A Stochastic Frontier Model with Correction for Sample Selection

William Greene*
Department of Economics, Stern School of Business,
New York University,
March, 2008

Abstract

Heckman's (1979) sample selection model has been employed in three decades of applications of linear regression studies. The formal extension of the method to nonlinear models, however, is of more recent vintage. A generic solution for nonlinear models is proposed in Terza (1998). We have developed simulation based approach in Greene (2006). This paper builds on this framework to obtain a sample selection correction for the stochastic frontier model. We first show a surprisingly simple way to estimate the familiar normal-half normal stochastic frontier model (which has a closed form log likelihood) using maximum simulated likelihood. The next step is to extend the technique to a stochastic frontier model with sample selection. Here, the log likelihood does not exist in closed form, and has not previously been analyzed. We develop a simulation based estimation method for the stochastic frontier model. In an application that seems superficially obvious, the method is used to revisit the World Health Organization data [WHO (2000), Tandon et al. (2000)] where the sample partitioning is based on OECD membership. The original study pooled all 191 countries. The OECD members appear to be discretely different from the rest of the sample. We examine the difference in a sample selection framework.

JEL classification: C14; C23; C25

Keywords: Stochastic Frontier, sample Selection, Simulation, Efficiency

---

* 44 West 4th St., Rm. 7-78, New York, NY 10012, USA, Telephone: 001-212-998-0876; e-mail: wgreene@stern.nyu.edu, URL www.stern.nyu.edu/~wgreene.

1 Introduction

Heckman's (1979) sample selection model has been employed in three decades of applications of linear regression studies. The formal extension of the method to nonlinear models, however, is of more recent vintage. The familiar approach [i.e., add "lambda" to the equation – e.g., Bradford et al. (2000) in a stochastic frontier model and the earliest attempt to extend to nonlinear models, Wynand and van Praag (1981) in a probit model] is not appropriate for nonlinear models such as the stochastic frontier. A method of incorporating "selectivity" in some nonlinear models, notably the Poisson and other loglinear models, was proposed in Terza (1995, 1998) and applied in Greene (1995, 1997). The four studies were all based on either nonlinear least squares or on using quadrature to approximate an open form log likelihood function. We have suggested an alternative approach for a class of nonlinear models in Greene (2006) that relies on a simulation based estimator. The current work builds on this to obtain a sample selection correction for the stochastic frontier model.

We first show a surprisingly simple way to estimate the familiar normal-half normal stochastic frontier model using maximum simulated likelihood. The method bears some similarity to Bayesian treatments of the stochastic frontier model in that the inefficiency component of the composed error is treated as data, then conditioned out of the likelihood function. This particular step is of minor consequence, since the closed form of the log likelihood is already known. The next step, which is somewhat more complicated, is to extend the technique to a model of sample selection. Here, the log likelihood does not exist in closed form, and has not previously been analyzed. We develop a simulation based estimation method for the stochastic frontier model.

In an application that seems superficially obvious, the method is used to revisit the World Health Organization (2000) data [see also Tandon et. al (2000)]where the sample partitioning is based on OECD membership. The original study pooled all 191 countries (in a panel, albeit one with negligible within groups variation). The OECD members appear to be discretely different from the rest of the sample. We examine the difference in a sample selection framework.

2. A Selection Corrected Stochastic Frontier Model

The canonical form of the stochastic frontier model [Aigner, Lovell and Schmidt (1977)] (ALS) is specified with

$$y_i = \boldsymbol{\beta}'\mathbf{x}_i + v_i - u_i$$
$$\text{where } u_i = |\sigma_u U_i| = \sigma_u |U_i|, \; U_i \sim \text{N}[0,1^2],$$
$$v_i = \sigma_v V_i, \; V_i \sim \text{N}[0,1^2].$$

The stochastic frontier model is documented elsewhere, for example at length in ALS (1977) and Greene (2008a). (The utility of isolating the scaling of u_i and v_i will emerge shortly.) A vast literature has explored variations in the specification to accommodate, e.g., heteroscedasticity, panel data formulations, etc. [See, e.g., Greene (2008a) for a survey.] It will suffice for present purposes to work with the simplest form. Extensions will be considered later. The model can be estimated by modifications of least squares [e.g., Greene (2008a)], the generalized method of moments [Kopp and Mullahy (1990)] or, as conventional in the current literature, by maximum likelihood (ALS). [A spate of Bayesian applications has also appeared in the recent literature, e.g., Koop and Steel (2001).] In this study, we will suggest (as a means to another end), a fourth estimator, maximum simulated likelihood.

2.1 Maximum Likelihood Estimation of the Stochastic Frontier Model

The log likelihood for the normal-half normal model for a sample of *N* observations is

$$\log L(\boldsymbol{\beta},\sigma,\lambda) = \sum_{i=1}^{N}\left[\tfrac{1}{2}\log\left(\tfrac{2}{\pi}\right) - \log\sigma - \tfrac{1}{2}(\varepsilon_i/\sigma)^2 + \log\Phi(-\lambda\varepsilon_i/\sigma)\right]$$

where $\varepsilon_i = y_i - \boldsymbol{\beta}'\mathbf{x}_i = v_i - u_i$,
$\lambda = \sigma_u/\sigma_v$,
$\sigma = \sqrt{\sigma_v^2 + \sigma_u^2}$

and $\Phi(.)$ denotes the standard normal cdf. Details on estimation of the model can be found in ALS and elsewhere. Estimation is a straightforward enough problem to have been installed in the standard menu of supported techniques in a variety of programs including *LIMDEP*, *Stata* and *TSP*. The density satisfies the standard regularity conditions, and maximum likelihood estimation of the model is a conventional problem handled with familiar methods. [See, e.g., Econometric Software (2007) and Greene (2008b, Chapter 16).]

2.2 Maximum Simulated Likelihood Estimation

Based on the specification above, conditioned on u_i, the central equation of the model would be a classical linear regression model with normally distributed disturbances. Thus,

$$f(y_i|\mathbf{x}_i,U_i) = \frac{\exp[-\tfrac{1}{2}(y_i - \boldsymbol{\beta}'\mathbf{x}_i + \sigma_u|U_i|)^2/\sigma_v^2]}{\sigma_v\sqrt{2\pi}}.$$

The unconditional log likelihood for the model is obtained by first integrating the unobserved random variable, $|U_i|$, out of the conditional density, then summing the logs of the resulting unconditional densities. Thus,

$$f(y_i|\mathbf{x}_i) = \int_{|U_i|}\frac{\exp[-\tfrac{1}{2}(y_i - \boldsymbol{\beta}'\mathbf{x}_i + \sigma_u|U_i|)^2/\sigma_v^2]}{\sigma_v\sqrt{2\pi}}p(|U_i|)d|U_i|$$

where $p(|U_i|) = \dfrac{2\exp[-\tfrac{1}{2}|U_i|^2]}{\sqrt{2\pi}}$.

