Editorial Manager(tm) for Journal of Productivity Analysis
Manuscript Draft

Manuscript Number:

Title: A Stochastic Frontier Model with Correction for Sample Selection

Article Type: Original Research

Keywords: Stochastic Frontier; Sample Selection; Simulation;
Efficiency

Corresponding Author: Dr William Greene, PhD

Corresponding Author's Institution: New York University, Stern School of Business

First Author: William Greene, PhD

Order of Authors: William Greene, PhD

A Stochastic Frontier Model with Correction for Sample Selection

William Greene*
Department of Economics, Stern School of Business,
New York University,
March, 2008
Revised April, 2009

Abstract

Heckman's (1976, 1979) sample selection model has been employed in three decades of applications of linear regression studies. 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 using maximum simulated likelihood. We then extend the technique to a stochastic frontier model with sample selection. 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: C13; C15; C21

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 pages.stern.nyu.edu/~wgreene.

1 Introduction

Heckman's (1976, 1979) sample selection model has been employed in three decades of applications of linear regression studies. Numerous applications have extended Heckman's approach to nonlinear settings such as the binary probit and Poisson regression models. The first is Wynand and van Praag's (1981) development of a probit model for insurance purchase. Among a number of other recent applications, Bradford et al. (2001) extended Heckman's method to a stochastic frontier model for hospital costs. The familiar approach in which a sample selection correction term is simply added to the model of interest (see (7) and (8)) is not appropriate for nonlinear models such as the stochastic frontier. In this study, we build on the maximum likelihood estimator of Heckman's sample selection corrected linear model and the extension to nonlinear models by Terza (1996, 2009) to obtain a counterpart 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 next step is to extend the technique to a stochastic frontier model in the presence of sample selection.

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 stochastic frontier model of Aigner, Lovell and Schmidt (1977) (ALS) is specified with

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

A vast literature has explored variations in the specification to accommodate, e.g., heteroscedasticity, panel data formulations, etc.[1] It will suffice for present purposes to

[1] See Greene (2008a) for further development of the model and a survey of extensions and applications.

work with the simplest form. Extensions will be considered later. The model can be estimated by modifications of ordinary least squares [e.g., Greene (2008a)], the generalized method of moments [Kopp and Mullahy (1990)] or, as is conventional in the recent 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, a fourth estimator, maximum simulated likelihood (MSL). The simulation based estimator merely replicates the conventional estimator for the base case, in which the closed form is already available. The log likelihood function for the sample selection model does not exist in closed form, so some approximation method, such as MSL is necessary.

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

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

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

and $\Phi(.)$ denotes the standard normal cdf. The density satisfies the standard regularity conditions, and maximum likelihood estimation of the model is a conventional problem handled with familiar methods. Estimation is straightforward and has been installed in the menu of supported techniques in a variety of programs including *LIMDEP*, *Stata* and *TSP*.[2]

Conditioned on u_i, the central equation of the model in (2.1) would be a classical linear regression model with normally distributed disturbances. Thus,

$$f(y_i|\mathbf{x}_i,|U_i|) = \frac{\exp[-\tfrac{1}{2}(y_i - \boldsymbol{\beta}'\mathbf{x}_i + \sigma_u|U_i|)^2/\sigma_v^2]}{\sigma_v\sqrt{2\pi}}. \qquad (3)$$

[2] Details on maximum likelihood estimation of the model can be found in ALS and elsewhere, e.g., Greene (2008b, Ch. 16).

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

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

$$\text{where } p(|U_i|) = \frac{\phi(|U_i|)}{\Phi(0)} = \exp[-\tfrac{1}{2}|U_i|^2]\sqrt{\frac{2}{\pi}}, |U_i| \geq 0, \quad (4)$$

$$\text{then } \log L(\boldsymbol{\beta},\sigma_u,\sigma_v) = \sum_{i=1}^{N} \log f(y_i|\mathbf{x}_i),$$

where ϕ is the standard normal density and Φ is the standard normal cdf . The closed form of the integral appears in (2).[3] Consider using simulation to approximate the integrals;

