

A Flexible Parametric Selection Model for Non-Normal Data with Application to Health Care Usage

James E. Prieger*
Department of Economics
University of California
One Shields Avenue
Davis, CA 95616-8578
jeprieger@ucdavis.edu

December 2000

Abstract

I examine the effects of insurance status and managed care on hospitalization spells, and develop a new approach for sample selection problems in parametric duration models. MLE of the *Flexible Parametric Selection* (FPS) model does not require numerical integration or simulation techniques. I discuss application to the exponential, Weibull, log-logistic and gamma duration models. Applying the model to the hospitalization data indicates that the FPS model may be preferred even in cases in which other parametric approaches are available.

Keywords: sample selection, Farlie-Gumbel-Morgenstern distribution, duration models, Lee's model, health care, managed care, health insurance, Medical Expenditure Panel Survey

*This paper benefitted from comments by A. Colin Cameron, Lung-fei Lee, seminar participants at UC Davis, and the anonymous referees.

Much recent popular and academic attention has focused on whether insurance status and health maintenance organization (HMO) membership affect health care usage. This study examines the length of individual hospital stays by participants in a nationally representative survey. Of primary interest is the effect of insurance coverage and managed care on the probability of admittance to the hospital and the duration of stay. If insurance status and HMO membership are important determinants of whether an individual is hospitalized, then sample selection is an issue when analyzing the length of stay. It is natural to analyze the length of stay with duration data models, which then requires a selection model for duration data. A major contribution of this paper is to present a new approach to selection in duration models, and to compare it with existing approaches.

HMOs are the most common form of *managed care* plans. Managed care is a catch-all term for mechanisms that attempt to reduce the cost of health care by moving away from unrestricted fee-for-service contracts with health care providers, and may save up to one-third relative to fee-for-service care (American Academy of Actuaries, 1996).¹ Under fee-for-service contracts and marginal cost pricing, the health care provider has the incentive to maximize the quality of care (where quality is broadly construed to include patient health, reputation of the institution, etc.), which has led to quality levels in the U.S. health care industry that have been criticized as inefficiently high. If the goal of the HMO is to reduce the quality of care (again, broadly construed), then one expects hospital stays to be less likely and shorter for individuals in HMO plans.²

Furthermore, whether the individual is insured and the type of insurance will affect the probability and duration of a stay. On the demand side, since insurance greatly reduces the price of health care for the insured, individuals would be more likely to agree to hospitalization (although many hospital admissions are not elective). On the supply side, health care institutions may be more willing to admit insured patients, believing that payment from uninsured patients is less

¹Managed care components may include limits on the length of hospital stays or full capitation, in which payment is received per head regardless of health care usage. See Frank and Lave (1989) for a discussion of reimbursement types and the incentives they provide to health care providers.

²Whether HMOs reduce *patient care* quality in particular is disputed. See Levinson and Ullman (1996) for an indication that infant health quality is preserved under managed care.

likely.³ Once in the hospital, the duration of the patient's stay may be presumed to be up to the medical staff. Whether the hospital has incentive to release the uninsured earlier than the insured will depend on their beliefs about the probability of payment from the uninsured and the form of contract with the insurance companies for the insured. There is evidence that hospitals respond to incentives to alter their care provided based on the expected generosity of the payer (Dor and Farley, 1996).

Measuring the effects of managed care and insurance on hospitalization requires untangling the selection effect. If unobserved health factors make individuals more likely to enter the hospital and more likely to stay longer, for example, then data from actual hospitalizations form a biased sample. Econometric models with selection effects are now commonplace in applied work. The appropriate selectivity model for the task at hand is the *sample selection* model.⁴ In the sample selection model, whether a response y is observed depends on the value of a selection variable d . In the present application, y represents the length of a hospital stay and d represents admission to the hospital. Estimation of the sample selection model, studied first by Gronau (1974) and Heckman (1974), usually proceeds by choosing a bivariate parametric model for (d, y) or by semi- or non-parametric procedures. For a recent survey of the numerous estimation procedures available for the sample selection problem, see Vella (1998). Examples of parametric models include Heckman's (1976) two-stage estimation procedure, developed for the bivariate normal model, and Lee's (1983) model, which can incorporate any pair of continuous distributions.⁵ There are trade-offs between the parametric and less parametric approaches. The parametric approach is efficient, typically easy to estimate, and allows inclusion of large numbers of explanatory variables. The nonparametric approach is robust, and the semi-parametric approach falls in between. In this paper, I present

³The uninsured tend to be low wage workers who earn just enough to disqualify themselves for Medicaid. Of those employed but earning less than \$20,000 in 1996, over half were uninsured; of the uninsured, 57% worked full time (or their spouse did) (Gardner, 1997).

⁴The model is also known as the *incidental truncation* model.

⁵Semi- and non-parametric approaches include Duncan (1986) and other articles in that issue, Manski (1989; 1990), Newey, Powell and Walker (1990), and Donald (1995).

a new parametric approach, the *Flexible Parametric Selection* (FPS) model, that provides an alternative to Lee's model for non-normal distributions.

To present the selectivity model, consider the standard linear form of the sample selection model (I generalize the model in section 2). Letting asterisks indicate latent, unobserved variables, the model is:

$$d_i^* = \alpha' z_i - \varepsilon_i, \tag{1}$$

$$y_i^* = \beta' x_i + u_i, \tag{2}$$

$$d_i = 1\{d_i^* > 0\}, \tag{3}$$

$$y_i \text{ observed as } y_i^* \text{ only if } d_i = 1. \tag{4}$$

It is well known that OLS performed on y_i is biased if ε_i and u_i are correlated. In Heckman's approach, one assumes that (ε_i, u_i) follow a bivariate normal distribution. Amemiya (1985) called this the Type 2 Tobit model. In Lee's generalization, one assumes instead that (ε_i, u_i) have cumulative distribution function

$$\Phi_2(J_\varepsilon(\varepsilon_i), J_u(u_i); \rho); \text{ where} \tag{5}$$

$$J_\varepsilon(\varepsilon_i) = \Phi^{-1}(F_\varepsilon(\varepsilon_i)), \tag{6}$$

$$J_u(u_i) = \Phi^{-1}(F_u(u_i)), \tag{7}$$

and where Φ^{-1} is the inverse of the standard normal cumulative density function (cdf), Φ_2 is the bivariate normal cdf with unit variances and correlation ρ , and F_a is the cdf of a , $a = \varepsilon, u$. In Lee's model one can specify any absolutely continuous cdf for F_ε and F_u . These cdfs are assumed to be known (up to a finite vector of parameters to be estimated). From the above one can derive the likelihood of a sample (d_i, y_i) (see Lee, 1983) and perform maximum likelihood estimation.

Thus Lee's model specifies a particular bivariate distribution that has marginal distributions f_u and f_ε and correlation that is increasing in ρ .⁶ While Lee's model has the advantage of allowing for

⁶When ε is normal with unit variance, as is often the case in applications of Lee's model, one can show that the correlation between ε and u in 5 is $(\rho/\sigma) \int u f_u(u) J_u(u) du$, where integration is over the support of u .

maximal correlation (Mardia, 1970*b*), one may be interested in alternative bivariate distributions for the purpose of comparison. Several properties may be desired of a general bivariate distribution for the latent variables in the sample selection model. In particular, one wants the two variables to exhibit correlation and the marginal distributions to take any form allowed for continuous data. Furthermore, for computational ease one may further wish that the likelihood of the observed variables does not require numerical integration or simulation techniques. The model proposed in this paper satisfies all three criteria, just as Lee's model does.⁷

In the next section I present the distribution for the latent variables in the model. The distribution allows correlation through a single parameter in an easily interpretable fashion: when correlation is positive, above-median values of the two variables are likely to appear together. The distribution does not allow for maximal correlation, however. Section 2 applies the distribution to the sample selection case, leading to a sample selection likelihood in closed form. In section 3 I apply the selection model to common duration distributions (exponential, Weibull, log-logistic, and gamma), and contrast the allowed correlation with that of other selection models. In each case the correlation allowed by the FPS model is limited, but is greater than that allowed by some of the competing models in some cases. In section 4 I return to the empirical application introduced above. The analysis of hospitalization incidence and duration shows that the model may be preferred over Lee's model in this application when evaluated by formal statistical criteria. The superiority of the model in the application is notwithstanding that both models are available and that Lee's model allows more correlation. The illustration reveals that selection effects are present in the hospitalization duration data, and that HMOs do not decrease hospitalizations but may reduce their duration. A final section discusses relaxing the parametric assumptions in the model and points to future work.

⁷Technically, the presence of Φ and Φ^{-1} in the likelihood of Lee's model preclude a closed form expression, but these functions are built into most programming packages.

1 The Latent Bivariate Distribution

I begin by specifying the bivariate distribution underlying the proposed selection model. The distribution is a simple form of the Farlie-Gumbel-Morgenstern (FGM) system of bivariate distributions (Kotz, Balakrishnan and Johnson, 2000, sec. 44.13), which has been available in the statistics literature for some time but has not previously been applied to econometrics. The marginal distributions in the FGM system may be of any form, as long as their cdfs are absolutely continuous. The latent selection random variable is D^* taking values $d^* \in \mathbb{R}$, and takes the observed value d in accordance with (3). The latent “selected” random variable is Y^* taking values $y^* \in \mathcal{Y}$, where \mathcal{Y} is a subset of \mathbb{R}_+ (duration data). Selection is as in (4).

Instead of specifying a joint distribution for the error terms in (1)–(2), I work directly with the distribution of (D^*, Y^*) , since many duration models do not naturally fit into the linear model (2). Let the marginal probability density function (pdf) of D^* , conditional on covariates z and finite parameter vector α be $f_{D^*}(d^*|z, \alpha)$, and let $E(D^*|z, \alpha) = \alpha'z$ and $Var(D^*|z, \alpha) = \tau^2$, where α is unknown and τ^2 is known. The two leading examples that I consider for f_{D^*} are the normal distribution, leading to a probit selection equation with $\tau^2 = 1$, and the logistic distribution, leading to a logit selection equation with $\tau^2 = \pi^2/3$. Fixing the variance of D^* to τ^2 is required for identification of α in the selection equation.

Likewise, let the marginal pdf of Y^* , conditional on covariates x and finite (unknown) parameter vector $\theta = (\beta, \gamma)$ be $f_{Y^*}(y^*|x, \theta)$, and let $E(Y^*|x, \theta) = \mu(\beta'x, \gamma)$ and $Var(Y^*|x, \theta) = \zeta^2(\beta'x, \gamma)$. In these formulations, the β is typically the coefficient of interest and γ is a nuisance parameter. I consider parametric forms for f_{Y^*} in section 3. The cdf of a random variable A will be denoted F_A , and let $\bar{F}_A = 1 - F_A$.

Then, suppressing the dependence on (x, z, α, θ) in most of the notation, the bivariate cdf of the FGM distribution is:

$$F_{D^*, Y^*}(d^*, y^*|x, z, \alpha, \theta) = F_{D^*}(d^*)F_{Y^*}(y^*) \{1 + \omega \bar{F}_{D^*}(d^*)\bar{F}_{Y^*}(y^*)\}, \quad -1 \leq \omega \leq 1, \quad (8)$$

from which the pdf is readily found as:

$$f_{D^*,Y^*}(d^*, y^*|x, z, \alpha, \theta) = f_{D^*}(d^*)f_{Y^*}(y^*) \{1 + \omega G_{D^*}(d^*)G_{Y^*}(y^*)\}, \quad (9)$$

where

$$G_A(a) = F_A(a) - \bar{F}_A(a), \quad A = D^*, Y^*. \quad (10)$$

There is a natural interpretation to the correlation parameter ω . First, notice from (9) that if $\omega = 0$ then D^* and Y^* are independent. If $\omega > 0$ and both D^* and Y^* are above their median values, then the bracketed term adds to the likelihood. So larger than median values of D^* and Y^* will tend to appear together when ω is positive, and the same for smaller than median values.