The closed form of the integral is in fact known. [See, e.g., ALS (1977).] (Its log was given earlier in the log likelihood function.) Consider using simulation to approximate the integration;

$$f(y_i|\mathbf{x}_i) \approx \frac{1}{R}\sum_{r=1}^{R}\frac{\exp[-\tfrac{1}{2}(y_i - \boldsymbol{\beta}'\mathbf{x}_i + \sigma_u|U_{ir}|)^2/\sigma_v^2]}{\sigma_v\sqrt{2\pi}},$$

where U_{ir} is a sequence of R random draws from the standard normal population. The simulated log likelihood is

$$\log L_S(\boldsymbol{\beta},\sigma_u,\sigma_v) = \sum_{i=1}^{N}\log\left\{\frac{1}{R}\sum_{r=1}^{R}\frac{\exp[-\frac{1}{2}(y_i-\boldsymbol{\beta}'\mathbf{x}_i+\sigma_u|U_{ir}|)^2/\sigma_v^2]}{\sigma_v\sqrt{2\pi}}\right\}$$

The maximum simulated likelihood estimators of the model parameters are obtained by maximizing this function with respect to the unknown parameters. The simplicity of the log likelihood for the linear regression model makes this straightforward. [See Gourieroux and Monfort (1996), Train (2003), Econometric Software (2007) and Greene (2008b).]

2.3 Sample Selection in a Linear Model

Heckman's (1979) generic sample selection model for the linear regression case is specified as

$d_i = 1[\boldsymbol{\alpha}'\mathbf{z}_i + w_i > 0]$, $w_i \sim N[0,1^2]$
$y_i = \boldsymbol{\beta}'\mathbf{x}_i + \varepsilon_i$, $\varepsilon_i \sim N[0,\sigma_\varepsilon^2]$
$(w_i,\varepsilon_i) \sim N_2[(0,1), (1, \rho\sigma_\varepsilon, \sigma_\varepsilon^2)]$
$(y_i,\mathbf{x}_i)$ observed only when $d_i = 1$.

[The selection model is documented elsewhere, e.g., Greene (2008b).] Two familiar methods have been developed for estimation of the model parameters. Heckman's (1979) two step method builds on the result

$$\begin{aligned}E[y_i|\mathbf{x}_i,d_i=1] &= \boldsymbol{\beta}'\mathbf{x}_i + E[\varepsilon_i|d_i=1]\\ &= \boldsymbol{\beta}'\mathbf{x}_i + \rho\sigma_\varepsilon\phi(\boldsymbol{\alpha}'\mathbf{z}_i)/\Phi(\boldsymbol{\alpha}'\mathbf{z}_i)\\ &= \boldsymbol{\beta}'\mathbf{x}_i + \theta\lambda_i.\end{aligned}$$

In the first step, $\boldsymbol{\alpha}$ in the probit equation is estimated by unconstrained single equation maximum likelihood and the inverse Mills ratio (IMR), $\hat{\lambda}_i = \phi(\hat{\boldsymbol{\alpha}}'\mathbf{z}_i)/\Phi(\hat{\boldsymbol{\alpha}}'\mathbf{z}_i)$ is computed for each observation, where ϕ is the standard normal density and Φ is the standard normal cdf. (We acknowledge the conflict in notation at this point. The use of the symbol λ for the two results given earlier is standard in both literatures. We will make distinction here by using λ without a subscript to denote the crucial parameter in the stochastic frontier model, $\lambda = \sigma_u/\sigma_v$, and λ_i with a subscript to indicate the inverse Mills ratio shown above for the selectivity model.) The second step in Heckman's procedure involves linear regression of y_i on the augmented regressor vector, $\mathbf{x}_i^* = (\mathbf{x}_i, \hat{\lambda}_i)$, using the observed subsample, with a correction of the OLS standard errors to account for the fact that an estimate of $\boldsymbol{\alpha}$ is used in the constructed regressor.

The maximum likelihood estimator for the same model is developed, e.g., in Maddala (1983). The log likelihood function for the same model is

$$\log L(\boldsymbol{\beta},\sigma_\varepsilon,\boldsymbol{\alpha},\rho) = \sum_{i=1}^{N}\log\left[d_i\frac{\exp\left(-\frac{1}{2}(\varepsilon_i^2/\sigma_\varepsilon^2)\right)}{\sigma_\varepsilon\sqrt{2\pi}}\Phi\left(\frac{(\rho\varepsilon_i/\sigma_\varepsilon)+\boldsymbol{\alpha}'\mathbf{z}_i}{\sqrt{1-\rho^2}}\right)+(1-d_i)\Phi(-\boldsymbol{\alpha}'\mathbf{z}_i)\right].$$

[See, e.g., Econometric Software (2007) for details.] This likewise has become a conventional, if relatively infrequently used estimator that is built into most contemporary software.

2.4 Sample Selection in Nonlinear Models

The received literature abounds with studies in which authors have ported Heckman's selectivity model to nonlinear settings, such as count data (e.g., Poisson), nonlinear regression, and binary choice models. The typical approach taken "to control for selection bias" is to fit the probit model as in the first step of Heckman's two step estimator, then append $\hat{\lambda}_i$ to the linear index part of the nonlinear model wherever it happens to appear. The first such application of this method was, in fact, the first application of the sample selection treatment in a nonlinear setting, Wynand and van Praag's (1981) development of a probit model for binary choice. The approach is in fact, inappropriate, as can be seen immediately in the specification. Note that $\hat{\lambda}_i$ arises as $E[\varepsilon_i|d_i=1]$ in a linear model. The expectation of some $g(\boldsymbol{\beta}'\mathbf{x}_i + \varepsilon_i)$ might well exist in the nonlinear setting as well, but it will not produce the form $E[g(\boldsymbol{\beta}'\mathbf{x}_i + \varepsilon_i)|d_i=1] = g(\boldsymbol{\beta}'\mathbf{x}_i + \lambda_i)$ which can then be imported back into the otherwise unchanged nonlinear model. [The precise expression, for example in a linear exponential model, is given in Terza (1995, 1998).] Indeed, in some cases, such as the probit and count data models, the ε_i for which the expectation given $d_i = 1$ is taken does not even appear in the original model; it is unclear as such what the "correction" is correcting. [Greene (1995, 1997) does the full development for the Poisson model by introducing ε_i into the Poisson mean as a form of latent heterogeneity.]