$$f(y_i|\mathbf{x}_i) \approx \frac{1}{R}\sum_{r=1}^{R} \frac{\exp[-\frac{1}{2}(y_i - \boldsymbol{\beta}'\mathbf{x}_i + \sigma_u |U_{ir}|)^2/\sigma_v^2]}{\sigma_v\sqrt{2\pi}}, \quad (5)$$

where U_{ir} is R random draws from the standard normal population. (There is no closed form for the extension of the model that appears below.) The simulated log likelihood is

$$\log L_S(\boldsymbol{\beta},\sigma_u,\sigma_v) = \sum_{i=1}^{N} \log\left\{\frac{1}{R}\sum_{r=1}^{R} \frac{\exp[-\frac{1}{2}(y_i - \boldsymbol{\beta}'\mathbf{x}_i + \sigma_u |U_{ir}|)^2/\sigma_v^2]}{\sigma_v\sqrt{2\pi}}\right\}. \quad (6)$$

The maximum simulated likelihood estimators of the model parameters are obtained by maximizing this function with respect to the unknown parameters.[4]

2.2 Sample Selection in the Linear Model

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

$$\begin{aligned} &d_i = 1[\boldsymbol{\alpha}'\mathbf{z}_i + w_i > 0], \ w_i \sim \mathrm{N}[0,1] \\ &y_i = \boldsymbol{\beta}'\mathbf{x}_i + \varepsilon_i, \ \varepsilon_i \sim \mathrm{N}[0,\sigma_\varepsilon^2] \\ &(w_i,\varepsilon_i) \sim \mathrm{N}_2[(0,1), (1, \rho\sigma_\varepsilon, \sigma_\varepsilon^2)] \\ &(y_i,\mathbf{x}_i) \text{ observed only when } d_i = 1. \end{aligned} \quad (7)$$

[3] See Weinstein (1964).

[4] See Gourieroux and Monfort (1996), Train (2003), Econometric Software (2007), Greene (2008b) and Greene and Misra (2004).

Two familiar methods have been developed for estimation of the model parameters. Heckman's (1979) two step, limited information method builds on the result

$$\begin{aligned} \mathrm{E}[y_i|\mathbf{x}_i,d_i=1] &= \boldsymbol{\beta}'\mathbf{x}_i + \mathrm{E}[\varepsilon_i|d_i=1] \\ &= \boldsymbol{\beta}'\mathbf{x}_i + \rho\sigma_\varepsilon\phi(\boldsymbol{\alpha}'\mathbf{z}_i)/\Phi(\boldsymbol{\alpha}'\mathbf{z}_i) \\ &= \boldsymbol{\beta}'\mathbf{x}_i + \theta\lambda_i. \end{aligned} \tag{8}$$

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. 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 full information maximum likelihood estimator for the model is developed in Heckman (1976) and Maddala (1983). The log likelihood function for the sample selection model is

$$\begin{aligned} \log L(\boldsymbol{\beta},\sigma_\varepsilon,\alpha,\rho) &= \sum_{i=1}^{N}\log\left[\begin{array}{l} d_i\left\{\begin{array}{l}\dfrac{\exp\left(-\frac{1}{2}((y_i-\boldsymbol{\beta}'\mathbf{x}_i)^2/\sigma_\varepsilon^2)\right)}{\sigma_\varepsilon\sqrt{2\pi}}\times \\ \Phi\left(\dfrac{(\rho(y_i-\boldsymbol{\beta}'\mathbf{x}_i)/\sigma_\varepsilon)+\boldsymbol{\alpha}'\mathbf{z}_i}{\sqrt{1-\rho^2}}\right)\end{array}\right\}+ \\ (1-d_i)\Phi(-\boldsymbol{\alpha}'\mathbf{z}_i)\end{array}\right] \\ &= \sum_{i=1}^{N}\log\left[d_i\frac{1}{\sigma_\varepsilon}\phi\left(\frac{\varepsilon_i}{\sigma_\varepsilon}\right)\Phi\left(\frac{(\rho\varepsilon_i/\sigma_\varepsilon+\boldsymbol{\alpha}'\mathbf{z}_i}{\sqrt{1-\rho^2}}\right)+(1-d_i)\Phi(-\boldsymbol{\alpha}'\mathbf{z}_i)\right]. \end{aligned} \tag{9}$$

This has become a conventional, if relatively less frequently used estimator that is built into most contemporary software.