Figure 1 plots illustrative isoprobability curves from f_{D^*,Y^*} for the case when D^* is normal and Y^* is exponential. In the figure the lines represents the same probability level for different values of ω . From Figure 1, note that when ω is negative, larger (smaller) than median values of D^* will tend to appear with smaller (larger) than median values of Y^* . Thus the correlation between D^* and Y^* has the same sign as ω . In fact, the (conditional) correlation is linear in ω :

$$\rho_P \equiv Corr(D^*, Y^*|x, z) = \omega \frac{H_{D^*}H_{Y^*}}{\tau_\zeta}, \quad (11)$$

where ρ_P is Pearson's cross-product measure of correlation and $H_A = \int F_A \bar{F}_A(a) da > 0$, for $A = D^*, Y^*$ (Johnson and Kotz, 1977).

The range of allowed correlation is dependent on the distributions chosen, but is $1/3$ at most (?). For the exponential duration models, the allowed correlation is about $(-0.3, 0.3)$, as will be shown in section 3, where I compare the Pearson correlation allowed by the FGM and competing models on a case by case basis. Another convenient characterization of correlation between D^* and Y^* is Kendall's tau, τ_K .⁸ For the FGM distribution (8), τ_K (conditional on x and z) is $2\omega/9$ no matter which marginal distributions are chosen, and thus is bounded on $[-2/9, 2/9]$.⁹ Thus the

⁸Kendall's tau is a measure of the "concordance" of two random variables, and is defined as $\tau_K \equiv 2 \Pr\{(D_1^* - D_2^*)(Y_1^* - Y_2^*) > 0\} - 1$, where A_i is an independent draw from F_A , $A = D^*, Y^*$ and $i = 1, 2$. τ_K is bounded between -1 and 1 and is 0 if D^* and Y^* are independent.

⁹From the definition of τ_K , it follows that $\tau_K = 4EF_{D^*,Y^*}(D^*, Y^*)$, from which (8) and (9) yield the result.

FPS model will be appropriate only when the marginal distributions are moderately correlated. By comparison, in Lee's model $\tau_K = (2/\pi) \sin^{-1}(\rho)$, taking values on $[-1, 1]$, a consequence of the model allowing maximal correlation (Mardia, 1970*b*); τ_K takes the extreme values -1 and 1 if and only if a bivariate distribution hits the Fréchet bounds for extreme correlation (Kruskal, 1958).¹⁰

Given that the main alternative to my proposed model for non-normal data is Lee's model, it is worth comparing the correlation patterns of the two models at a deeper level than Pearson's or Kendall's summary statistics. For convenience, in this section I refer to the distribution underlying Lee's model, (5), as the TBN (for Translated Bivariate Normal) distribution.¹¹ Figures 2 and 3 plot the density functions from the FGM and TBN distributions for the case of uniform marginals and positive correlation. Such plots, known as the *uniform representation* of a bivariate distribution, are convenient for comparing bivariate distributions (Kimeldorf and Sampson, 1975*b*). In these plots any spikes or dips are purely a consequence of correlation;¹² in the absence of correlation the pdfs would be flat.

For both distributions, positive correlation has the effect of concentrating probability mass toward the (0,0) and (1,1) corners. However, from the plots one can see that the conditional distributions in the FGM model inherit certain characteristics of the original uniform marginals, more so than in the TBN model. For example, when either variable takes the median value of 0.5 in the FGM distribution, the other variable is uniformly distributed (notwithstanding the positive correlation). Furthermore, in the FGM distribution, the pdf is positive at the extreme values (the borders of the unit square), just as in the uniform marginals. The conditional distribution of one variable given the other is a straight line, with positive slope if the conditioning variable is above the median (and vice versa). On the other hand, the conditional distributions in the TBN

¹⁰Pearson's ρ_P may be strictly inside $[-1, 1]$ at the Fréchet bounds. For example, when D^* is normal and Y^* is exponential, ρ_P is bounded on $[-0.903, 0.903]$.

¹¹The TBN appears to have originated with Nataf (1962); correlation is studied by Mardia (1970*b*).

¹²In the figures, $\omega = 1$ and $\rho = 0.5$. The plots are qualitatively the same for any positive correlation, except that at $\rho = 1$ all probability mass in the TBN collapses to the line $D^* = Y^*$. Negative correlation rotates the graphs 90 degrees.

model, by construction as translates of a normal distribution, inherit the normal's strong central tendencies; the pdf is zero at all extreme values. The FGM distribution incorporates correlation in a fashion that preserves characteristics of the marginals in the conditional distributions, which may be advantageous when the marginal distributions are chosen to reflect desired characteristics—as would be the case in most duration data applications, for example. The price for the preservation of the marginals' features is the limited correlation allowed.

Finally, note that one advantage of the FGM distribution is its tractability. The additive form of the correlation terms in (8) and (9) leads to explicit analytical forms for most expressions of interest, such as the sample selection likelihood. No bivariate distribution allowing maximal correlation, that I have found, is as simple to work with.¹³

2 Sample Selection

I now generalize the sample selection model, (1)–(4), using the new notation. I term the resulting model the *Flexible Parametric Selection* (FPS) model. The probability that $D = 0$ is

$$\Pr\{D = 0|z, \alpha\} = F_{D^*}(0) \quad (12)$$

The likelihood function for $D = 1$ and $Y = y$ is given by the integral of the joint density over the region where Y is observed:

$$\int_0^\infty f_{D^*, Y^*}(t, y) dt = f_{Y^*}(y) \bar{F}_{D^*}(0) [1 + \omega F_{D^*}(0) G_{Y^*}(y)]. \quad (13)$$

From (12) and (13) the joint distribution can be written as

$$f_{D, Y}(d, y) = \left[F_{D^*}(0)^{1-d} \bar{F}_{D^*}(0)^d \right] \{f_{Y^*}(y) [1 + \omega F_{D^*}(0) G_{Y^*}(y)]\}^d \quad (14)$$

$$= f_D(d) f_{Y|D}(y|d=1) \quad (15)$$

As seen from (15), the joint density may be decomposed into the marginal density of the binary random variable D (the first, bracketed term of (14)) and the conditional density of the observed

¹³See Kimeldorf and Sampson (1975a) for a discussion of bivariate distributions allowing maximal correlation.

random variable Y conditional on observation (the second, braced term). The conditional density has an intuitive interpretation. When ω is positive, then Y stochastically dominates Y^* (in the sense that $F_{Y|D=1}(y) < F_{Y^*}(y) \forall y$).¹⁴ So an observed y is larger on average than is a y^* from the latent population. The opposite holds if ω is negative. Finally, as noted above, if there is no correlation ($\omega = 0$) then the conditional density reduces to the marginal density of Y^* .

Given fixed τ^2 (the variance of the latent selection variable), all other parameters are identified (subject to the usual multicollinearity concerns for the data). No exclusion restriction is required between x and z : the two may consist entirely of common variables. As usual in probit and logit models, the restriction on τ^2 implies that the coefficients in the selection equation are identified only up to scale.

With a pairwise iid sample from $f_{D,Y}(d, y)$, estimation may proceed by FIML on (14), which requires no numerical integration. I use FIML in the empirical implementation of the model in section 4. However, Newey et al. (1990) note that the joint likelihood in a selection model may be ill determined in practice, causing FIML to be computationally cumbersome. In such cases estimation may also proceed by LIML on f_D for $\hat{\alpha}$ (which will be a standard probit or logit problem) or $f_{Y|D}$ for $(\hat{\alpha}, \hat{\theta})$.

The conditional expectation of the observed y is:

$$E(y|d = 1, x, z, \alpha, \theta, \omega) = \mu(\beta'x, \gamma) + \omega F_{D^*}(0|z, \alpha) H_{Y^*}(x, \beta, \gamma), \quad (16)$$

where H_{Y^*} is as in (11).¹⁵ Equation (16) reveals why inference based on the observed y 's and μ only is biased if $\omega \neq 0$. The second term on the right side of (16) is a non-ignorable selection term. For example, if $\omega > 0$ then the selection term is positive and the mean in the observed sample is greater than the true mean of the latent population, as noted above. Incidentally, equation (16)

¹⁴To see this, note that when y is at the median of Y^* , y^m , we have $G_{Y^*}(y^m) = 0$ and $f_{Y|D}(y^m|d = 1) = f_{Y^*}(y^m)$. For $\omega > 0$ and $y > y^m$, we have $f_{Y|D}(y^m|d = 1) > f_{Y^*}(y^m)$. For $\omega > 0$ and $y < y^m$, we have $f_{Y|D}(y^m|d = 1) < f_{Y^*}(y^m)$.

¹⁵For some distributions H_{Y^*} may not have a closed form expression. Note however that H_{Y^*} does not appear in (14), and so is not required for FIML.

may be used as the basis for NLLS.¹⁶ The nonlinearity of the selection term implies that NLLS may be as computationally intensive as FIML, however.¹⁷

3 Application to Duration Models

In this section I apply the FPS model to selection in specific parametric duration models. If one is willing to assume that the durations of interest are lognormally distributed, then one may develop a standard probit selection model based on the bivariate normal distribution. The lognormal distribution is not suitable for many applications, however, given that it exhibits a nonmonotonic hazard rate and does not admit a constant hazard rate as a special case. The other standard parametric duration models are the exponential, Weibull, log-logistic, and (less commonly) gamma. The densities, means, and variances of these distributions are presented in Table 1. The FPS model can readily incorporate any of these, coupled with either the logit or probit form of the selection equation.

The exponential model is often used as a baseline duration model because it exhibits a constant hazard rate. What correlation is allowed? As defined in (11), we have $H_{Y^*}/\varsigma = 0.5$. For probit selection, $H_{D^*}/\tau \simeq 0.564$. For logit selection, $H_{D^*}/\tau = \sqrt{3}/\pi \simeq .551$. Thus allowed correlation between the latent variables is 0.282ω for the probit exponential model and 0.276ω for the logit exponential model. Table 2 contains the correlation allowed by the FPS model and also gives the Fréchet bounds for comparison.

Another common duration model is log-logistic model. In the log-logistic model, $\log(y^*)$ follows the logistic distribution, and also has a shape parameter $\gamma > 0$. The log-logistic distribution has finite mean if $\gamma < 1$ and finite variance if $\gamma < 1/2$. The hazard rate is decreasing for $\gamma \geq 1$ and

¹⁶Two-stage approaches are also available. First, estimate $\hat{\alpha}$ by MLE based on $f_D(d)$ (probit or logit). Then perform NLLS on $E(y|d = 1, x, z, \hat{\alpha}, \theta, \omega)$ in (16) to find $(\hat{\theta}, \hat{\omega})$. For improved efficiency, one can use the resulting $(\hat{\alpha}, \hat{\theta}, \hat{\omega})$ and estimated variance to perform NLWLS.

¹⁷If the integral in H_{Y^*} cannot be solved analytically, then NLLS is actually more computationally intensive than FIML.

has a \cap shape for $\gamma < 1$. The Weibull and gamma models add a shape parameter $\gamma > 0$ to the exponential model. When $\gamma = 1$, they reduce to the exponential model. When $\gamma > 1$ in the Weibull model, the hazard is monotonically decreasing and the durations exhibit negative duration dependence. When $\gamma < 1$ in the Weibull model, the hazard is monotonically increasing and shows positive duration dependence. The opposite pattern holds for the gamma model. The allowed correlation for these models depends on the nuisance parameter γ . Table 2 also lists the correlation for a few values of γ for the Weibull and log-logistic models.

It appears that the allowed correlation in the FPS model is quite limited, compared to the Fréchet bounds. It is important to note, however, that even if one develops a bivariate duration selection model based on the normal distribution, the correlation between the duration variable and the selection variable is much less than unity in general. To be precise, consider the bivariate normal duration selection (BNDS) model, which consists of (1), (3), (4), and

$$\log(y_i^*) = \beta'x_i + u_i, \tag{17}$$

where (u_i, ε_i) are distributed mean zero bivariate normal, with covariance matrix

$$\Sigma = \begin{bmatrix} \sigma^2 & \sigma\rho \\ \sigma\rho & 1 \end{bmatrix}. \tag{18}$$

The BNDS model is a natural transformation of the Type 2 Tobit Model for duration data, and is in fact a special case of Lee's model with $\log(y_i^*)$ replacing y_i^* in (2), $F_\varepsilon = \Phi$ in (6), and $F_u(u_i) = \Phi(u_i/\sigma)$ in (7).¹⁸ Although the transformed duration variable $\log(y_i^*)$ has the full range of correlation with d_i^* , the correlation between y_i^* and d_i^* can be shown to be $r(\sigma) = \rho\sigma \left(e^{\sigma^2} - 1\right)^{-1/2}$ (see footnote 6).¹⁹ The correlation function r goes to zero rapidly as σ increases.