A second defect in the common strategy for nonlinear models is that the distribution of the observed random variable conditioned on the selection will not be what it was without the selection (plus the addition of the inverse Mills ratio, λ_i to the index function). Thus, one cannot just add λ_i to the same likelihood function. (In fact, this can be seen even for the linear case. The least squares estimator (with λ_i) is not the MLE; it is merely a feasible consistent estimator. The appropriate log likelihood, which corresponds to a skewed distribution, appears above. (One might then ask, since OLS is consistent in the linear case, would "conventional MLE" be consistent in the nonlinear case? This remains an area for research, but it seems unlikely. Terza (1995, 1998) and Greene (1997) have examined this in detail for the Poisson regression model, where it appears not to be the case. One well worked out special case does appear in the literature already. Maddala (1983) and Boyes, Hoffman and Lowe (1989) obtained the appropriate closed form log likelihood for a bivariate probit model subject to sample selection. The other well known example is the open form result for the Poisson regression model obtained by Terza (1995, 1998) and Greene (1995, 1997).

A generic log likelihood for nonlinear models with sample selection is developed in Terza (1998) and Greene (2006). The model will take the form

$$
\begin{aligned}
d_i &= 1(\boldsymbol{\alpha}'\mathbf{z}_i + w_i > 0) \quad w_i \sim N[0,1], \\
g_i|\varepsilon_i &= g(\boldsymbol{\beta}'\mathbf{x}_i, \sigma_\varepsilon \varepsilon_i) \quad \varepsilon_i \sim N[0,1] \\
y_i\,|\mathbf{x}_i, \varepsilon_i &\sim f[y_i \mid g(\boldsymbol{\beta}'\mathbf{x}_i, \sigma_\varepsilon \varepsilon_i)] \\
[w_i, \varepsilon_i] &\sim N[(0,1),(1,\rho,1)] \\
y_i, \mathbf{x}_i &\text{ are observed only when } z_i = 1.
\end{aligned}
$$

Note that the model is assumed to involve an index function, $\boldsymbol{\beta}'\mathbf{x}_i$ and the normally distributed heterogeneity, ε_i, not necessarily, albeit usually, combined in a term $\boldsymbol{\beta}'\mathbf{x}_i+\varepsilon_i$. The density that enters

the log likelihood will then be.

$$f(y_i,d_i|\mathbf{x}_i,\mathbf{z}_i) = \int_{-\infty}^{\infty} \{(1 - d_i) + d_i f[y_i|\, g(\boldsymbol{\beta}'\mathbf{x}_i,\sigma_\varepsilon \varepsilon_i)]\}\, \Phi\left((2d_i - 1)[\boldsymbol{\alpha}'\mathbf{z}_i + \rho\varepsilon_i]/\sqrt{1-\rho^2}\right)\, \phi(\varepsilon_i)d\varepsilon_i ,$$

Since the integrals do not exist in closed form, they are approximated either by quadrature [Terza(1998)] or by simulation [Greene (2006)].

2.5 Estimating a Stochastic Frontier Model with Sample Selection.

The combination of efficiency estimation and sample selection appears in several studies. Bradford, et. al (2000) studied patient specific costs for cardiac revascularization in a large hospital. They state "... the patients in this sample were not randomly assigned to each treatment group. Statistically, this implies that the data are subject to sample selection bias. Therefore, we utilize a standard Heckman two-stage sample-selection process, creating an inverse Mill's ratio from a first-stage probit estimator of the likelihood of CABG or PTCA. This correction variable is included in the frontier estimate...." (page 306). (The authors opt for the GMM estimator based on Kopp and Mullahy's (1990) (KM) relaxation of the distributional assumptions in the standard frontier model and, it is suggested, that KM "find that the traditional maximum likelihood estimators tend to overestimate the average inefficiency." (Page 304) KM did not, in fact, make the latter argument, and we can find no evidence to support it in the since received literature. KM's support for the GMM estimator is based on its more general, distribution free specification. We do note, Newhouse (1994), whom Bradford et. al cite, has stridently argued against the stochastic frontier model, but not based on the properties of the MLE.)

Sipiläinen and Oude Lansink (2005) have utilized a stochastic frontier, translog model to analyze technical efficiency for organic and conventional farms. They state "Possible selection bias between organic and conventional production can be taken into account [by] applying Heckman's (1979) two step procedure." (Page 169.) In this case, the inefficiency component in the stochastic frontier translog distance function is distributed as the truncation at zero of a U_i with a heterogeneous mean. [See Battese and Coelli (1995).] The IMR is added to the deterministic (production function) part of the frontier function.

Other authors have acknowledged the sample selection issue in stochastic frontier studies. Kaparakis, Miller and Noulas (1994) in an analysis of commercial banks and Collins and Harris (2005) in their study of UK chemical plants both suggested that "sample selection" was a potential issue in their analysis. Neither of these formally modified their stochastic frontier models to accommodate the result, however.

If we specify that the unobservables in the selection model are correlated with the "noise" in the stochastic frontier model, then the combination of the two models is

$d_i = 1[\boldsymbol{\alpha}'\mathbf{z}_i + w_i > 0],\ w_i \sim N[0,1^2]$
$y_i = \boldsymbol{\beta}'\mathbf{x}_i + \varepsilon_i,\ \varepsilon_i \sim N[0,\sigma_\varepsilon^2]$
$(y_i,\mathbf{x}_i)$ observed only when $d_i = 1$.
$\varepsilon_i = v_i - u_i$
$u_i = |\sigma_u U_i| = \sigma_u |U_i|$ where $U_i \sim N[0,1^2]$
$v_i = \sigma_v V_i$ where $V_i \sim N[0,1^2]$.
$(w_i,v_i) \sim N_2[(0,1), (1, \rho\sigma_v, \sigma_v^2)]$

("Sample selection bias" arises as a consequence of the correlation of the unobservables in the main equation with those in the sample selection equation. Thus, the ambiguity in adding an IMR to a model that contains no such unobservables, such as the probit model [Wynand and van Praag

(1981) or the base case Poisson model [Greene (1994).]) The conditional density for an observation in this specification is

$$f\,[y_i|x_i,|U_i|,\mathbf{z}_i,d_i,]= \left[d_i \frac{\exp\left(-\frac{1}{2}(y_i-\beta' x_i+\sigma_u\,|U_i\,|)^2/\sigma_v^2\right)}{\sigma_v\sqrt{2\pi}}\Phi\left(\frac{\rho(y_i-\beta' x_i+\sigma_u\,|U_i\,|)/\sigma_\varepsilon+\boldsymbol{\alpha}'\mathbf{z}_i}{\sqrt{1-\rho^2}}\right) + (1-d_i)\Phi(-\boldsymbol{\alpha}'\mathbf{z}_i)\right]$$