2.3 Estimating a Stochastic Frontier Model with Sample Selection.

The received literature contains many studies in which authors, have extended Heckman's selectivity model to nonlinear settings, such as count data (e.g., Poisson regression – Greene (1994)), nonlinear regression, and binary choice models. The first application of the sample selection treatment in a nonlinear setting was Wynand and van

Praag's (1981) development of a probit model for binary choice. The typical approach taken to *control for selection bias*, motivated by (8), is to fit the probit model in (7), as in the first step of Heckman's two step estimator, then append $\hat{\lambda}_i$ (from (8)) to the linear index part of the nonlinear model wherever it happens to appear. The approach is inappropriate. The term $\hat{\lambda}_i$ in (8) arises as $E[\varepsilon_i|d_i=1]$ in a linear model. The expectation of some nonlinear $g(\boldsymbol{\beta}'\mathbf{x}_i + \varepsilon_i)$ subject to selection will generally not produce the form $E[g(\boldsymbol{\beta}'\mathbf{x}_i + \varepsilon_i)|d_i=1] = g(\boldsymbol{\beta}'\mathbf{x}_i + \theta\lambda_i)$ which can then be carried back into the otherwise unchanged nonlinear model. See, e.g., Terza (1994, 1996, 1998) who develops the result in detail for nonlinear regressions such as the exponential conditional mean case. 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.

The distribution of the observed random variable conditioned on the selection will generally not be what it was without the selection (with or without the addition of the inverse Mills ratio, λ_i to the index function). Thus, the addition of λ_i to the original likelihood function generally does not produce the appropriate log likelihood in the presence of the sample selection. This can be seen even for the linear case in (9). The least squares estimator of $\boldsymbol{\beta}$ (with λ_i added to the equation) is not the MLE in (9); it is merely a feasible consistent estimator. Two well worked out specific cases do appear in the literature. Maddala (1983) and Boyes, Hoffman and Lowe (1989) obtained the appropriate closed form log likelihood for a probit model subject to sample selection. The resulting formulation is a type of bivariate probit model, not a univariate probit model based on $(\mathbf{x}_i,\lambda_i)$. Another well known example is the open form result for the Poisson regression model obtained by Terza (1996,1998).[5]

The combination of efficiency estimation and sample selection appears in several studies. Bradford, et al. (2001) 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

[5] See, also, Winkelman (1998).

process, creating an IMR from a first-stage probit estimator of the likelihood of CABG or PTCA. This correction variable is included in the frontier estimate...." (page 306).[6]

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.[7] 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, to motivate the sample selection treatment, we specify that the unobservables in the selection model are correlated with the noise in the stochastic frontier model, then the stochastic frontier model with sample selection can be cast as an extension of Heckman's specification for the linear regression model. The combination of the models in (1) and (7) is

$$
\begin{aligned}
&d_i = 1[\boldsymbol{\alpha}'\mathbf{z}_i + w_i > 0],\ w_i \sim N[0,1] \\
&y_i = \boldsymbol{\beta}'\mathbf{x}_i + \varepsilon_i,\ \varepsilon_i \sim N[0,\sigma_\varepsilon^2] \\
&(y_i,\mathbf{x}_i) \text{ observed only when } d_i = 1. \\
&\varepsilon_i = v_i - u_i \\
&u_i = |\sigma_u U_i| = \sigma_u |U_i| \text{ where } U_i \sim N[0,1] \\
&v_i = \sigma_v V_i \text{ where } V_i \sim N[0,1]. \\
&(w_i,v_i) \sim N_2[(0,1), (1, \rho\sigma_v, \sigma_v^2)]
\end{aligned}
\qquad (10)
$$

The conditional density for an observation in this specification is

[6] The authors opt for a GMM estimator based on Kopp and Mullahy's (1990) (KM) relaxation of the distributional assumptions in the standard frontier model. 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 as well, but not based on the properties of the MLE.