The BNDS correlation function is plotted in Figure 4. The correlation functions for the FPS model and Lee's model for exponential durations (both with probit selection) are also given in

¹⁸Most of the original labor applications of the sample selection model (e.g., Heckman, 1974) used the BNDS, although because y^* was a wage, it was not interpreted as a duration model.

¹⁹In general for Lee's model, $|\text{corr}(Y^*, D^*)| \leq \rho$, with equality only when (Y^*, D^*) are bivariate normal (Mardia, 1970a, p.33).

Figure 4.²⁰ The comparison reveals that the BNDS model allows for more correlation than the exponential FPS model when σ is less than about two (and more correlation than Lee's exponential model when σ is less than 0.6) but admits less correlation for higher levels of σ . The lesson from Table 2 and Figure 4 is twofold. First, the correlation bounds in these models may be much less than $[-1, 1]$. Second, if the data exhibit both high variance and high correlation, the usual BNDS model is a poor choice.

The correlation functions for the Weibull models (FPS and Lee) are in Figure 5. The correlation functions for the Gamma models (not shown) are similar in shape to those for the Weibull models when plotted on a comparable scale. The log-logistic models (Figure 6) have severely limited correlation compared to the other distributions, due to the rapidity with which the variance of the log-logistic distribution approaches infinity as γ increases. Lee's model clearly allows more correlation in each case than the FPS models. Allowed correlation is only one dimension along which to judge a model, however, and the FPS model may be preferred over Lee's model for other reasons, which proves to be the case in the next section.

4 The Incidence and Duration of Hospitalization

4.1 Data

With the FPS model in hand, I now return to the hospitalization data. The duration data are the length of individual hospital stays in 1996 by participants in the Medical Expenditure Panel Survey (MEPS), a nationally representative survey of U.S. medical care and expenditures.²¹ The selection variable represents whether an individual had a hospital stay. What is the effect of insurance

²⁰The FPS correlations are functions of γ . To make them comparable to $r(\sigma)$, I reparameterized them to be functions of δ such that $\log(y)$ has variance δ^2 , just as $\log(y)$ for the lognormal model has variance σ^2 . For the Weibull model $\delta = \gamma\pi/\sqrt{6}$; for the log-logistic model $\delta = \gamma\pi/\sqrt{3}$; for the gamma model $\delta = \sqrt{\psi'(\gamma)}$. The x -axis in Figures 4–6 is σ for the lognormal curves and δ for the other curves.

²¹See <<http://www.meps.ahrp.gov/>> for more information on MEPS. Various data from the 1987 wave of the survey have been studied by many authors (e.g., Madrian, 1994; Deb and Trivedi, 1997; Gilleskie, 1998). The duration of hospitalization has been previously studied by Welch (1985), Frank and Lave (1989), and Rosenman (1993).

coverage and managed care on the probability of admittance to the hospital and the duration of stay? As mentioned in the introduction, if insurance status and HMO membership determine whether an individual is hospitalized, then sample selection affects the length of stay.

Three major forms of insurance are present in the sample: Medicare (MEDICARE), Medicaid (MEDICAID), and private insurance (PRIVINS) (see Table 3). Medicare is available to all U.S. residents who are 65 or older. Medicare participants may also purchase additional private insurance known as Medigap insurance (PRIVMCARE). Medicaid is available to low-income individuals; few individuals on Medicaid also have private insurance (PRIVMCAID). Some have both Medicaid and Medicare insurance (MCAREMCAID). Each of these types of coverage may be through an HMO (HMOPRIV, HMOMCARE, HMOMCAID). About half of those in the sample with private insurance are enrolled in HMOs. For Medicare and Medicaid enrollees, 22% and 38%, respectively, are in HMOs.

Given that individuals are likely to take into account their future expected health when choosing private insurance (and to a lesser extent Medicaid) (see, e.g., Ettner, 1997), I control for health status as much as possible. Medicare coverage may be considered exogenous because of its automatic enrollment procedure (Deb and Trivedi, 1997), although I include it in the endogeneity testing below. I include several measures of health status reported at the beginning of the survey:²² two self-perceived measures of the individual's health (POORHLTH and EXCLHLTH), the number of reported medical conditions (CONDN), the number of those conditions that are on a priority list (PRIOLIST), and an indicator for disability (ADLHELP).²³ The priority list contains conditions deemed important due to their prevalence, expense, or relevance to policy, so that PRIOLIST may be viewed as a proxy for the number of severe conditions an individual has.²⁴ Demographic

²²The first round of the survey took place from March to August 1996. The hospital stay data cover the entire calendar year.

²³ADLHELP is a dummy variable indicating that the individual requires help or supervision carrying out Instrumental Activities of Daily Living (IADL) or Activities of Daily Living (ADL). IADL includes using the telephone, paying bills, taking medications, preparing light meals, doing laundry, or going shopping. ADL includes personal care such as bathing, dressing, or getting around the house.

²⁴Conditions on the priority list include long-term life-threatening conditions (cancer, diabetes, emphysema, high

variables such as region (MIDWEST, SOUTH, WEST), sex (FEMALE),²⁵ age (AGE),²⁶ years of education (EDUC), race (BLACK, HISPANIC), and marriage and employment status (MARRIED, EMPLOYED) were included to capture additional heterogeneity among respondents. These covariates are available for 14,955 individuals out of the 15,692 adults in the survey with complete insurance information.²⁷ A list of variables, definitions, and summary statistics is in Table 4. No set of controls can capture all heterogeneity in factors influencing preference for health care (e.g., fear of doctors or hypochondria). Therefore, these controls notwithstanding, I return to the question of endogenous insurance choice in section 4.3 below.

4.2 Choice of a Duration Model

What is an appropriate duration model to use for the length of hospital stays? Figure 7 contains a non-parametric estimate of the baseline integrated hazard of the duration of a single hospital stay.²⁸ The integrated hazard is linear if the hazard rate is constant, and concave if the hazard rate is declining. The figure suggests a constant hazard rate for the first 10 days (which includes 93% of these durations), and a possibly smaller rate for the few durations longer than 10 days. Thus an appropriate model would allow for a declining hazard rate and allow testing down to exponential; either the gamma or Weibull models suffice. For reasons explained below, I choose the gamma distribution. The lognormal and finite-mean log-logistic distributions, with their \cap -shaped hazard rates, do not appear to be appropriate. Because the lognormal distribution is not appropriate, the

cholesterol, HIV/AIDS, hypertension, stroke), chronic manageable conditions (arthritis, asthma, gall bladder disease, stomach ulcers, and back problems), and certain mental health conditions (Alzheimers disease, dementias, depression, and anxiety disorders).

²⁵ I exclude pregnancy and pregnancy-related conditions from the data set.

²⁶ My sample excludes children (below 18 years of age).

²⁷ A further nine observations are dropped because of contradictory hospitalization data, reducing the sample size to 14,946.

²⁸ The estimate is from the Cox (1972; 1975) semiparametric proportional hazards model and the Breslow's estimate (Fleming and Harrington, 1984) of the survival curve. The Cox model takes the hazard rate to be $h_0(t)e^{\beta'x_i}$, where h_0 is an unspecified baseline hazard rate common to all individuals and x_i includes all explanatory variables in table 4. The survival curve is $S_i(t) = \exp(-H_i(t))$, where H_i is the integrated hazard. Given the Cox coefficient estimates, the survival curve is estimated, from which the baseline integrated hazard, $H_0(t) = \int_0^t h_0(s)ds$, is recovered. The integrated hazard is plotted instead of the hazard rate because H_0 is much smoother than h_0 . Only single-incidence hospitalizations were used in the estimation.

BNDS model is not as attractive as the FPS or Lee models in this case, which allow more flexibility in the specification of the hazard.

The MEPS data include the number of hospitalizations (HOSPNUM) and the sum of the length of all hospital stays in 1996 (HOSPDUR).²⁹ The duration of the individual hospital stays are not known when the individual is hospitalized more than once. For these relatively few cases (2.5% of the sample) the likelihood of an observation is calculated from convolutions of the density of a single spell. Given the choice between the Weibull and the gamma distributions, the gamma is more attractive because convolutions admit an analytical solution (sums of gamma random variables are also gamma),³⁰ whereas for the Weibull they do not. The log likelihood of the sample with probit selection, conditional on n_i , the number of hospitalizations per individual (HOSPNUM), is

$$L(\alpha, \beta, \gamma) = \sum_{i=1}^N (1 - d_i) \log \Phi(-\alpha' z_i) + d_i \log \Phi(\alpha' z_i) + d_i \left\{ (\gamma n_i - 1) y_i \log(y_i) - \lambda_i y_i + \gamma n_i \log(\lambda_i) - \log \Gamma(\gamma n_i) + \log \left(1 + \omega \Phi(-\alpha' z_i) \left[2 \frac{\Gamma_{\lambda y}(\gamma n_i)}{\Gamma(\gamma n_i)} - 1 \right] \right) \right\} \quad (19)$$

where $\gamma > 0$ is the gamma distribution shape parameter, and λ_i and the gamma function notation is as in Table 1. The model contains the special case of exponential durations for individual hospitalization spells when $\gamma = 1$. Hypothesis tests for $\gamma = 1$ fail to reject the hypothesis that the durations are exponentially distributed.³¹ The marginal effects of the covariates are robust between the gamma and exponential assumptions on individual durations; the marginal effects from the gamma model differ from those from the exponential model (reported in Table 7 below) by no more than 0.07. Further specification testing of the exponential assumption, based on conditional moments and designed to be sensitive to neglected heterogeneity in the hazard rate, yielded mixed

²⁹The restriction of the data to 1996 introduces a minuscule amount of censoring (0.4% of sample) into the sample when hospital stays begin before or end after the calendar year. Furthermore, the length of stay is recorded as number of nights spent in the hospital. Hospitalizations without a night spent in the hospital (0.7% of sample) were set to length 0.5.

³⁰The sum of n independent and identically distributed Gamma(λ, γ) random variables (notation as in table 1) is distributed Gamma($\lambda, n\gamma$).

³¹For the hypothesis tests, MLE assuming gamma-distributed individual durations was performed on the observed durations (results not reported). The shape parameter γ was estimated to be 1.05 (0.04). Neither a t -test (p -value = 0.17) nor a LR test (p -value = 0.17) rejects the null hypothesis that $\gamma = 1$.

results.³² As might be expected from Figure 7, there tend to be more observations in the upper tail of the distribution than the exponential model accounts for, even after controlling for the covariates. Nevertheless, for the sake of tractability all further work assumes that individual spells follow the exponential distribution.³³ When interpreting the fitted model, therefore, the reader is cautioned that the variance of the actual durations may be somewhat higher than that predicted by the fitted model.

4.3 Estimation of the Models

Many theoretical and empirical studies demonstrate that insurance choice may be endogenous in estimations in which health outcomes are the dependent variable (see, e.g., Cameron, Trivedi, Milne and Piggott (1988) for a theoretical model and econometric application). If the health insurance variables included in α and β are endogenous, then MLE based on (19) will yield inconsistent estimates. Accordingly, I test the insurance choice variables for endogeneity using standard methods for endogenous dummy variables. The details of the procedures and testing results are in the appendix. The testing fails to reject the null hypothesis that the insurance variables are exogenous in both the selection and the duration equations. While there is no question that some individuals take into account their expected future health outcomes when choosing insurance, the resulting endogeneity is not statistically discernible in these data.³⁴ For the rest of the estimations I assume the exogeneity of the insurance variables, a considerable econometric simplification, although caution should remain in interpreting the results.