The log likelihood is formed by integrating out the unobserved $|U_i|$ then maximizing with respect to the unknown parameters. Thus,

$$\log L(\boldsymbol{\beta},\sigma_u,\sigma_v,\boldsymbol{\alpha},\rho) = \sum_{i=1}^{N}\log) p(|U_i\,|)d\,|U_i\,|\,.$$

In this instance, the integral is not known; it must be approximated, once again either by quadrature or by simulation. Using the latter, we obtain the simulated log likelihood,

$$\log L_S(\boldsymbol{\beta},\sigma_u,\sigma_v,\boldsymbol{\alpha},\rho)= \sum_{i=1}^{N}\log\frac{1}{R}\sum_{r=1}^{R}\left[d_i \frac{\exp\left(-\frac{1}{2}(y_i-\boldsymbol{\beta}'\mathbf{x}_i+\sigma_u\,|U_{ir}\,|)^2/\sigma_v^2\right)}{\sigma_v\sqrt{2\pi}}\times \Phi\left(\frac{\rho(y_i-\boldsymbol{\beta}'\mathbf{x}_i+\sigma_u\,|U_{ir}\,|)/\sigma_\varepsilon+\boldsymbol{\alpha}'\mathbf{z}_i}{\sqrt{1-\rho^2}}\right) +(1-d_i)\Phi(-\boldsymbol{\alpha}'\mathbf{z}_i)\right].$$

To simplify the estimation, we have used a two step approach. The single equation MLE of $\boldsymbol{\alpha}$ in the probit equation is consistent, albeit inefficient. For purposes of estimation of the parameters of the stochastic frontier model, however, $\boldsymbol{\alpha}$ need not be reestimated. We take the estimates of $\boldsymbol{\alpha}$ as given in the simulated log likelihood at the second step, then use the Murphy and Topel (2002) correction to adjust the standard errors (in essentially the same fashion as Heckman's correction of the canonical selection model). Thus, our conditional simulated log likelihood function is

$$\log L_{S,C}(\boldsymbol{\beta},\sigma_u,\sigma_v,\rho)= \sum_{i=1}^{N}\log\frac{1}{R}\sum_{r=1}^{R}\left[d_i \frac{\exp\left(-\frac{1}{2}(y_i-\boldsymbol{\beta}'\mathbf{x}_i+\sigma_u\,|U_{ir}\,|)^2/\sigma_v^2\right)}{\sigma_v\sqrt{2\pi}}\times \Phi\left(\frac{\rho(y_i-\boldsymbol{\beta}'\mathbf{x}_i+\sigma_u\,|U_{ir}\,|)/\sigma_\varepsilon+a_i}{\sqrt{1-\rho^2}}\right) +(1-d_i)\Phi(-a_i)\right].$$

where $a_i = \hat{\boldsymbol{\alpha}}'\mathbf{z}_i$. With this simplification, the nonselected observations (those with $d_i = 0$) do not contribute information about the parameters to the simulated log likelihood. Thus, the function we maximize becomes

$$\log L_{S,C}(\boldsymbol{\beta},\sigma_u,\sigma_v,\rho) = \sum_{d_i=1} \log \frac{1}{R} \sum_{r=1}^{R} \left[\frac{\exp\left(-\frac{1}{2}(y_i - \beta' x_i + \sigma_u |U_{ir}|)^2 / \sigma_v^2\right)}{\sigma_v \sqrt{2\pi}} \times \Phi\left(\frac{\rho(y_i - \beta' x_i + \sigma_u |U_{ir}|)/\sigma_\varepsilon + a_i}{\sqrt{1-\rho^2}} \right) \right].$$

The parameters of the model were estimated using a conventional gradient based approach, the BFGS method. The derivatives of the function must be simulated as well. The BHHH estimator is used to estimate the standard errors for the parameter estimators. Note that when $\rho = 0$, the maximand reduces to that of the maximum simulated likelihood estimator of the basic frontier model shown earlier. This provides us with a method of testing the specification of the selectivity model against the simpler model using a (simulated) likelihood ratio test.

2.6 Estimating Observation Specific Inefficiency

The end objective of the estimation process is to characterize the *inefficiency* in the sample, u_i or the *efficiency*, $\exp(-u_i)$. Aggregate summary measures, such as the sample mean and variance are often provided (e.g., Bradford, et. al (2000) for hospital costs). Researchers also compute individual specific estimates of the conditional means based on the Jondrow et al. (1982) (JLMS) result

$$E[u_i | \varepsilon_i] = \frac{\sigma\lambda}{1+\lambda^2}\left[\mu_i + \frac{\phi(\mu_i)}{\Phi(\mu_i)} \right], \ \mu_i = \frac{-\lambda\varepsilon_i}{\sigma}, \ \varepsilon_i = y_i - \boldsymbol{\beta}'\mathbf{x}_i.$$

The standard approach computes this function after estimation based on the maximum likelihood estimates. In principle, we could repeat this computation with the maximum simulated likelihood estimates. An alternative approach takes advantage of the simulation of the values of u_i during estimation. Using Bayes theorem, we can write

$$p(u_i | \varepsilon_i) = \frac{p(u_i, \varepsilon_i)}{p(\varepsilon_i)} = \frac{p(\varepsilon_i | u_i) p(u_i)}{\int_{u_i} p(\varepsilon_i | u_i) p(u_i) du_i}.$$

Recall $u_i = \sigma_u |U_i|$. Thus, equivalently,

$$p[(\sigma_u |U_i|) \ | \varepsilon_i] = \frac{p[(\sigma_u |U_i|), \varepsilon_i]}{p(\varepsilon_i)} = \frac{p[\varepsilon_i | (\sigma_u |U_i|)] p(\sigma_u |U_i|)}{\int_{u_i} p[\varepsilon_i | (\sigma_u |U_i|)] p(\sigma_u |U_i|) d(\sigma_u |U_i|)}.$$

The desired expectation is, then

$$E[(\sigma_u |U_i|) | \varepsilon_i] = \frac{\int_{\sigma_u |U_i|} (\sigma_u |U_i|) p[\varepsilon_i | (\sigma_u |U_i|)] p(\sigma_u |U_i|) d(\sigma_u |U_i|)}{\int_{\sigma_u |U_i|} p([\varepsilon_i | (\sigma_u |U_i|)] p(\sigma_u |U_i|) d(\sigma_u |U_i|)}.$$

These are the terms that enter the simulated log likelihood for each observation. The simulated denominator would be

$$\hat{B}_i = \frac{1}{R}\sum_{r=1}^{R}\left[\begin{array}{l}\dfrac{\exp\left(-\frac{1}{2}(y_i - \hat{\boldsymbol{\beta}}'\mathbf{x}_i + \hat{\sigma}_u |U_{ir}|)^2 / \hat{\sigma}_v^2)\right)}{\hat{\sigma}_v\sqrt{2\pi}}\times \\ \Phi\left(\dfrac{\hat{\rho}(y_i - \hat{\boldsymbol{\beta}}'\mathbf{x}_i + \hat{\sigma}_u |U_{ir}|)/\hat{\sigma}_\varepsilon + a_i}{\sqrt{1-\hat{\rho}^2}}\right)\end{array}\right] = \frac{1}{R}\sum_{r=1}^{R}\hat{f}_{ir}$$

while the numerator is simulated with $\hat{A}_i = \frac{1}{R}\sum_{r=1}^{R}(\hat{\sigma}_u |U_{ir}|)\hat{f}_{ir}$. The estimate of E[$u_i$|$\varepsilon_i$] is then $\hat{A}_i / \hat{B}_i$. These are computed for each observation using the estimated parameters, the raw data and the same pool of random draws as were used to do the estimation. As shown below, this gives a strikingly similar answer to the JLMS plug in result suggested at the outset.