[7] See Battese and Coelli (1995).

$$f(y_i|x_i,|U_i|,\mathbf{z}_i,d_i,) = \left[\begin{array}{l} d_i \left\{ \begin{array}{l} \dfrac{\exp\left(-\frac{1}{2}(y_i - \boldsymbol{\beta}'\mathbf{x}_i + \sigma_u |U_i|)^2/\sigma_v^2\right)}{\sigma_v\sqrt{2\pi}} \times \\ \Phi\left(\dfrac{\rho(y_i - \boldsymbol{\beta}'\mathbf{x}_i + \sigma_u |U_i|)/\sigma_\varepsilon + \boldsymbol{\alpha}'\mathbf{z}_i}{\sqrt{1-\rho^2}}\right) \end{array} \right\} + \\ (1-d_i)\Phi(-\boldsymbol{\alpha}'\mathbf{z}_i) \end{array} \right] \tag{11}$$

Save for the appearance of the unobserved inefficiency term, $\sigma_u|U_i|$, (11) is the same as (9). Terza (1996, 2009) develops the log likelihood function for a generic extension of Heckman's result in (9) to nonlinear models. The result in (11) shows an application to the stochastic frontier case – see (34:SS) in Terza (2009).

Sample selection arises as a consequence of the correlation of the unobservables in the production or cost equation, v_i, with those in the sample selection equation, w_i. Two other applications of this general approach to modeling sample selection or endogenous switching in the stochastic frontier model have appeared in the recent literature. In Kumbhakar, Tsionas and Similainen (2009), the model framework is very similar to that in (10), but the selection mechanism is assumed to operate through u_i rather than v_i. In particular, the disturbance in their counterpart to the equation for d_i is $w_i + \delta u_i$; in essence, the inefficiency in the production process produces an "inclination" towards, in their case, organic farming. In Lai, Polachek and Wang's (2009) application to a wage equation, the w_i in the selection mechanism is correlated (through a copula function) with ε_i, not specifically with v_i or u_i. In both of these cases, the log likelihood is substantially more complicated than the one used here. More importantly, the difference in the assumption of the impact of the selection effect is substantive.

The log likelihood for the model in (10) is formed by integrating out the unobserved $|U_i|$ then maximizing with respect to the unknown parameters. Thus, as in (4) and (5),

$$\log L(\boldsymbol{\beta},\sigma_u,\sigma_v,\boldsymbol{\alpha},\rho) = \sum_{i=1}^{N} \log \int_{|U_i|} f(y_i \,|\, \mathbf{x}_i, \mathbf{z}_i, d_i, |U_i|)\, p(|U_i|) d\,|U_i|\,. \tag{12}$$

The integral in (12) is not known; it must be approximated. The simulated log likelihood function is

$$\log L_S(\boldsymbol{\beta},\sigma_u,\sigma_v,\boldsymbol{\alpha},\rho)= \sum_{i=1}^{N}\log\frac{1}{R}\sum_{r=1}^{R}\left[\begin{array}{l} d_i\left[\dfrac{\exp\left(-\frac{1}{2}(y_i-\boldsymbol{\beta}'\mathbf{x}_i+\sigma_u|U_{ir}|)^2/\sigma_v^2)\right)}{\sigma_v\sqrt{2\pi}}\times\right. \\ \left.\quad \Phi\left(\dfrac{\rho(y_i-\boldsymbol{\beta}'\mathbf{x}_i+\sigma_u|U_{ir}|)/\sigma_\varepsilon+\boldsymbol{\alpha}'\mathbf{z}_i}{\sqrt{1-\rho^2}}\right)\right] \\ +(1-d_i)\Phi(-\boldsymbol{\alpha}'\mathbf{z}_i)\end{array}\right]. \qquad (13)$$