³²The hypothesis test is based on second-degree Laguerre polynomials in the generalized residuals (the integrated hazard) (Sharma, 1992) and is computed via an auxiliary regression. The test is equivalent to Lancaster's (1985) LM test for unobserved heterogeneity. Given the well-known size problems of the auxiliary regression form of these conditional moment tests (see Prieger (2000a) for an in-depth exploration), the test statistic is referred to bootstrap critical values. The test does not reject the exponential distribution when an empirical residual bootstrap (in which draws from the fitted residuals from equation (22) are used to form the bootstrap samples) is performed (p -value = 0.55). Given the heteroskedasticity of these errors, draws for an observation with a given n_i were made only from residuals from observations with the same n_i . The test rejects the exponential distribution when a parametric bootstrap is performed (p -value = 0.00). Each test used 399 bootstrap replications.

³³When $\gamma = 1$, then $\Gamma_{\lambda y}(\gamma n_i)$ in (19) has a closed form solution since $n_i \in \mathbb{N}$.

³⁴Note that Cameron et al. (1988) also failed to reject exogeneity of insurance choice in their estimations for hospital admissions and length of stay for wealthy individuals (using data from Australia).

The estimation results are presented in Tables 5 (main variables) and 6 (demographic variables). All results presented are for probit selection and the exponential model for individual durations. The first estimation, presented in the first two columns, is the baseline model, in which the selection and duration equations are assumed independent. This is the equivalent of fixing $\omega = 0$ in the FPS model (or $\rho = 0$ in Lee’s model).

The second estimation, in columns three and four, is the FPS model with ω (the correlation parameter) free, from MLE based on (19).³⁵ Correlation between admittance and length of stay might have the following interpretation. If unobserved factors cause a person to have poorer health than average, that individual may be both more likely to enter the hospital and to need to stay longer than average, resulting in positive correlation. Negative correlation might arise from a correction effect: if an individual enters the hospital when the observables indicate that he should not (on average), then his condition may not be as severe as the average condition of admitted patients and the hospital stay consequently may be shorter. In this application, we have no *a priori* expectation on the sign of the correlation.

Estimates from Lee’s model (also with exponential durations) are in the last two columns. The coefficient estimates are generally similar across the models, with the exception of the constant and correlation terms. The marginal effects of the covariates from the baseline and FPS models are presented in Table 7. The major results are as follows.

- The correlation between the selection and duration variables is significantly negative in the FPS model. The estimate $\hat{\omega} = -0.87$ implies that the correlation between the latent selection and duration variables is about -0.24 (see Table 2). The negative correlation implies that the observed durations are shorter than those in the latent population; correcting for selection leads to a longer predicted mean individual stay (Table 7). The correlation and the different estimated duration constant in the FPS model lead to marginal duration effects that are

³⁵Since ω and ρ are restricted to the interval $[-1, 1]$, it is computationally convenient to reparametrize as $\tilde{\xi} = \Phi^{-1} [(\xi + 1)/2] \in \mathbb{R}$, for $\xi = \omega, \rho$. Standard errors reported in table 5 are calculated by the delta method.

markedly different than those from the baseline model (Table 7). Ignoring the selection effect in these data generally leads to marginal duration effects that are biased toward zero, compared to the selection-corrected effects.

- Enrollment in an HMO has no discernible effect on hospital admittance, possibly reflecting a lack of substitution possibilities between inpatient and outpatient care. HMOPRIV decreases the length of stay by 1.5 days on average, out of the predicted mean individual stay of 6.1 days (see Table 7). This finding may be consistent with the claim that HMOs reduce health care expenditure, although cost and readmittance data would need to be examined directly for substantiation. If so, however, then expenditure is reduced not by declining admittance but by shortening the stay. HMO status has no significant effect on duration when the individual is also enrolled in Medicare or Medicaid.
- The insurance variables PRIVINS, MEDICARE, and MEDICAID all increase the probability of admittance to the hospital. These primary effects are reinforced or tempered by the multiple-insurance indicators PRIVMCARE, PRIVMCAID, and MCAREMCAID (all significant). That insured individuals are more likely to receive medical care has been documented in other studies for measures such as visits to doctors' offices (Deb and Trivedi, 1997).³⁶ Of more novel interest is that persons with private insurance (PRIVINS) do not appear to stay in the hospital any longer than uninsured persons. MEDICARE and MEDICAID decrease the duration of stay; the latter effect is reversed if the individual also has private insurance (PRIVMCAID is positive and outweighs MEDICAID). Compare these findings with Dor and Farley's (1996) finding that hospitals tend to spend differing amounts on patients covered by Medicare, Medicaid, private insurance, and not covered, depending on the generosity of the payer. Again, direct observation on cost is not available in these data, but the present find-

³⁶Endogeneity of insurance choice would also lead to this result, if the health status controls do not adequately deal with the potential problem. Note however that the increased likelihood of hospitalization is as pronounced for MEDICARE as for the other two types of insurance, even though Medicare enrollment is nearly certainly exogenous.

ings are consistent with hospitals reducing expenditure on Medicare and Medicaid patients by shortening their stays.

- The health status controls all have expected signs in the hospitalization equation. The number of medical and priority conditions (CONDN and PRIOLIST), disability status (ADLHELP), and self-perceived poor health (POORHLTH) all increase the probability of hospital admittance; self-perceived good health (EXCLHLTH) has the opposite impact. In the duration equation, only POORHLTH and ADLHELP have significant impacts; both lengthen the stay.

A final issue is the relative performance of the FPS model versus the baseline model and Lee’s model. Recall that Lee’s model allows for more correlation; in this application, however, Lee’s model actually finds less correlation.³⁷ The estimates from Lee’s model are very similar to the independent model in general; the correlation parameter $\hat{\rho}$ is insignificant. Thus the marginal effects in Lee’s model are similar to those given for the baseline model in Table 7. Because neither the FPS model nor Lee’s model nests the other but both nest the baseline model, I use information criteria and Vuong’s test for model selection. The Akaike Information Criterion (AIC), Schwarz’s Bayesian Information Criterion (BIC), and the Consistent AIC (CAIC) all lead one to choose the FPS model over both Lee’s model and the baseline model.³⁸ All measures would also lead one to choose the baseline model over Lee’s model. These information criteria are not entirely satisfactory for distinguishing between the FPS and Lee models, however, since in that case they reduce to comparing the likelihoods. A more formal guide to model selection, Vuong’s (1989) test for non-nested hypotheses, also prefers the FPS model.³⁹ Thus the FPS model appears to be the

³⁷This finding is possible because Lee’s model incorporates correlation through deviations from mean, and the FPS models incorporate correlation through deviations from median (see section 1). For duration distributions, the median generally differs from the mean.

³⁸The AIC (Akaike, 1974) is $-2 \log L + 2k$, the BIC (Schwarz, 1978) is $-2 \log L + k \log n$, and the CAIC (Bozdogan, 1987) is $-2 \log L + (1 + \log n)k$, where L is the likelihood, k is the number of parameters, and n is the number of observations. The criteria give increasingly large penalties in k and n .

³⁹These models are *overlapping*, because although neither nests the other, when $\omega = \rho = 0$ they are equivalent. To implement Vuong’s (1989) two step test for overlapping models, I first reject the null hypothesis that the models are equivalent. In the present case, such rejection is immediate because ω differs significantly from zero both by t -test and LR test (Vuong, 1989, footnote 6). In the second step, the two models are discriminated based on their

most appropriate model in this application.

5 Conclusion

The FPS model provides a useful alternative to Lee’s model for continuous selected variables, provided the correlation exhibited by the data is not too high. If higher correlation is required, one option would be to replace the FGM distribution in the FPS with a generalized FGM distribution (Johnson and Kotz, 1975) that allows more correlation, although I have not explored that alternative here. The FPS model is more flexible than alternatives based on the bivariate normal distribution (i.e., the BNDS model). Although Lee’s model allows for more correlation between the selection disturbance and the selected variable, the FPS model may provide a better fit to the data, as the application in the previous section demonstrated.

Although I have not directly compared the FPS model to semiparametric approaches, the ability of the FPS model to incorporate any functional form gives it much flexibility. For example, $f_{Y^*}(y^*)$ could take the “semi-nonparametric” series expansion form of Gallant and Nychka (1987). Such methods blur the distinction between parametric and semiparametric inference and lend an arbitrary amount of flexibility to maximum likelihood estimation. Another future research avenue is the application of the FPS distribution to other sample selection problems, such as when D is not binary but incidentally censored (Amemiya’s (1985) Type 3 Tobit model) or the non-random assignment treatment effects model (also termed the endogenous dummy variable model). Other possible applications include any requiring bivariate distributions, such as selection in count data models,⁴⁰ bivariate count models, bivariate multinomial choice problems, and the like.

Kullback-Leibler information content. By this metric and using Vuong’s terminology, the FPS model is (statistically significantly) *better* than Lee’s model (p -value: 0.007).

⁴⁰See Prieger (2000*b*) for the extension of the FGM distribution to count data.

Appendix: The Endogeneity of Health Insurance Choice

This appendix contains the details of the exogeneity testing mentioned in section 4.3 performed on the health insurance choice variables. For the selection equation for hospitalization, we have a probit equation with potentially endogenous dummy variables. Given the difficulties with estimating multivariate probit models of high order, I collapse all the insurance indicators into a single variable (INSURED), an indicator for insurance of any type. The latent model for insurance choice and hospitalization is (1), (3), and

$$l_i^* = \delta' w_i - \eta_i \quad (20)$$

$$\iota_i = 1\{l_i^* > 0\} \quad (21)$$

where ι_i is the indicator INSURED, ι_i is included in z_i in (1), and (ε_i, η_i) is mean-zero bivariate normal with unit variances and correlation ρ . This is model 6 of Maddala (1983, p.122), who provides the likelihood for MLE. If $\rho \neq 0$, then INSURED is endogenous and identification of α in (1) requires that w_i include at least one instrument not in z_i . I use two instruments: indicators measuring whether insurance was offered to the individual through a current (OFFINSCUR) or previous (OFFINSPREV) job (regardless of whether the individual procured insurance through the offer). When OFFINSCUR and OFFINSPREV are included in z_i in a univariate probit estimation of (1) and (3) (results not reported), neither is significant, corroboration that the exclusion restriction is valid. Two estimations are reported in Table 9. In the first column, ρ is fixed at zero. In the second column, ρ is allowed to vary and the estimates are MLE of the system of equations: (1), (3), (20) and (21). The coefficient on INSURED is a bit lower in the joint estimation, although still positive. In both estimations, OFFINSCUR and OFFINSPREV are highly significant determinants of INSURED. Neither a LR test (p -value = 0.40) nor a t -test for $\rho = 0$ (p -value = 0.40) reject the hypothesis that INSURED is exogenous.

For the length of stay estimation, we have a gamma distributed variable (the sum of all exponentially-distributed hospital stays for an individual) with potentially endogenous dummy

variables ν_i . In these estimations ν_i is a vector comprising PRIVINS, MEDICAID, MEDICARE, HMOPRIV, HMOMCARE, HMOMCAID, PRIVMCARE, PRIVMCAID, and MCAREMCAID. An appropriate transformation of the dependent variable yields a linear equation:

$$\log(y_i) - \psi(n_i) = \beta_1'x_i + \beta_2'\nu_i + \nu_i, \quad (22)$$

where ψ is the digamma function, n_i is as in (19), and ν_i is a mean-zero random variable with variance $\psi'(n_i)$ (Johnson, Kotz and Balakrishnan, 1995, p.382-3), where ψ' is the trigamma function. The exogeneity testing here is via a Hausman test. Under the null hypothesis that ν_i is exogenous, $\hat{\theta} = (\hat{\beta}_1, \hat{\beta}_2)$, the ML estimates based on (22), are consistent and efficient. These estimates have already been reported in the bottom half of column one in Table 5. Under the null and the alternative hypotheses that ν_i is endogenous, a two-stage IV estimation (resulting in estimates $\tilde{\theta}$) is consistent. The IV estimation is OLS based on (22), where $\hat{\nu}_i$, the expected value of ν_i resulting from single-equation probit estimations, replaces ν_i (Maddala, 1983, pp.120-2). The first stage estimation was performed both with and without the extra OFFINSCUR and OFFINSPREV instruments. The second stage estimations are reported in Table 10. Of the significant insurance coefficients from Table 5, all have the same sign in the IV estimation (except HMOPRIV in the *without extra instruments* estimation). Since the second stage estimation involves an embedded estimator, the estimate of the variance, \tilde{V} (robust to heteroskedasticity), is incorrect. If \tilde{V} were correct, then $H_N = (\hat{\theta} - \tilde{\theta})'(\tilde{V} - \hat{V})^{-1}(\hat{\theta} - \tilde{\theta})$ would have a $\chi^2(\text{rank}(\tilde{V} - \hat{V}))$ asymptotic distribution. Since critical values based on the χ^2 distribution are not appropriate for H_N even asymptotically because the statistic is not appropriately studentized, I bootstrap the test statistic (999 iterations). In this case the bootstrap is applied not for asymptotic refinement but merely to obtain critical values correct to the usual $O(N^{-1/2})$ (Beran, 1987). When referred to the bootstrap critical values, neither test rejects the hypothesis that the insurance variables are exogenous.