3. Application

In 2000, the World Health Organization published its millennium edition of the *World Health Report* (WHR) [WHO (2000).] The report contained Tandon at al's (2000) (TMLE) frontier analysis of the efficiency of health care delivery for 191 countries. The frontier analysis attracted a surprising amount of attention in the popular press (given its small page length, minor role in the report and highly technical nature), notably for its assignment of a rank of 37 to the United States's health care system. [Seven years after the report, it still commanded attention, e.g., *New York Times* (2007).] The authors provided their data and methodology to numerous researchers who have subsequently analyzed, criticized, and extended the WHO study. [E.g., Gravelle et al. (2002a,b), Hollingsworth and Wildman (2002) and Greene (2004).]

TMLE based their analysis on COMP, a new measure of health care attainment that they created: (The standard measure at the time was DALE, disability adjusted life expectancy.) "In order to assess overall efficiency, the first step was to combine the individual attainments on all five goals of the health system into a single number, which we call the composite index. The composite index is a weighted average of the five component goals specified above. First, country attainment on all five indicators (i.e., health, health inequality, responsiveness-level, responsiveness-distribution, and fair-financing) were rescaled restricting them to the [0,1] interval. Then the following weights were used to construct the overall composite measure: 25% for health (DALE), 25% for health inequality, 12.5% for the level of responsiveness, 12.5% for the distribution of responsiveness, and 25% for fairness in financing. These weights are based on a survey carried out by WHO to elicit stated preferences of individuals in their relative valuations of the goals of the health system." (TMLE, page 4.) (It is intriguing that in the public outcry over the results, it was never reported that the WHO study did not, in fact, rank countries by health care attainment, COMP, but rather by the efficiency with which countries attained their COMP. (That is, countries were ranked by the difference between their COMP and a constructed country specific optimal COMP*.) In terms of COMP, itself, the U.S. ranked 15th in the study, not 37th, and France did not rank first as widely reported, Japan did. The full set of results needed to reach these conclusions are contained in TMLE.)

The data set used by TMLE contained five years (1993-1997) of observations on the time varying variables COMP, per capita health care expenditure and average educational attainment, and time invariant, 1997 observations on the set of variables listed in Table 1. TMLE used a linear fixed effects translog production model,

$$\log COMP_{it} = \beta_1 + \beta_2 \log HExp_{it} + \beta_3 \log Educ_{it} + \beta_4 \log^2 Educ_{it} + \beta_5 \log^2 HExp_{it} + \beta_6 \log HExp_{it} \log \mathrm{Educ}_{it} - u_i + v_{it}.$$

in which health expenditure and education enter loglinearly and quadratically. (They ultimately dropped the last two terms in their specification.) Their estimates of u_i were computed from the estimated constant terms in the linear fixed effects regression. Since their analysis was based on the fixed effects regression, they did not use the time invariant variables in their regressions or subsequent analysis. [See Greene (2004) for discussion.] Their overall efficiency indexes for the 191 WHO member countries are published in the report (Table 1, pages 18-21) and used in the analysis below.

Table 1 lists descriptive statistics for the TMLE efficiencies and for the variables present in the WHO data base. The COMP, education and health expenditure are described for the 1997 observation. Although these variables are time varying, the amount of within group variation ranges from very small to trivial. [See Gravelle et. al (2002a) for discussion.] The time invariant variables were not used in their analysis. The data in Table 1 are segmented by OECD membership. The OECD members are primarily 30 of the wealthiest countries (thought not specifically *the* 30 wealthiest countries). The difference between OECD countries and the rest of the world is evident. Figure 1 below plots the ETML efficiency estimates versus per capita GDP, segmented by OECD membership. The figure is consistent with the values in Table 1. This suggests (but, of course, does not establish) that OECD membership may be a substantive selection mechanism. OECD membership is based on more than simply per capita GDP. The selectivity issue is whether other factors related to OECD membership are correlated with the stochastic element in the production function.

Figure 1 plots TMLE's estimated efficiency scores against per capita GDP for the 191 countries stratified by OECD membership. The difference is stark. The layer of points at the top of the figure for the OECD countries suggests that, as might be expected, wealth produces efficiency in the outcome. The question for present purposes is whether the selection based on the observed GDP value is a complete explanation of the difference, or whether there are latent factors related to OECD membership that also impact the placement of the frontier function. We will use the sample selection model developed earlier to examine the issue. We note, it is not our intent here to replace the results of the WHO study. Rather, this provides a setting for demonstrating the selection model. Since we will be using a stochastic frontier model while they used a fixed effects linear regression, there is no reason to expect the resulting efficiency scores to be similar or even comparable. It is interesting to compare the rankings produced by the two methodologies, though we will do so without naming names.

We have estimated the stochastic frontier models for the logCOMP measure using TMLE's truncated specification of the translog model. Since the time invariant data are only observed for 1997, we have used the country means of the logs of the variables COMP, HExp and Educ in our estimation. Table 2 presents the maximum likelihood and maximum simulated likelihood estimates of the parameters of the frontier models. The MSL estimates are computed using 200 Halton draws for each observation for the simulation. [See Greene (2008) or Train (2003) for discussion of Halton sequences.] By using Halton draws rather than pseudorandom numbers, we can achieve replicability of the estimates. To test the specification of the selection model, we have fit the sample selection model while constraining ρ to equal zero. The log likelihood functions can then be compared using the usual chi squared statistic. The results provide two statistics for the test, then, the Wald statistic (t ratio) associated with the estimate of

ρ and the likelihood ratio statistic. Both Wald statistics fail to reject the null hypothesis of no selection. For the LR statistics (with one degree of freedom) we do not reject the base model for the non-OECD countries, but we do for the OECD countries, in conflict with the t test. Since the sample is only 30 observations, the statistic may be suspect. We would conclude that the evidence does not strongly support the selection model. It would seem that the selection is dominated by the observables, presumably primarily by per capita income. Figure 2 plots the estimated efficiency scores from the stochastic frontier model versus those in the WHO report. As anticipated in Greene (2004), the impact of the fixed effects regression is too attribute to inefficiency effects that might be better explained by cross country heterogeneity. These effects would be picked up by the noise term in the frontier model. Figure 3 shows a plot of the two estimators of the inefficiency scores in the selectivity corrected frontier model, the JLMS estimator and the simulated values of $E[u|\varepsilon]$ computed during the estimation. As noted earlier, they are strikingly similar. Finally, figure 4 shows a plot of the country ranks based on the stochastic frontier model versus the country ranks implicit in the WHO estimates for the non-OECD countries. The essential lack of correlation in the two sets of results should cast at least some suspicion on the original study; the results depend crucially on the specification.