To simplify the estimation, we will use a two step approach. The single equation MLE of $\boldsymbol{\alpha}$ in the probit equation in (7) 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 in (13), 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 in (8). Thus, the conditional simulated log likelihood function is

$$\log L_{S,C}(\boldsymbol{\beta},\sigma_u,\sigma_v,\rho)= \sum_{i=1}^{N}\log\frac{1}{R}\sum_{r=1}^{R}\left[\begin{array}{l} d_i\left\{\dfrac{\exp\left(-\frac{1}{2}(y_i-\boldsymbol{\beta}'\mathbf{x}_i+\sigma_u|U_{ir}|)^2/\sigma_v^2)\right)}{\sigma_v\sqrt{2\pi}}\times\right. \\ \left.\quad \Phi\left(\dfrac{\rho(y_i-\boldsymbol{\beta}'\mathbf{x}_i+\sigma_u|U_{ir}|)/\sigma_\varepsilon+a_i}{\sqrt{1-\rho^2}}\right)\right\} \\ +(1-d_i)\Phi(-a_i)\end{array}\right]. \qquad (14)$$

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

$$\log L_{S,C}(\boldsymbol{\beta},\sigma_u,\sigma_v,\rho) = \sum_{d_i=1}\log\frac{1}{R}\sum_{r=1}^{R}\left[\begin{array}{l}\dfrac{\exp\left(-\frac{1}{2}(y_i-\boldsymbol{\beta}'\mathbf{x}_i+\sigma_u|U_{ir}|)^2/\sigma_v^2)\right)}{\sigma_v\sqrt{2\pi}}\times \\ \Phi\left(\dfrac{\rho(y_i-\boldsymbol{\beta}'\mathbf{x}_i+\sigma_u|U_{ir}|)/\sigma_\varepsilon+a_i}{\sqrt{1-\rho^2}}\right)\end{array}\right]. \qquad (15)$$

The parameters of the model are estimated using a conventional gradient based approach, the BFGS method. We use the BHHH estimator to estimate the asymptotic standard errors for the parameter estimators. When ρ equals zero, 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.

2.4 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. (2001) for hospital costs). Researchers also compute individual specific estimates of the conditional means based on the Jondrow et al. (1982) (JLMS) result,

$$E[u_i \mid \varepsilon_i] = \frac{\sigma\lambda}{1+\lambda^2}\left[\mu_i + \frac{\phi(\mu_i)}{\Phi(\mu_i)}\right],\ \mu_i = \frac{-\lambda\varepsilon_i}{\sigma},\ \varepsilon_i = y_i - \boldsymbol{\beta}'\mathbf{x}_i. \tag{16}$$

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

$$p(u_i \mid \varepsilon_i) = \frac{p(u_i, \varepsilon_i)}{p(\varepsilon_i)} = \frac{p(\varepsilon_i \mid u_i)\,p(u_i)}{\int_{u_i} p(\varepsilon_i \mid u_i)\,p(u_i)\,du_i}. \tag{17}$$

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

$$p[(\sigma_u \mid U_i \mid) \mid \varepsilon_i] = \frac{p[(\sigma_u \mid U_i \mid), \varepsilon_i]}{p(\varepsilon_i)} = \frac{p[\varepsilon_i \mid (\sigma_u \mid U_i \mid)]\,p(\sigma_u \mid U_i \mid)}{\int_{u_i} p[\varepsilon_i \mid (\sigma_u \mid U_i \mid)]\,p(\sigma_u \mid U_i \mid)\,d(\sigma_u \mid U_i \mid)}. \tag{18}$$

The desired expectation is, then

$$E[(\sigma_u \mid U_i \mid) \mid \varepsilon_i] = \frac{\int_{\sigma_u|U_i|} (\sigma_u \mid U_i \mid)\,p[\varepsilon_i \mid (\sigma_u \mid U_i \mid)]\,p(\sigma_u \mid U_i \mid)\,d(\sigma_u \mid U_i \mid)}{\int_{\sigma_u|U_i|} p([\varepsilon_i \mid (\sigma_u \mid U_i \mid)]\,p(\sigma_u \mid U_i \mid)\,d(\sigma_u \mid U_i \mid)}. \tag{19}$$

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

$$\hat{B}_i = \frac{1}{R}\sum_{r=1}^{R}\left[\begin{array}{l}\dfrac{\exp\left(-\frac{1}{2}(y_i - \hat{\boldsymbol{\beta}}'\mathbf{x}_i + \hat{\sigma}_u|U_{ir}|)^2/\hat{\sigma}_v^2\right)}{\hat{\sigma}_v\sqrt{2\pi}}\times \\ \Phi\left(\dfrac{\hat{\rho}(y_i - \hat{\boldsymbol{\beta}}'\mathbf{x}_i + \hat{\sigma}_u|U_{ir}|)/\hat{\sigma}_\varepsilon + a_i}{\sqrt{1-\hat{\rho}^2}}\right)\end{array}\right] = \frac{1}{R}\sum_{r=1}^{R}\hat{f}_{ir} \qquad (20)$$