References

- Akaike, Hirotugu (1974), 'A New Look at the Statistical Identification Model', *IEEE Transactions on Automatic Control* **19**, 716–723.
- Amemiya, Takeshi (1985), *Advanced Econometrics*, Cambridge: Harvard University Press.
- American Academy of Actuaries (1996), *Medicare Managed Care: Savings, Access, and Quality*, AAA, Washington, D.C.
- Beran, Rudolf (1987), 'Prepivoting to Reduce Level Error of Confidence Sets', *Biometrika* **74**(3), 457–468.
- Bozdogan, Hamparsum (1987), 'Model Selection and Akaike's Information Criterion (AIC): The General Theory and its Analytical Extensions', *Psychometrika* **52**(3), 345–370.
- Cameron, A. Colin, Trivedi, Pravin K., Milne, Frank and Piggott, J. (1988), 'A Microeconomic Model of the Demand for Health Care and Health Insurance in Australia', *Review of Economic Studies* **55**(1), 85–106.
- Cox, David R. (1972), 'Regression Models and Life-Tables', *Journal of the Royal Statistical Society, Series B* **34**, 187–202.
- Cox, David R. (1975), 'Partial Likelihood', *Biometrika* **62**, 269–276.
- Deb, Partha and Trivedi, Pravin K. (1997), 'Demand for Medical Care by the Elderly: A Finite Mixture Approach', *Journal of Applied Econometrics* **12**(3), 313–336.
- Donald, Stephen G. (1995), 'Two-Step Estimation of Heteroskedastic Sample Selection Models', *Journal of Econometrics* **65**(2), 347–380.

- Dor, Avi and Farley, Dean E. (1996), 'Payment Source and the Cost of Hospital Care: Evidence from a Multiproduct Cost Function With Multiple Payers', *Journal of Health Economics* **15**(1), 453–481.
- Duncan, Gregory M. (1986), 'Continuous/Discrete Econometric Models with Unspecified Error Distribution', *Journal of Econometrics* **36**, 1.
- Ettner, Susan L. (1997), 'Adverse Selection and the Purchase of Medigap Insurance by the Elderly', *Journal of Health Economics* **16**(5), 543–562.
- Fleming, Thomas R. and Harrington, David P. (1984), 'Nonparametric Estimation of the Survival Distribution in Censored Data', *Communications in Statistics: Theory and Methods* **13**(20), 2469–2486.
- Frank, Richard G. and Lave, Judith R. (1989), 'A Comparison of Hospital Responses to Reimbursement Policies for Medicaid Psychiatric Patients', *Rand Journal of Economics* **20**(4), 588–600.
- Gallant, A. Ronald and Nychka, Douglas W. (1987), 'Semi-Nonparametric Maximum Likelihood Estimation', *Econometrica* **55**(2), 363–390.
- Gardner, Jonathan (1997), 'No Plans to Boost Coverage', *Modern Healthcare* p. 24, December 15.
- Gilleskie, Donna B. (1998), 'A Dynamic Stochastic Model of Medical Care Use and Work Absence', *Econometrica* **66**(1), 1–45.
- Gronau, Reuben (1974), 'Wage Comparisons—A Selectivity Bias', *Journal of Political Economy* **82**(6), 1119–1143.
- Heckman, James J. (1974), 'Shadow Prices, Market Wages, and Labor Supply', *Econometrica* **42**(4), 679–694.

- Heckman, James J. (1976), ‘The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models’, *Annals of Economic and Social Measurement* **5**, 475–492.
- Johnson, Norman L. and Kotz, Samuel (1975), ‘On some Generalized Farlie-Gumbel-Morgenstern Distributions’, *Communications in Statistics* **4**, 415–427.
- Johnson, Norman L. and Kotz, Samuel (1977), ‘On some Generalized Farlie-Gumbel-Morgenstern Distributions — II: Regression, Correlation and Further Generalizations’, *Communications in Statistics A: Theory and Methods* **A6**(6), 485–496.
- Johnson, Norman L., Kotz, Samuel and Balakrishnan, N. (1995), *Continuous Univariate Distributions*, Vol. 1 of *Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics section.*, 2nd edn, New York: John Wiley & Sons.
- Kimeldorf, George and Sampson, Allan R. (1975a), ‘One Parameter Families of Bivariate Distributions with Fixed Marginals’, *Communications in Statistics* **4**(3), 293–301.
- Kimeldorf, George and Sampson, Allan R. (1975b), ‘Uniform Representations of Bivariate Distributions’, *Communications in Statistics* **4**, 617–627.
- Kotz, Samuel, Balakrishnan, N. and Johnson, Norman L. (2000), *Continuous Multivariate Distributions: Volume 1: Models and Applications*, Wiley Series in Probability and Statistics: Applied Probability and Statistics Section, John Wiley and Sons, New York.
- Kruskal, William H. (1958), ‘Ordinal Measures of Association’, *Journal of the American Statistical Association* **53**(284), 814–861.
- Lancaster, Tony (1985), ‘Generalised Residuals and Heterogeneous Duration Models: With Applications to the Weibull Model’, *Journal of Econometrics* **28**(1), 155–169.

- Lee, Lung-Fei (1983), ‘Generalized Econometric Models with Selectivity’, *Econometrica* **51**(2), 507–512.
- Levinson, Arik and Ullman, Frank (1996), ‘Medicaid Managed Care and Infant Health’, *Journal of Health Economics* **17**(3), 351–368.
- Maddala, G. S. (1983), *Limited-Dependent and Qualitative Variables in Econometrics*, Econometric Society Monographs, Cambridge University Press, Cambridge.
- Madrian, Brigitte C. (1994), ‘Employment-Based Health Insurance and Job Mobility: Is There Evidence of Job-Lock?’, *Quarterly Journal of Economics* **109**(1), 27–54.
- Manski, Charles F. (1989), ‘Anatomy of the Selection Problem’, *Journal of Human Resources* **24**(3), 343–360.
- Manski, Charles F. (1990), ‘Nonparametric Bounds on Treatment Effects’, *American Economic Review* **80**(2), 319–323.
- Mardia, Kanti V. (1970a), *Families of Bivariate Distributions*, Vol. 27 of *Griffin’s Statistical Monographs & Courses*, London: Charles Griffin.
- Mardia, Kanti V. (1970b), ‘A Translation Family of Bivariate Distributions and Fréchet’s Bounds’, *Sankhyā A* **32**, 119–122.
- Nataf, A. (1962), ‘Déterminations Des Distributions de Probabilités Dont Les Marges Sont Données’, *Comptes Rendus des Séances de l’Académie des Sciences* **255**, 42–3.
- Newey, Whitney K., Powell, James L. and Walker, James R. (1990), ‘Semiparametric Estimation of Selection Models: Some Empirical Results’, *American Economic Review* **80**(2), 324–328.
- Prieger, James E. (2000a), Conditional Moment Tests for Parametric Duration Models, Working Paper 00-10, University of California, Davis.

- Prieger, James E. (2000*b*), A Generalized Parametric Selection Model for Non-Normal Data, Working Paper 00-09, University of California, Davis, Department of Economics.
- Rosenman, Robert E. (1993), 'Health Plan Effects on Inpatient Resource Use: Some Contrary Evidence About IPAs', *The Journal of Socio-Economics* **22**, 131–140.
- Schwarz, Gideon (1978), 'Estimating the Dimension of a Model', *Annals of Statistics* **6**(2), 461–464.
- Sharma, Sunil (1992), 'On Specification Diagnostics for Econometric Models of Durations', *Journal of Quantitative Economics* **8**(2), 285–307.
- Vella, Francis (1998), 'Estimating Models with Sample Selection Bias: A Survey', *Journal of Human Resources* **33**(1), 127–169.
- Vuong, Quang H. (1989), 'Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses', *Econometrica* **57**(2), 307–333, March.
- Welch, W. P. (1985), 'Health Care Utilization in HMO's: Results from Two National Samples', *Journal of Health Economics* **4**(4), 293–308.

Distribution	PDF $f_{Y^*}(y^* \beta'x, \theta)$	CDF $F_{Y^*}(y^* \beta'x, \theta)$	Mean $\mu(\beta'x, \theta)$	Variance $\zeta^2(\beta'x, \theta)$
Exponential	$\lambda e^{-\lambda y^*}$	$1 - e^{-\lambda y^*}$	$1/\lambda$	$1/\lambda^2$
Gamma	$\frac{(\lambda y^*)^\gamma \exp(-\lambda y^*)}{y^* \Gamma(\gamma)}$	$\Gamma_{\lambda y^*}(\gamma)/\Gamma(\gamma)$	γ/λ	γ/λ^2
Log-logistic	$g(y^*) \left[\frac{1}{1+(\lambda y^*)^{1/\gamma}} \right]^2$	$1 - \frac{1}{1+(\lambda y^*)^{1/\gamma}}$	$\gamma \pi \csc(\pi\gamma) / \lambda$	$2\gamma \pi \csc(2\pi\gamma) / \lambda^2 - \mu^2$
Lognormal	$\frac{1}{\sigma y^*} \phi\left(\frac{\log y^* - \beta'x}{\sigma}\right)$	$\Phi\left(\frac{\log y^* - \beta'x}{\sigma}\right)$	$e^{\frac{1}{2}\sigma^2} / \lambda$	$(e^{2\sigma^2} - e^{\sigma^2}) / \lambda^2$
Weibull	$g(y^*) e^{-(\lambda y^*)^{1/\gamma}}$	$1 - e^{-(\lambda y^*)^{1/\gamma}}$	$\gamma \Gamma(\gamma) / \lambda$	$2\gamma \Gamma(2\gamma) / \lambda^2 - \mu^2$

For all models, $\lambda = e^{-\beta'x}$. ϕ and Φ are the pdf and cdf, resp., of the standard normal distribution, Γ is the gamma function, $\Gamma_c(\gamma) = \int_0^c t^{\gamma-1} e^{-t} dt$ is the incomplete gamma function, and $g(y^*) = (\gamma y^*)^{-1} (\lambda y^*)^{1/\gamma}$.

Table 1: Duration Distributions

Distribution	Logit Selection		Probit Selection	
	<i>FPS model</i>	<i>Fréchet bound</i>	<i>FPS model</i>	<i>Fréchet bound</i>
Exponential	0.276 ω	0.907	0.282 ω	0.905
Gamma				
$\gamma = 0.05$	0.115 ω	0.371	0.118 ω	0.312
$\gamma = 0.25$	0.210 ω	0.677	0.215 ω	0.660
$\gamma = 1$	0.276 ω	0.900	0.282 ω	0.899
$\gamma = 2$	0.292 ω	0.947	0.299 ω	0.948
Log-logistic				
$\lim_{\downarrow 0} \gamma$	0.304 ω	1.000	0.311 ω	1.000
$\gamma = 0.1$	0.297 ω	0.985	0.304 ω	0.980
$\gamma = 0.25$	0.264 ω	0.902	0.270 ω	0.887
$\gamma = 0.4$	0.183 ω	0.667	0.187 ω	0.642
$\lim_{\uparrow 0.5} \gamma$	0.000	0.000	0.000	0.000
Weibull				
$\gamma = 0.25$	0.313 ω	0.997	0.320 ω	0.999
$\gamma = 0.5$	0.309 ω	0.980	0.316 ω	0.987
$\gamma = 1$	0.276 ω	0.907	0.282 ω	0.903
$\gamma = 2$	0.185 ω	0.702	0.189 ω	0.670
$\gamma = 5$	0.034 ω	0.208	0.034 ω	0.171

Table note: the Fréchet bound is the theoretical maximum correlation possible between the two specified random variables.