4. Conclusions

We have developed a maximum simulated likelihood estimator ALS's the normal – half normal stochastic frontier model. The normal exponential model, a normal –t model, or normal anything else model would be trivial modifications. The manner in which the values of u_i are simulated is all the changes. The identical simulation based estimator of the inefficiencies is used as well. We note that in a few other cases, such as the t distribution, simulation (or MCMC) is the only feasible method of proceeding. [See Tsionas, Kumbhakar and Greene (2008).]

Replication of the Pitt and Lee (1981) random effects form of the model, again with any distribution from which draws can be simulated, is simple. The term B_i defined earlier that enters the log likelihood becomes

$$B_i = \frac{1}{R}\sum_{r=1}^{R}\prod_{t=1}^{T_i}\left[\begin{array}{l}\dfrac{\exp\left(-\frac{1}{2}(y_{it}-\boldsymbol{\beta}'\mathbf{x}_{it}+\sigma_u|U_{ir}|)^2/\sigma_v^2)\right)}{\sigma_v\sqrt{2\pi}}\times \\ \Phi\left(\dfrac{\rho(y_{it}-\boldsymbol{\beta}'\mathbf{x}_{it}+\sigma_u|U_{ir}|)/\sigma_\varepsilon+a_i}{\sqrt{1-\rho^2}}\right)\end{array}\right] = \frac{1}{R}\sum_{r=1}^{R}\prod_{t=1}^{T_i}\hat{f}_{irt}$$

Further refinements, such as a counterpart to Battese and Coelli (1992, 1995) and the Stevenson's (1980) truncation model may be possible as well. This remains to be investigated.

The assumption that the unobservables in the selection equation are correlated with the heterogeneity in the production function but uncorrelated with the inefficiency is an important feature of the model. It seems natural and appropriate in this setting – one might expect that observations are not selected into the sample based on their being inefficient to begin with. Nonetheless, that, as well, is an issue that might be further considered. A related question is whether it is reasonable to assume that the heterogeneity and the inefficiency in the production model should be assumed to be uncorrelated. Some progress has been made in this regard, e.g., in Smith (2003), but the analysis is tangential to the model considered here.

We have re(-re)visited the WHO (2000) study, and found, once again, that the results vary greatly depending on the specification. It does appear that our expectation that selection on OECD membership was not supported statistically, however.

Table 1 Descriptive Statistics for WHO Variables, 1997 Observations*

	Non-OECD		OECD		All	
	Mean	**Std. Dev.**	**Mean**	**Std. Dev.**	**Mean**	**Std. Dev**
COMP	70.30	10.96	89.42	3.97	73.30	12.34
HEXP	249.17	315.11	1498.27	762.01	445.37	616.36
EDUC	5.44	2.38	9.04	1.53	6.00	2.62
GINI	0.399	0.0777	0.299	0.0636	0.383	0.0836
VOICE	-0.195	0.794	1.259	0.534	0.0331	0.926
GEFF	-0.312	0.643	1.166	0.625	-0.0799	0.835
TROPICS	0.596	0.492	0.0333	0.183	0.508	0.501
POPDEN	757.9	2816.3	454.56	1006.7	710.2	2616.5
PUBFIN	56.89	21.14	72.89	14.10	59.40	20.99
GDPC	4449.8	4717.7	18199.07	6978.0	6609.4	7614.8
Efficiency	0.5904	0.2012	0.8831	0.0783	0.6364	0.2155
Sample	161		30		191	

* Variables in the data set are as follows:

COMP = WHO health care attainment measure.
HEXP = Per capita health expenditure in PPP units.
EDUC = Average years of formal education.
GINI = World bank measure of income inequality.
VOICE = World bank measure of democratization.
GEFF = World bank measure of government effectiveness.
TROPICS = Dummy variable for tropical location.
POPDEN = Population density in persons per square kilometer.
PUBFIN = Proportion of health expenditure paid by government.
GDPC = Per capita GDP in PPP units.
Efficiency = TMLE estimated efficiency from fixed effects model.

Table 2 Estimated Stochastic Frontier Models[a] (Estimated standard errors in parentheses)

	Non-OECD Countries		OECD Countries	
	Stochastic Frontier	**Sample Selection**	**Stochastic Frontier**	**Sample Selection**
Constant	3.76162 (0.05429)	3.74915 (0.05213)	3.10994 (1.15519)	3.38244 (1.42161)
LogHexp	0.08388 (0.01023)	0.08842 (0.010228)	0.04765 (0.006426)	0.04340 (0.008805)
LogEduc	0.09096 (0.075150)	0.09053 (0.073367)	1.00667 (1.06222)	0.77422 (1.2535)
Log^2Educ	0.00649 (0.02834)	0.00564 (0.02776)	-0.23710 (0.24441)	-0.18202 (0.28421)
σ_u	0.12300	0.12859	0.02649	0.01509
σ_v	0.05075	0.04735	0.00547	0.01354
λ	2.42388	2.71549	4.84042	1.11413
σ	0.13306	0.13703	0.02705	0.02027
ρ	0.0000	0.63967 (1.4626)	0.0000	-0.73001 (0.56945)
logL	160.2753	161.0141	62.96128	65.44358
LR test	1.4776		4.9646	
N	161		30	

[a]The estimated probit model for OECD membership (with estimated standard errors in parentheses) is
OECD = -9.2404 (3.369) + 0.7388 (0.3820) + 0.6098 (0.4388) + 0.7291 (0.3171)