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

$$\hat{A}_i / \hat{B}_i = \hat{\sigma}_u \sum_{r=1}^{R}\hat{c}_{ir}|U_{ir}|, \text{ where } 0 < \hat{c}_{ir} = \frac{\hat{f}_{ir}}{\Sigma_{r=1}^{R}\hat{f}_{ir}} < 1. \qquad (21)$$

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.

The immediate advantage of this alternative approach is only that the whole set of computations is done at once, during the estimation of the parameters. However, as noted below, the estimators in (15) and (21) can be employed with other distributions for which the JLMS result in (16) is not available. The simulation estimator suggested here can, in principle, be used with any inefficiency distribution that can be simulated.

2.5 Panel Data and Other Extensions

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 in (20) that enters the log likelihood becomes

$$B_i = \frac{1}{R}\sum_{r=1}^{R}\prod_{t=1}^{T_i}\left[\begin{array}{l}\dfrac{\exp\left(-\frac{1}{2}(y_{it} - \boldsymbol{\beta}'\mathbf{x}_{it} + \sigma_u|U_{ir}|)^2/\sigma_v^2\right)}{\sigma_v\sqrt{2\pi}}\times \\ \Phi\left(\dfrac{\rho(y_{it} - \boldsymbol{\beta}'\mathbf{x}_{it} + \sigma_u|U_{ir}|)/\sigma_\varepsilon + a_i}{\sqrt{1-\rho^2}}\right)\end{array}\right] = \frac{1}{R}\sum_{r=1}^{R}\prod_{t=1}^{T_i}\hat{f}_{irt} \qquad (22)$$

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

3. Applications

In 2000, the World Health Organization published its millennium edition of the *World Health Report* (WHR) [WHO (2000).] The report contained Tandon et 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' health care system. [Seven years after its publication, the report 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 disability adjusted life expectancy, *DALE*.) "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 (2000).

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,

$$\begin{aligned}\log COMP_{it} = {} & \beta_1 + \beta_2 \log HExp_{it} + \beta_3 \log Educ_{it} \\ & + \beta_4 \log^2 Educ_{it} + \beta_5 \log^2 HExp_{it} \\ & + \beta_6 \log HExp_{it} \times \log \mathrm{Educ}_{it} - u_i + v_{it}\end{aligned} \qquad (23)$$

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 plots the TMLE 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 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, it will be difficult to make a direct comparison of the results. [The issue is examined in detail in Greene (2004).] TMLE also used an elaborate normalization based on a turn of the last century benchmark to anchor their efficiency estimates to a "minimal" level of health care. And, of course, they used a panel data (fixed effects) estimator whereas we have used a cross section. As such, it seems unlikely that the specific estimates of inefficiency would be very similar. We can, however, see whether general conclusions do hold up in the two settings. For example, if both approaches are addressing the same broad concept of efficiency relative to the production function in (23), then the rankings of countries might well be broadly similar. It is interesting to compare the rankings of countries 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 (2008b) 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 standard normal and chi squared limiting distribution used for the test 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. (We did not reestimate the TMLE values; those shown in the figure appear in the tables in the WHO report.) As anticipated in Greene (2004), the impact of the fixed effects regression is to 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. The heavy diagonal line in the figure shows the effect; save for the very largest values, the MSL estimates of $E[u_i|\varepsilon_i]$ are well below their counterparts computed using the TMLE fixed effects estimator.