Table 2: Allowed Correlation for the FPS Duration Models

	Insured	Private Insurance	Medicare	Medicaid
Insured	12,524			
Private Insurance	10,224	10,224		
Medicare	2,736	1,482	2,736	
Medicaid	1,345	96	391	1,345
HMO	5,787	5,129	598	519

Table note: Cell entries are the number of individuals (out of 15,692 observations) who have insurance of both the column and the row type. Categories are not mutually exclusive. An additional 150 persons were insured but not by Medicare, Medicaid, or private sources.

Table 3: MEPS Insurance Data

Variable	Description	Mean	SD
HOSPSTAY	Binary variable: 1=individual had hospital stay	0.09	0.29
HOSPDUR	Length of all hospitalizations, given HOSPSTAY=1	7.43	11.36
HOSPNUM	Number of hospital stays, given HOSPSTAY=1	1.42	0.85
ADLHELP	1 = requires assistance with daily living tasks	0.04	0.20
AGE	Age	44.40	17.31
BLACK	1 = black (not hispanic)	0.12	0.33
CONDN	Number of self-reported medical conditions	1.68	1.91
EDUC	Years of education	12.38	3.16
EMPLOYED	Employment status: 1=currently employed	0.65	0.48
EXCLHLTH	1 = individual reports health to be “excellent”	0.29	0.45
FEMALE	1 = female	0.54	0.50
HISPANIC	1 = of hispanic ethnicity	0.18	0.38
HMOMCAID	1 = enrolled in a HMO and covered by Medicaid	0.03	0.18
HMOMCARE	1 = enrolled in a HMO and covered by Medicare	0.04	0.19
HMOPRIV	1 = enrolled in a HMO and covered by private insurance	0.33	0.47
MARRIED	Marital status: 1 = currently married	0.57	0.49
MCAREMCAID	1 = currently covered by Medicaid and Medicare	0.02	0.16
MEDICAID	1 = currently covered by Medicaid	0.09	0.28
MEDICARE	1 = currently covered by Medicare	0.17	0.38
MIDWEST	Regional indicator (EAST is the excluded dummy)	0.22	0.42
POORHLTH	1 = individual reports health to be “poor”	0.04	0.20
PRIOLIST	Number of conditions on the priority list	0.54	1.00
PRIVINS	1 = covered by private insurance of any type	0.66	0.47
PRIVMCAID	1 = covered by private insurance and Medicaid	0.01	0.08
PRIVMCARE	1 = covered by private insurance and Medicare	0.10	0.29
SOUTH	Regional indicator (EAST is the excluded dummy)	0.35	0.48
WEST	Regional indicator (EAST is the excluded dummy)	0.23	0.42

Table note: all hospitalization variables are for 1996.

Table 4: MEPS Data: Variable Definitions and Summary Statistics

Variable	Baseline Model		FPS Model		Lee's Model	
	Estimate	s.e.	Estimate	s.e.	Estimate	s.e.
<i>Probit Selection</i>						
CONSTANT	-1.697	(0.108)***	-1.702	(0.108)***	-1.697	(0.113)***
PRIVINS	0.178	(0.054)***	0.177	(0.054)***	0.178	(0.055)***
MEDICARE	0.440	(0.081)***	0.441	(0.081)***	0.440	(0.080)***
MEDICAID	0.452	(0.078)***	0.452	(0.078)***	0.452	(0.079)***
HMOPRIV	0.028	(0.042)	0.027	(0.042)	0.028	(0.043)
HMOMCARE	0.011	(0.077)	0.015	(0.077)	0.011	(0.078)
HMOMCAID	-0.032	(0.092)	-0.033	(0.092)	-0.032	(0.092)
PRIVMCARE	-0.176	(0.080)**	-0.179	(0.080)**	-0.176	(0.081)**
PRIVMCAID	0.403	(0.157)**	0.395	(0.157)**	0.403	(0.163)**
MCAREMCAID	-0.524	(0.111)***	-0.532	(0.111)***	-0.524	(0.111)***
CONDN	0.081	(0.010)***	0.080	(0.010)***	0.081	(0.010)***
PRIOLIST	0.062	(0.019)***	0.063	(0.019)***	0.062	(0.019)***
EXCLHLTH	-0.143	(0.039)***	-0.143	(0.039)***	-0.143	(0.040)***
POORHLTH	0.177	(0.064)***	0.174	(0.063)***	0.177	(0.061)***
ADLHELP	0.373	(0.062)***	0.380	(0.062)***	0.373	(0.061)***
Demographic variables	included		included		included	
<i>Exponential Durations</i>						
CONSTANT	1.316	(0.136)***	1.712	(0.182)***	1.314	(0.144)***
PRIVINS	-0.064	(0.074)	-0.128	(0.094)	-0.064	(0.075)
MEDICARE	-0.286	(0.095)***	-0.302	(0.121)**	-0.285	(0.097)***
MEDICAID	-0.230	(0.113)**	-0.261	(0.121)**	-0.229	(0.116)**
HMOPRIV	-0.274	(0.053)***	-0.217	(0.072)***	-0.274	(0.054)***
HMOMCARE	0.131	(0.079)*	0.063	(0.103)	0.131	(0.080)
HMOMCAID	0.008	(0.118)	0.030	(0.130)	0.008	(0.120)
PRIVMCARE	0.083	(0.093)	0.109	(0.119)	0.083	(0.095)
PRIVMCAID	0.729	(0.253)***	0.720	(0.191)***	0.729	(0.255)***
MCAREMCAID	0.196	(0.130)	0.217	(0.149)	0.195	(0.132)
CONDN	0.004	(0.009)	-0.009	(0.013)	0.004	(0.010)
PRIOLIST	0.005	(0.017)	-0.006	(0.023)	0.005	(0.018)
EXCLHLTH	0.027	(0.055)	0.001	(0.074)	0.027	(0.057)
POORHLTH	0.192	(0.062)***	0.128	(0.072)*	0.192	(0.063)***
ADLHELP	0.247	(0.059)***	0.185	(0.073)**	0.247	(0.060)***
Demographic variables	included		included		included	
Corr. parameter (ω or ρ)			-0.873	(0.052)***	0.001	(0.040)
Log likelihood	-7674.24		-7641.4		-7674.25	

N = 14,946. * indicates 10% level significance, ** 5% level significance, and *** 1% level significance. All estimates are MLE. See Table 6 for the demographic coefficients, Table 7 for the marginal effects, and Table 8 for model selection criteria.

Table 5: Hospitalization Incidence and Duration: Estimation Results

Variable	Baseline Model		FPS Model		Lee's Model	
	Estimate	s.e.	Estimate	s.e.	Estimate	s.e.
<i>Probit Selection</i>						
MIDWEST	0.023	(0.047)	0.025	(0.047)	0.023	(0.048)
SOUTH	0.007	(0.043)	0.008	(0.043)	0.007	(0.044)
WEST	-0.083	(0.048)*	-0.080	(0.048)*	-0.083	(0.049)*
FEMALE	0.154	(0.032)***	0.153	(0.032)***	0.154	(0.034)***
AGE	-0.002	(0.001)	-0.001	(0.001)	-0.002	(0.001)
BLACK	-0.041	(0.050)	-0.039	(0.050)	-0.040	(0.051)
HISPANIC	0.083	(0.046)*	0.082	(0.046)*	0.083	(0.046)*
EDUC	-0.004	(0.005)	-0.004	(0.005)	-0.004	(0.005)
MARRIED	0.069	(0.033)**	0.067	(0.033)**	0.069	(0.034)**
EMPLOYED	-0.197	(0.040)***	-0.198	(0.040)***	-0.197	(0.040)***
<i>Exponential Durations</i>						
MIDWEST	-0.086	(0.057)	-0.063	(0.074)	-0.086	(0.058)
SOUTH	-0.049	(0.050)	-0.030	(0.068)	-0.049	(0.051)
WEST	-0.277	(0.063)***	-0.252	(0.080)***	-0.277	(0.064)***
FEMALE	-0.252	(0.041)***	-0.198	(0.053)***	-0.252	(0.041)***
AGE	0.013	(0.001)***	0.013	(0.002)***	0.013	(0.002)***
BLACK	0.272	(0.063)***	0.259	(0.081)***	0.272	(0.064)***
HISPANIC	0.109	(0.061)*	0.103	(0.076)	0.109	(0.062)*
EDUC	-0.002	(0.006)	-0.001	(0.008)	-0.002	(0.006)
MARRIED	-0.189	(0.041)***	-0.190	(0.053)***	-0.189	(0.041)***
EMPLOYED	-0.130	(0.051)**	-0.089	(0.066)	-0.130	(0.052)**

Notes: this table is a continuation of Table 5, which contains the other coefficients. See notes to Table 5.

Table 6: Hospitalization Incidence and Duration: Estimation Results (Demographic Variables)

Variable	Probit Slope or Δ in $\Pr(d = 1)$		Duration Slope or Δ in $E(y)$	
	<i>Baseline</i>	<i>FPS</i>	<i>Baseline</i>	<i>FPS</i>
	PRIVINS	0.023	0.023	-0.276
MEDICARE	0.048	0.048	-1.377	-2.157
MEDICAID	0.049	0.049	-1.074	-1.826
HMOPRIV	0.004	0.004	-1.311	-1.483
HMOMCARE	0.002	0.002	0.510	0.374
HMOMCAID	-0.005	-0.005	0.032	0.181
PRIVMCARE	-0.028	-0.029	0.333	0.634
PRIVMCAID	0.045	0.045	2.154	3.145
MCAREMCAID	-0.102	-0.104	0.740	1.196
CONDN	0.012	0.012	0.016	-0.055
PRIOLIST	0.009	0.009	0.021	-0.034
EXCLHLTH	-0.023	-0.023	0.111	0.006
POORHLTH	0.023	0.023	0.726	0.737
ADLHELP	0.043	0.043	0.910	1.036
MIDWEST	0.003	0.004	-0.373	-0.400
SOUTH	0.001	0.001	-0.211	-0.187
WEST	-0.013	-0.012	-1.327	-1.758
FEMALE	0.020	0.020	-1.192	-1.342
AGE	0.000	0.000	0.055	0.080
BLACK	-0.006	-0.006	0.992	1.397
HISPANIC	0.011	0.011	0.428	0.602
EDUC	-0.001	-0.001	-0.008	-0.005
MARRIED	0.010	0.009	-0.868	-1.281
EMPLOYED	-0.032	-0.032	-0.577	-0.571
Predicted Mean	0.090	0.090	4.163	6.128

Notes: the figures are slopes (i.e., the derivative of the predicted conditional mean $[E(d|z) = \Phi(\alpha'z)$ for probit, $E(y|x, z) = \exp(\beta'x)$ for duration] with respect to the covariate) for continuous covariates, and discrete changes in the conditional mean for indicator variables (i.e., the change in the conditional mean when the indicator takes the value 1). Duration units are days. Reported figures are averages in the entire sample of 14,946 individuals.

Table 7: Hospitalization Incidence and Duration: Marginal Effects of Covariates and Predicted Means

Selection Criterion	Baseline Model	FPS Model	Lee's Model
Log likelihood	-7674.3	-7641.4	-7674.3
Parameters	50.0	51.0	51.0
Observations	14946	14946.0	14946.0
AIC	15448.5	15384.8*	15450.5
BIC	15829.1	15773.1*	15838.7
CAIC	15879.1	15824.1*	15889.7
Vuong's test		preferred*	not preferred

Table notes: * denotes preferred model by a particular criterion. See footnotes 38 and 39.