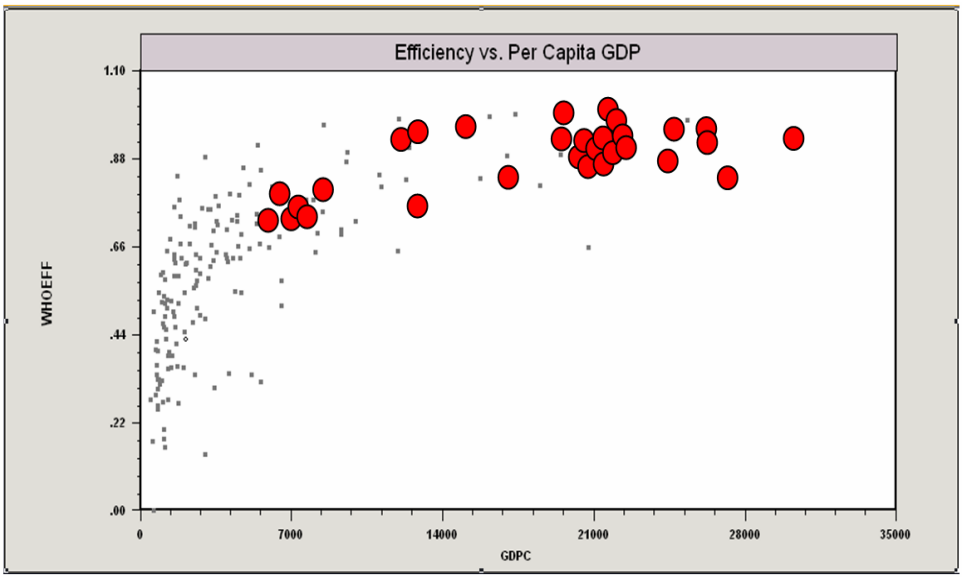

Figure 1 Efficiency Scores Related to Per Capita GDP.

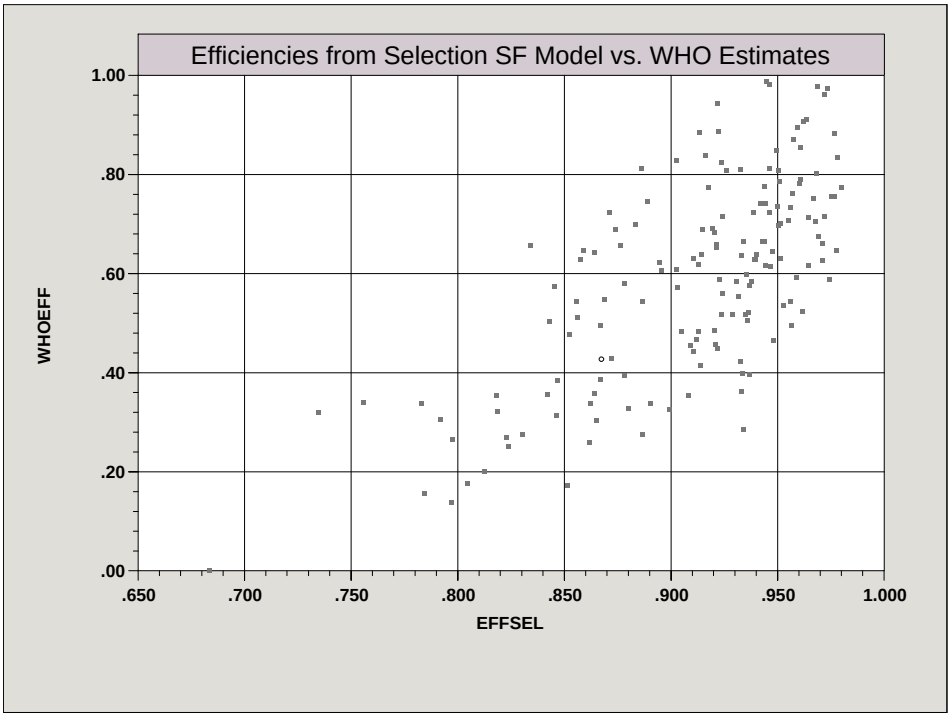

Figure 2 Estimated Efficiency Scores

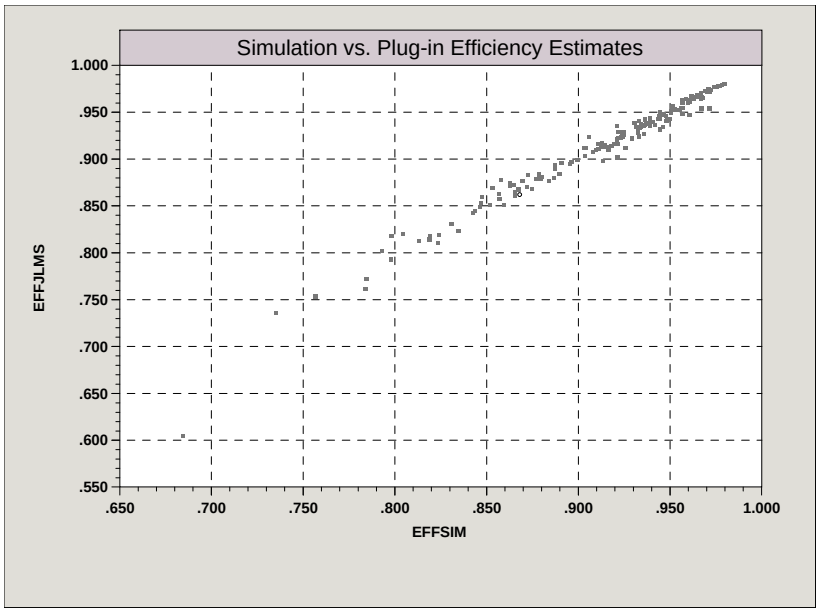

Figure 3 Alternative Estimators of Efficiency Scores

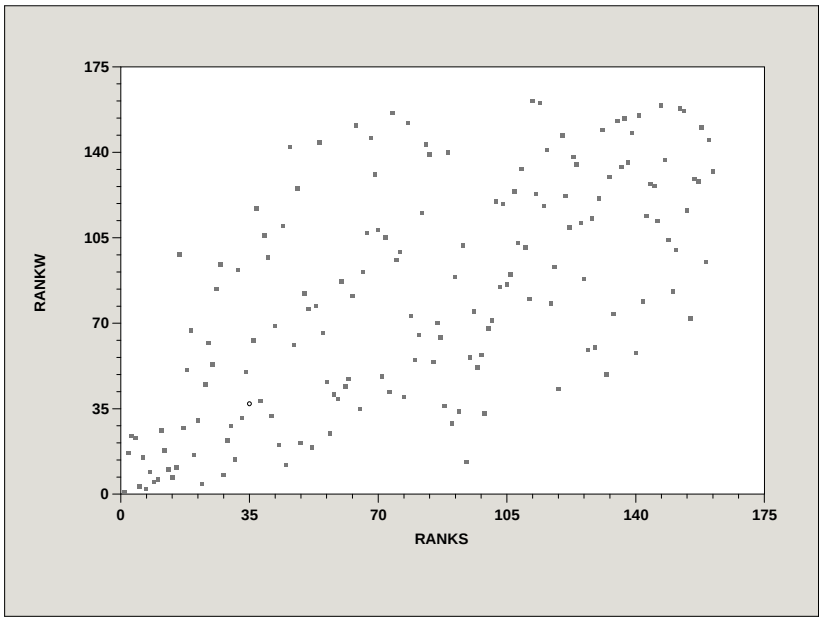

Figure 4 Ranks of Countries Based on WHO and Simulation Efficiency Estimates

References

Aigner, D., K. Lovell, and P. Schmidt, 1977, "Formulation and Estimation of Stochastic Frontier Production Function Models," *Journal of Econometrics*, 6, pp. 21-37.