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. These are based on the parameters of the selectivity model in (11) 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 Spearman rank correlation of the two series is 0.66, which seems higher than the figure would suggest. The (visually) quite weak correlation in the two sets of results conflicts with our earlier suggestion. In sum, there are a long list of substantive differences between the approach taken here and the one in TMLE. There are at least three sources of difference. First, TMLE used a fixed effects linear regression whereas we have used a stochastic frontier model. Second, we have used the time invariant variables in Table 1 to control for cross country heterogeneity whereas ETML did not make use of these. Third, we have accounted for the nonrandom sample selection in the OECD and NonOECD subsamples. None of these, alone or together should necessarily produce a change in the rankings of observations. The impacts of each source of variation might be the subject of some fruitful further analysis. The TMLE study was ultimately

focused on the ranks of the counties, not on the inefficiency levels themselves. The disparity in the ranks produced by the methods considered here should be of significant concern.

The analysis described here is essentially microeconomic, behavioral in nature. One might question the theoretical underpinnings of a behavioral model of optimization and efficiency applied to macroeconomic data such as these. Another recent study, Rahman, Wiboonpongse, Sriboonchitta and Chaovanapoonphol (2009) used the methods described in this paper to analyze production efficiency of rice producers in Thailand. In this study, the authors analyzed the switch by Thai farmers from lower quality rice varieties to a higher quality, Jasmine variety. Their sample included 207 farmers in the former group and 141 in the latter. They were able to examine the production process in much greater detail than we have here. In their results, the "correction" for selection into the high quality market produced quite marked differences in the estimated production frontier and a highly significant "selection effect."

4. Conclusions

We have proposed a maximum simulated likelihood estimator for ALS's normal – half normal stochastic frontier model. The normal–exponential model, a normal–t model, or normal–anything else model would all be trivial modifications. The manner in which the values of u_i are simulated is all that changes from one to the next. 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).] The model is then extended using Heckman's (1976) formulation for the linear model and Terza's (1986,2009) extension to nonlinear models to produce a sample selection correction for the stochastic frontier model.

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. (Note, again, the alternative approaches by Kumbhakar, Tsionas and Sipilainen (2009) and by Lai, Polachek and Wang (2009).) 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), and, by implication, Lai et al. (2009), but the analysis is tangential to the model considered here.

We have revisited the WHO (2000) study, and found that the results vary greatly depending on the specification. However, it does appear that our expectation that 'selection' on OECD membership is an important element of the measured inefficiency in the data was not supported statistically. The results suggest that the obvious pattern in Figure 1 that separates OECD from nonOECD members is explained by observables (such as per capita GDP) and not unobservables as would be implied by the sample selection model.

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^{2}Educ	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 = -8.2404 (3.369) + 0.7388LogPerCapitaGDP (0.3820)
+ 0.6098GovernmentEffectiveness (0.4388)
+ 0.7291Voice (0.3171)

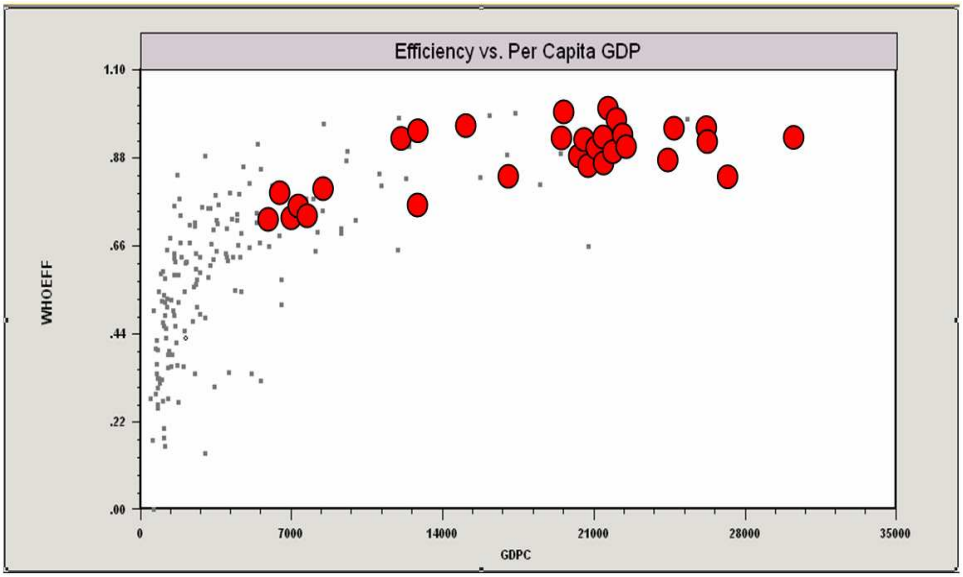

Figure 1 Efficiency Scores Related to Per Capita GDP. Larger points indicate OECD members