Table 8: Hospitalization Incidence and Duration: Model Selection Criteria

Variable	Single Equation Probit		Bivariate Probit	
	Estimate	s.e.	Estimate	s.e.
<i>Dep. variable: HOSPSTAY</i>				
INSURED	0.315	(0.050)***	0.220	(0.125)*
CONDN	0.078	(0.011)***	0.080	(0.011)***
PRIOLIST	0.074	(0.020)***	0.073	(0.019)***
EXCLHLTH	-0.147	(0.040)***	-0.144	(0.041)***
POORHLTH	0.153	(0.067)**	0.152	(0.064)**
ADLHELP	0.372	(0.064)***	0.379	(0.062)***
MIDWEST	0.016	(0.048)	0.015	(0.049)
SOUTH	0.012	(0.044)	0.007	(0.046)
WEST	-0.070	(0.050)	-0.073	(0.050)
FEMALE	0.159	(0.033)***	0.161	(0.035)***
AGE	0.000	(0.001)	0.000	(0.001)
BLACK	-0.015	(0.051)	-0.017	(0.051)
HISPANIC	0.107	(0.047)**	0.095	(0.049)*
EDUC	-0.011	(0.005)**	-0.010	(0.006)*
MARRIED	0.041	(0.033)	0.047	(0.034)
EMPLOYED	-0.291	(0.037)***	-0.286	(0.037)***
CONSTANT	-1.645	(0.107)***	-1.608	(0.120)***
<i>Dep. variable: INSURED</i>				
OFFINSCUR	1.426	(0.037)***	1.426	(0.037)***
OFFINSPREV	0.358	(0.051)***	0.357	(0.051)***
CONDN	0.111	(0.013)***	0.111	(0.011)***
PRIOLIST	-0.014	(0.026)	-0.013	(0.025)
EXCLHLTH	0.159	(0.034)***	0.159	(0.033)***
POORHLTH	-0.114	(0.079)	-0.115	(0.078)
ADLHELP	0.584	(0.103)***	0.585	(0.101)***
MIDWEST	-0.076	(0.048)	-0.077	(0.047)
SOUTH	-0.294	(0.042)***	-0.295	(0.042)***
WEST	-0.163	(0.046)***	-0.164	(0.045)***
FEMALE	0.129	(0.030)***	0.127	(0.030)***
AGE	0.020	(0.001)***	0.020	(0.001)***
BLACK	-0.197	(0.044)***	-0.197	(0.044)***
HISPANIC	-0.437	(0.038)***	-0.437	(0.039)***
EDUC	0.049	(0.005)***	0.049	(0.005)***
MARRIED	0.235	(0.031)***	0.234	(0.031)***
EMPLOYED	-0.482	(0.037)***	-0.482	(0.037)***
CONSTANT	-0.820	(0.092)***	-0.819	(0.094)***
Correlation parameter (ρ)	0	fixed	0.062	(0.073)
Log likelihood	-8660.85		-8660.50	

N = 14,041. * indicates 10% level significance, ** 5% level significance, and *** 1% level significance. All estimates are MLE.

Table 9: Estimation Results for Endogeneity Testing: Selection Equation

Variable	Second Stage IV Estimation			
	without extra instruments		with extra instruments	
	<i>estimate</i>	<i>s.e.</i>	<i>estimate</i>	<i>s.e.</i>
CONSTANT	1.608	(0.346)***	1.546	(0.353)***
$\widehat{\text{PRIVINS}}$	0.662	(0.866)	0.420	(0.391)
$\widehat{\text{MEDICARE}}$	-0.149	(0.270)	-0.080	(0.280)
$\widehat{\text{MEDICAID}}$	-0.373	(0.625)	-0.254	(0.524)
$\widehat{\text{HMOPRIV}}$	0.240	(1.062)	-0.463	(0.505)
$\widehat{\text{HMOMCARE}}$	-0.598	(0.697)	-1.138	(0.609)*
$\widehat{\text{HMOMCAID}}$	-0.202	(0.887)	-0.801	(0.780)
$\widehat{\text{PRIVMCARE}}$	0.397	(0.307)	0.377	(0.286)
$\widehat{\text{PRIVMCAID}}$	0.752	(3.644)	2.999	(2.858)
$\widehat{\text{MCAREMCAID}}$	0.403	(0.765)	0.799	(0.571)
CONDN	-0.034	(0.021)	-0.028	(0.016)*
PRIOLIST	0.022	(0.030)	0.024	(0.029)
EXCLHLTH	-0.171	(0.102)*	-0.116	(0.070)*
POORHLTH	0.182	(0.107)*	0.180	(0.096)*
ADLHELP	0.416	(0.166)**	0.294	(0.099)***
MIDWEST	-0.044	(0.093)	-0.117	(0.083)
SOUTH	0.082	(0.115)	0.040	(0.084)
WEST	-0.117	(0.141)	-0.066	(0.115)
FEMALE	-0.158	(0.068)**	-0.139	(0.059)**
AGE	0.006	(0.005)	0.006	(0.004)
BLACK	0.405	(0.153)***	0.291	(0.104)***
HISPANIC	0.219	(0.127)*	0.225	(0.092)**
EDUC	-0.035	(0.035)	-0.020	(0.011)*
MARRIED	-0.303	(0.200)	-0.117	(0.069)*
EMPLOYED	-0.403	(0.350)	-0.135	(0.114)
R^2		0.144		0.141

N = 1,346. * indicates 10% level significance, ** 5% level significance, and *** 1% level significance. The *with extra instruments* estimation includes the variables OFFINSCUR and OFFINSPREV as instruments in the first round probit estimations. Hatted variables are the estimated expected values from the first stage. S.e.'s (in parentheses) for the IV estimates are robust to heteroskedasticity but are not adjusted for the embedded estimators. See text for details.

Table 10: Estimation Results for Endogeneity Testing: Duration Equation

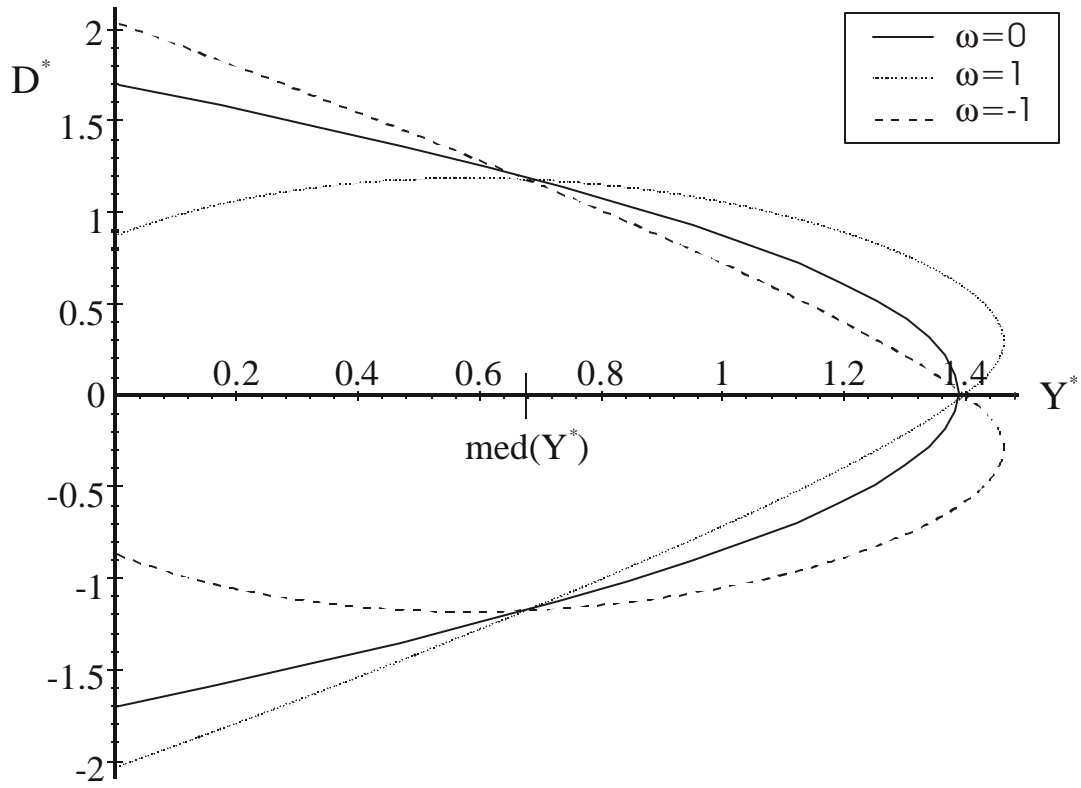


Figure 1: Isoprobability curves of the FGM distribution (D^* normal, Y^* exponential)

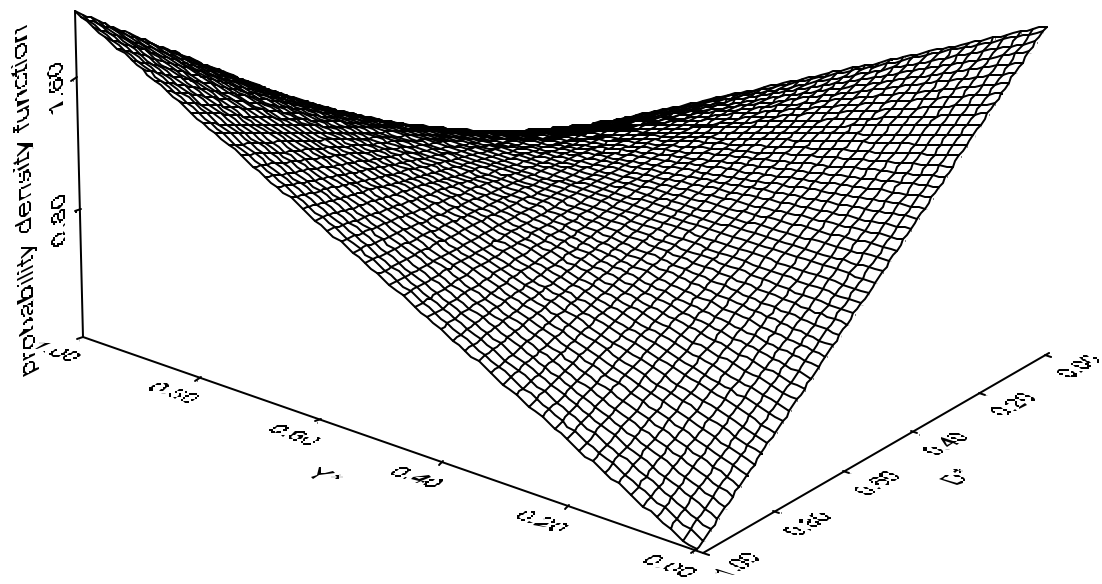


Figure 2: Uniform representation of the FGM distribution with positive correlation

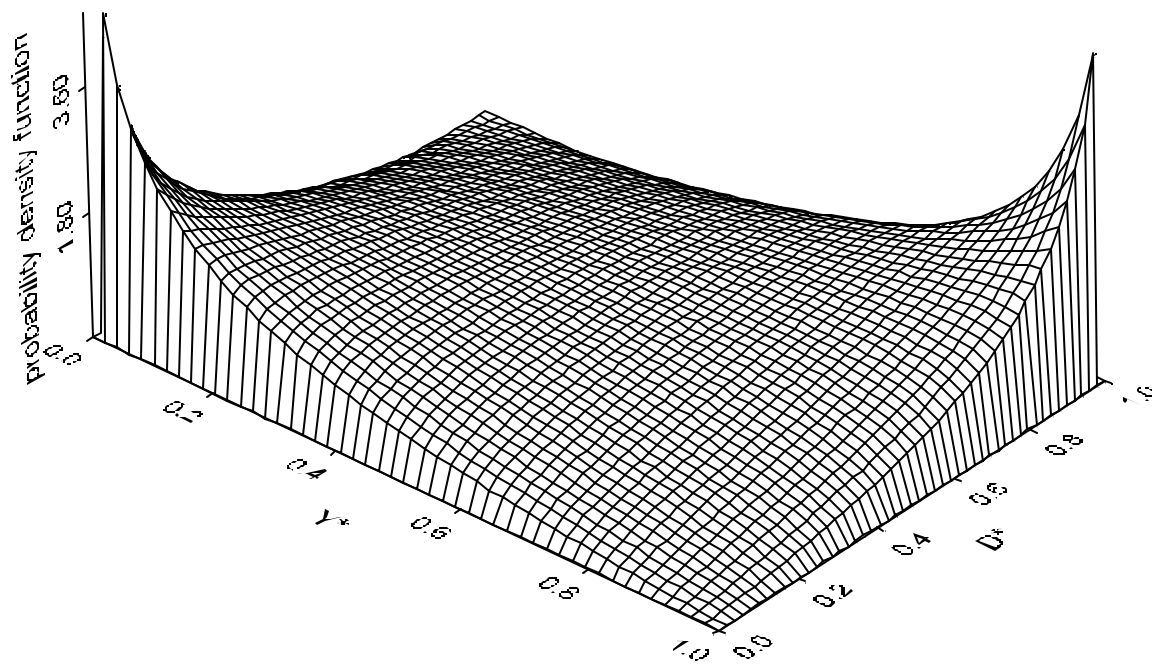


Figure 3: Uniform representation of the TBN distribution with positive correlation