Battese, G. and T. Coelli, 1995, "A Model for Technical Inefficiency Effects in a Stochastic Frontier Production for Panel Data," *Empirical Economics*, 20, pp. 325-332.

Bradford, D., Kleit, A., Krousel-Wood, M. and Re, R., "Stochastic Frontier Estimation of Cost Models within the Hospital," *Review of Economics and Statistics*, 83, 2, 2000, pp. 302-309.

Econometric Software, Inc., *LIMDEP Version 9.0*, Plainview, New York, 2007.

Collins, A. and R. Harris, "The Impact of Foreign Ownership and Efficiency on Pollution Abatement Expenditures by Chemical Plants: Some UK Evidence," *Scottish Journal of Political Economy*, 52, 5, 2005, pp. 757-768.

Gourieroux, C. and A. Monfort, *Simulation Based Econometric Methods*, Oxford: Oxford University Press, 1996.

Gravelle H, Jacobs R, Jones A, Street, "Comparing the Efficiency of National Health Systems: Econometric Analysis Should Be Handled with Care," University of York, Health Economics, UK. Manuscript , 2002a.

Gravelle H, Jacobs R, Jones A, Street, "Comparing the Efficiency of National Health Systems: A Sensitivity Approach," University of York, Health Economics, Manuscript, UK, 2002b.

Greene, W., "Accounting for Excess Zeros and Sample Selection in Poisson and Negative Binomial Regression Models," Stern School of Business, NYU, Working Paper EC-94-10, 1994.

Greene, W., "Sample Selection in the Poisson Regression Model," Stern School of Business, Department of Economics, Working paper #95-06, 1995.

Greene, W., FIML Estimation of Sample Selection Models for Count Data, Stern School, Department of Economics, Working Paper 97-02, 1997

Greene, W., 2004, "Distinguishing Between Heterogeneity and Inefficiency: Stochastic Frontier Analysis of the World Health Organization's Panel Data on National Health Care Systems," *Health Economics*, 13, pp. 959-980.

Greene, W., "A General Approach to Incorporating 'Selectivity' in a Model," Working Paper EC-06-10, Stern School of Business, New York University, 2006.

Greene, W., "The Econometric Approach to Efficiency Analysis," in K Lovell and S. Schmidt, eds. *The Measurement of Efficiency*, H Fried, , Oxford University Press, 2008a.

Greene, W., *Econometric Analysis, 6th ed*., Prentice Hall, Englewood Cliffs, 2008b.

Heckman, J. "Sample Selection Bias as a Specification Error." *Econometrica,* 47, 1979, pp. 153–161.

Hollingsworth J, Wildman B., 2002, The Efficiency of Health Production: Re-estimating the WHO Panel Data Using Parametric and Nonparametric Approaches to Provide Additional Information. *Health Economics*, 11, pp. 1-11.

Jondrow, J., K. Lovell, I. Materov, and P. Schmidt, 1982, "On the Estimation of Technical Inefficiency in the Stochastic Frontier Production Function Model," *Journal of Econometrics*, 19, pp. 233-238.

Kaparakis, E., S. Miller and A. Noulas, "Short Run Cost Inefficiency of Commercial Banks: A Flexible Stochastic Frontier Approach," Journal of Money, Credit and Banking, 26, 1994, pp. 21-28.

Kopp, R. and J. Mullahy, "Moment-based Estimation and Testing of Stochastic Frontier Models," *Journal of Econometrics*, 46, 1/2, 1990, pp. 165-184.

Koop, G. and M. Steel, 2001, "Bayesian Analysis of Stochastic Frontier Models," in B. Baltagi, ed., *Companion to Theoretical Econometrics*, Blackwell Publishers, Oxford.

Maddala, G., *Limited Dependent and Qualitative Variables in Econometrics*, Cambridge: Cambridge University Press, 1983.

New York Times, Editorial: "World's Best Medical Care?" August 12, 2007.

Newhouse, J., "Frontier Estimation: How Useful a Tool for Health Economics?" Journal of Health Economics, 13, 1994, pp. 317-322.

Pitt, M., and L. Lee, 1981, "The Measurement and Sources of Technical Inefficiency in the Indonesian Weaving Industry," *Journal of Development Economics*, 9, pp. 43-64.

Sipiläinen, T. and A. Oude Lansink, "Learning in Switching to Organic Farming," Nordic Association of Agricultural Scientists, NJF Report Volume 1, Number 1, 2005. http://orgprints.org/5767/01/N369.pdf

Smith, M., "Modelling Sample Selection Using Archimedean Copulas," *Econometrics Journal*, 6, 2003, pp. 99-123.

Stevenson, R., 1980, "Likelihood Functions for Generalized Stochastic Frontier Estimation," *Journal of Econometrics*, 13, pp. 58-66.

Tandon, A., C. Murray, J. Lauer and D. Evans, "Measuring the Overall Health System Performance for 191 Countries," World Health Organization, GPE Discussion Paper, EIP/GPE/EQC Number 30, 2000. http://www.who.int/entity/healthinfo/paper30.pdf

Terza, J., "Estimating Count Data Models with Endogenous Switching and Sample Selection," Department of Economics, Penn State University, Working Paper IPRE-95-14, 1995.

Terza, J. "Estimating Count Data Models with Endogenous Switching: Sample Selection and Endogenous Treatment Effects." *Journal of Econometrics*, 84, 1, 1998, pp. 129–154.

Train, K., *Discrete Choice Methods with Simulation*, Cambridge: Cambridge University Press, 2003.

Tsionas, E., S. Kumbhakar and W. Greene, "Non-Gaussian Stochastic Frontier Models," Manuscript, Department of Economics, University of Binghamton, 2008.

World Health Organization, *The World Health Report*, WHO, Geneva, 2000

Wynand, P., and B. van Praag. "The Demand for Deductibles in Private Health Insurance: A Probit Model with Sample Selection." *Journal of Econometrics*, 17, 1981, pp. 229–252.