psi(\varepsilon_{2i}|\varepsilon_{1i})\mathrm{d}\varepsilon_{2i} & \text{if} \quad d_i = 0 \\ \int_{-z_i'\beta_2}^{\infty} \psi(\varepsilon_{2i}|\varepsilon_{1i})\mathrm{d}\varepsilon_{2i} & \text{if} \quad d_i = 1 \end{cases}. \tag{13}$$

After some algebraic computation, Equation (13) 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 β_1 (or β_d), a small change in the value of β_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

[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).

(1994), and Keane (1994) to evaluate the log-likelihood in the censored part.[3] The GHK simulator is described as follows:

1) 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.
2) Calculate $\left[1 - \Phi\left(\underline{\varepsilon}_{1i}/\sigma_1\right)\right]\Psi(\varepsilon_{1s})/P$.
3) Repeat the steps 1 to 2 S times, and calculate the simulated probability: $\left[1 - \Phi\left(\underline{\varepsilon}_{1i}/\sigma_1\right)\right]\sum_{s=1}^{S}\Psi(\varepsilon_{1s})/(P \times S)$.

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 $\left[1 - \Phi\left(\underline{\varepsilon}_{1i}/\sigma_1\right)\right]$ for $\Psi(\varepsilon_{1s})/P$. Unlike the GH quadrature or SML, the GHK simulator calculates the log-likelihood on *fixed* evaluation points.

Moreover, this paper 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, ...). Similar sequences are defined for other prime numbers, such as 3 ($1/3$, $2/3$, $1/9$, $4/9$, $7/9$, $2/9$, $5/9$, $8/9$, ...). In order to obtain corresponding standard normal points from each Halton draw, we take the inverse standard normal distribution transformation: $\Phi^{-1}(1/2) = 0$, $\Phi^{-1}(1/4) = -0.67$, $\Phi^{-1}(3/4) = 0.67$, ..., where Φ^{-1} is an inverse of the cumulative density function of the standard normal.

4. 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 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 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

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

(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 3.

<<Insert Table 3>>

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 private insurance, medicare, medicaid, and HMO; this is done so as to avoid the difficulties involved in estimating multivariate probit models of high 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 4 and 5 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 $K = 2$.[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.

<<Insert Table 4>>

<<Insert Table 5>>

Figure 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.

<<Insert Figure 1>>

5. Conclusion

This paper 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 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

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

[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.

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

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.

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.

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.

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.

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.

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.

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.

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

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

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.

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

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

Table 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

Table2: Monte Carlo Results

(a) Monte Carlo Results of Probit Models with an Endogenous Continuous Variable

		ρ = 0.3		ρ = 0.6		ρ = 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)

(b) Monte Carlo Results of Probit Models with an Endogenous Binary Variable

		ρ = 0.3		ρ = 0.6		ρ = 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)

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 3: 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 typec	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/).

Table 4: Estimated Results of Hospital Stays (Selection Equation)

	non-censored data				artificial censored data at t=30			
	parametric		SNP (K=2)		parametric		SNP (K=2)	
selection equation								
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.

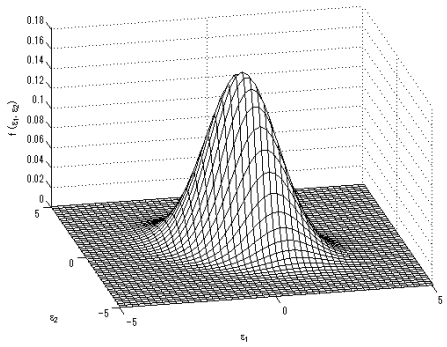
Table 5: Estimated Results of Hospital Stays (Duration Equation)

	non-censored data				artificial censored data at t=30			
	parametric		SNP (K=2)		parametric		SNP (K=2)	
duration equation								
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)

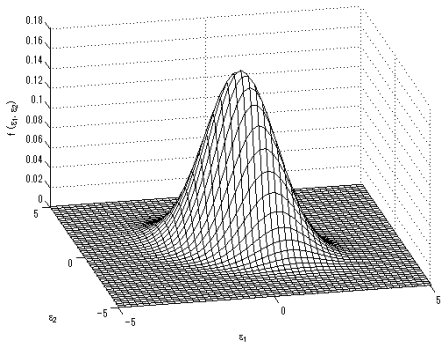
Notes: SNP denotes the semi-nonparametric duration model; standard errors are in parentheses.

Figure 1: Estimated Densities of Heterogeneity

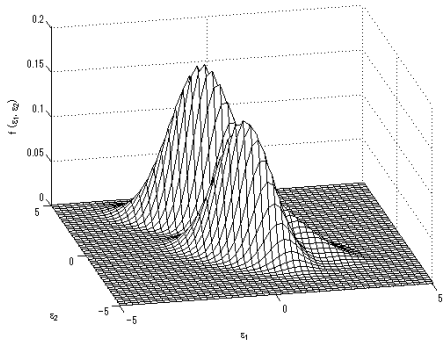
(a) non-censored (parametric)



(b) artificial censored (parametric)



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



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

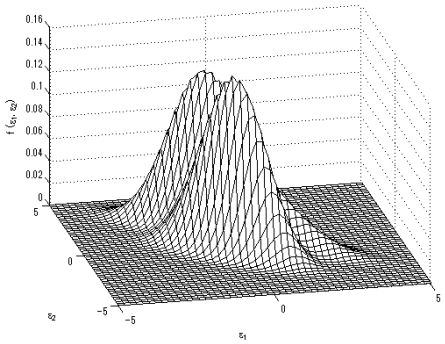


Table 1

	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

Table 2

(a) Monte Carlo Results of Probit Models with an Endogenous Continuous Variable

		ρ = 0.3		ρ = 0.6		ρ = 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)

(b) Monte Carlo Results of Probit Models with an Endogenous Binary Variable

		ρ = 0.3		ρ = 0.6		ρ = 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)

Notes: Figures without parentheses are mean bias (BIAS) values. Root mean squared errors (RMSE) appear

Table 3 MEPS Data: 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 typec	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 dum	0.238	0.426	0	1
SOUTH	Regional indicator (EAST is the excluded dum	0.363	0.481	0	1
WEST	Regional indicator (EAST is the excluded dum	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 mair	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

CENSOR		0.041	0.199	0	1
HOSPDUR2		6.373	7.762	0.5	31
HOSPSTAY	1 = individual had hospital stays	1.000	-----	1	1

Table 4

	non-censored data				artificial censored data at t =30			
	parametric		SNP (K =2)		parametric		SNP (K =2)	
duration equation								
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.

Table 5

	non-censored data				artificial censored data at t=30			
	parametric		SNP (K=2)		parametric		SNP (K=2)	
selection equation								
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-likelihoo	-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.