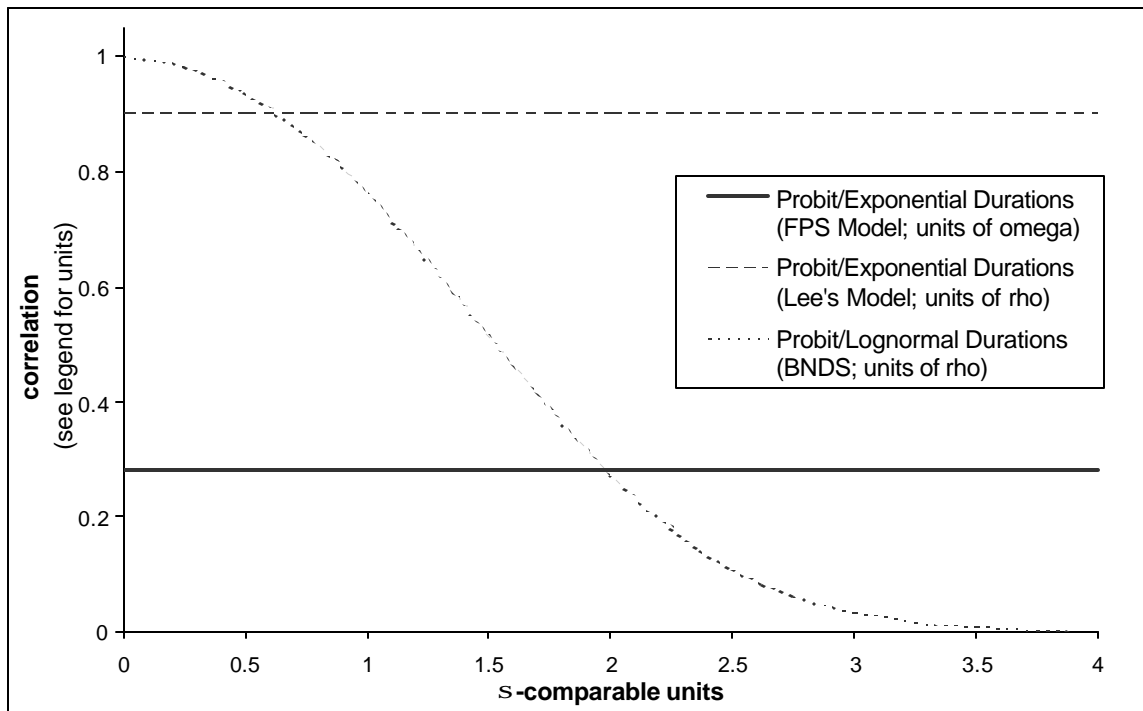


Figure 4: Comparison of allowed correlation in the exponential and BNDS duration selection models

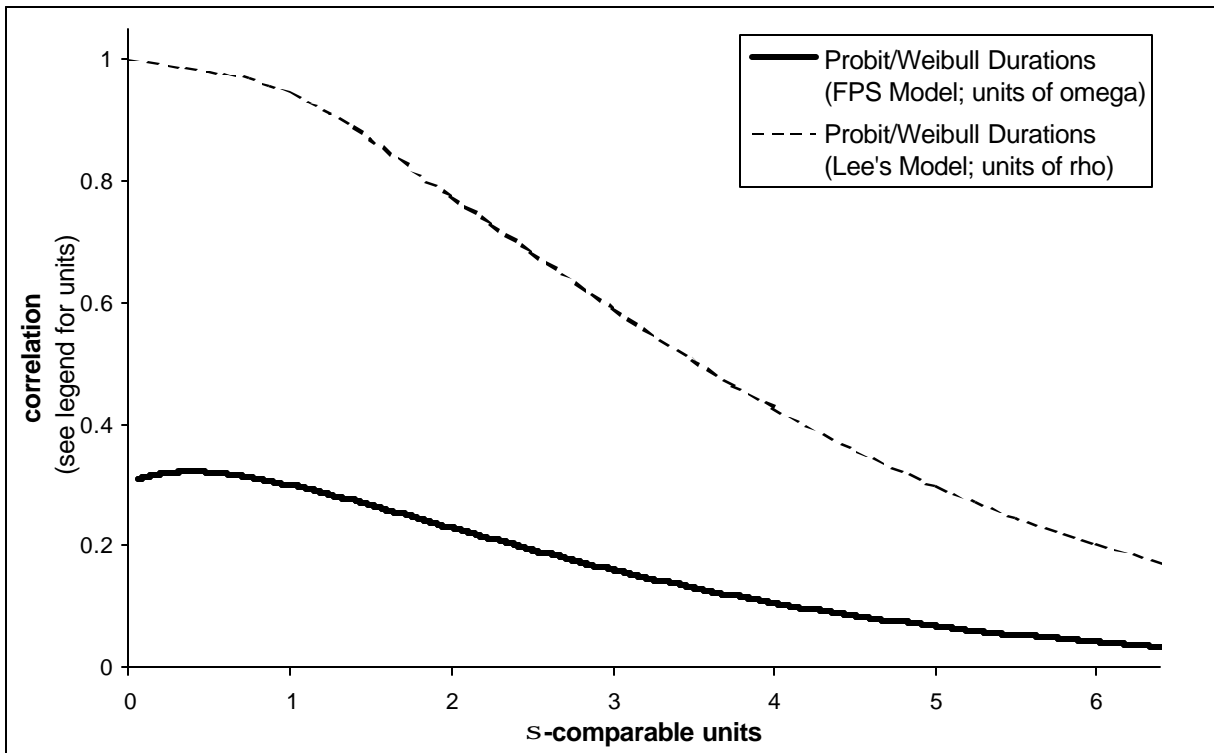


Figure 5: Comparison of allowed correlation in the Weibull duration selection models

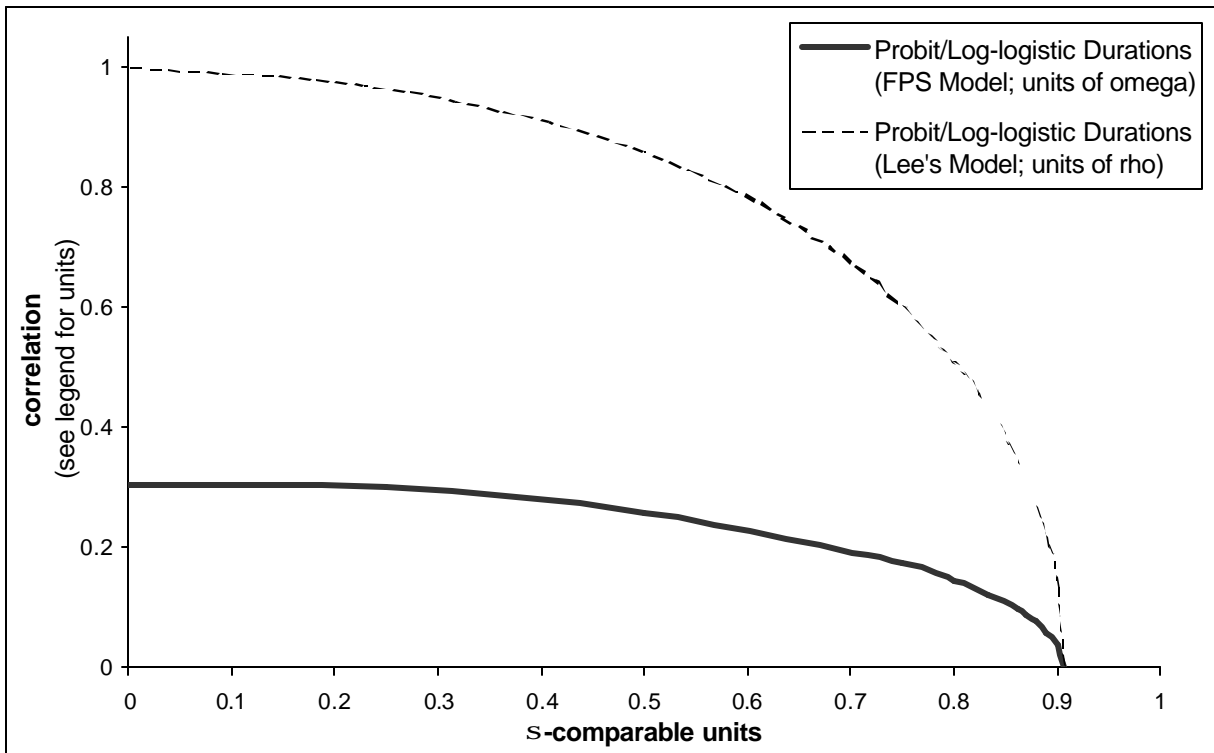


Figure 6: Comparison of allowed correlation in the log-logistic duration selection models

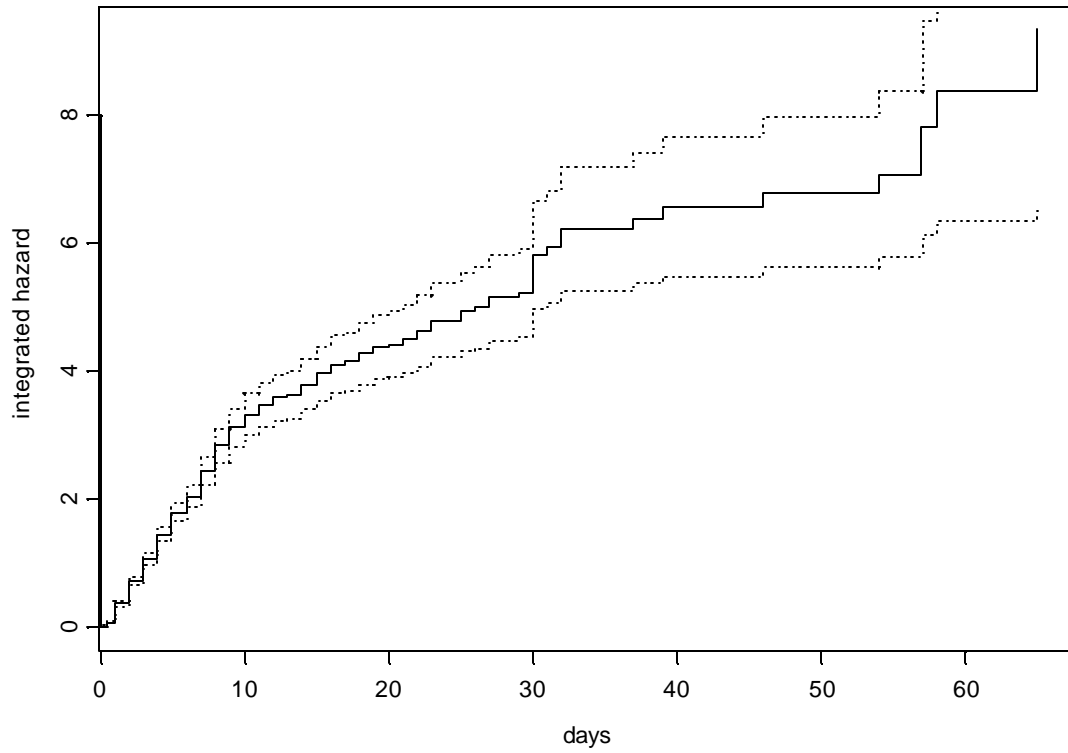


Figure 7: Duration of individual hospitalization spells: nonparametric estimate of the baseline integrated hazard (with 95% confidence band)