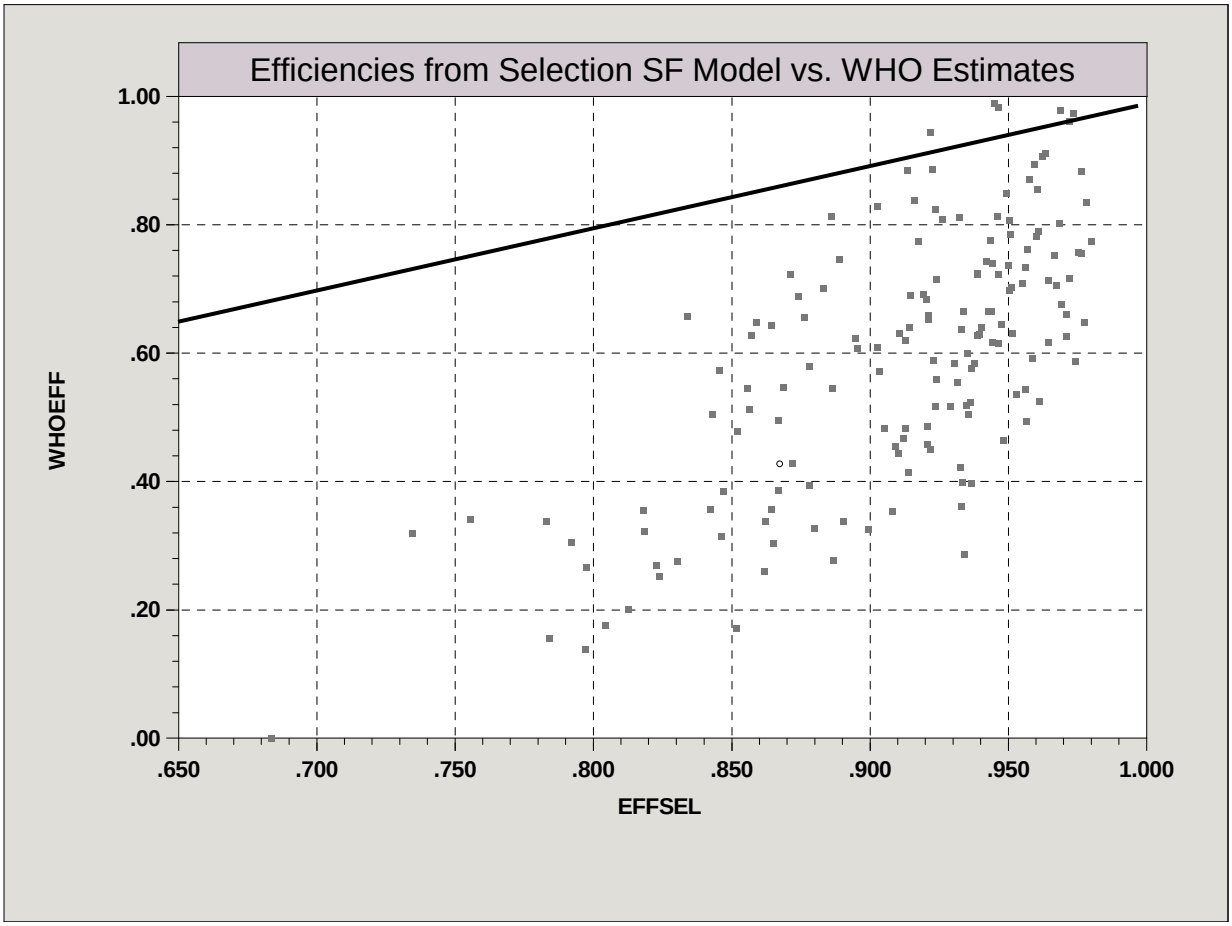

Figure 2 Estimated Efficiency Scores

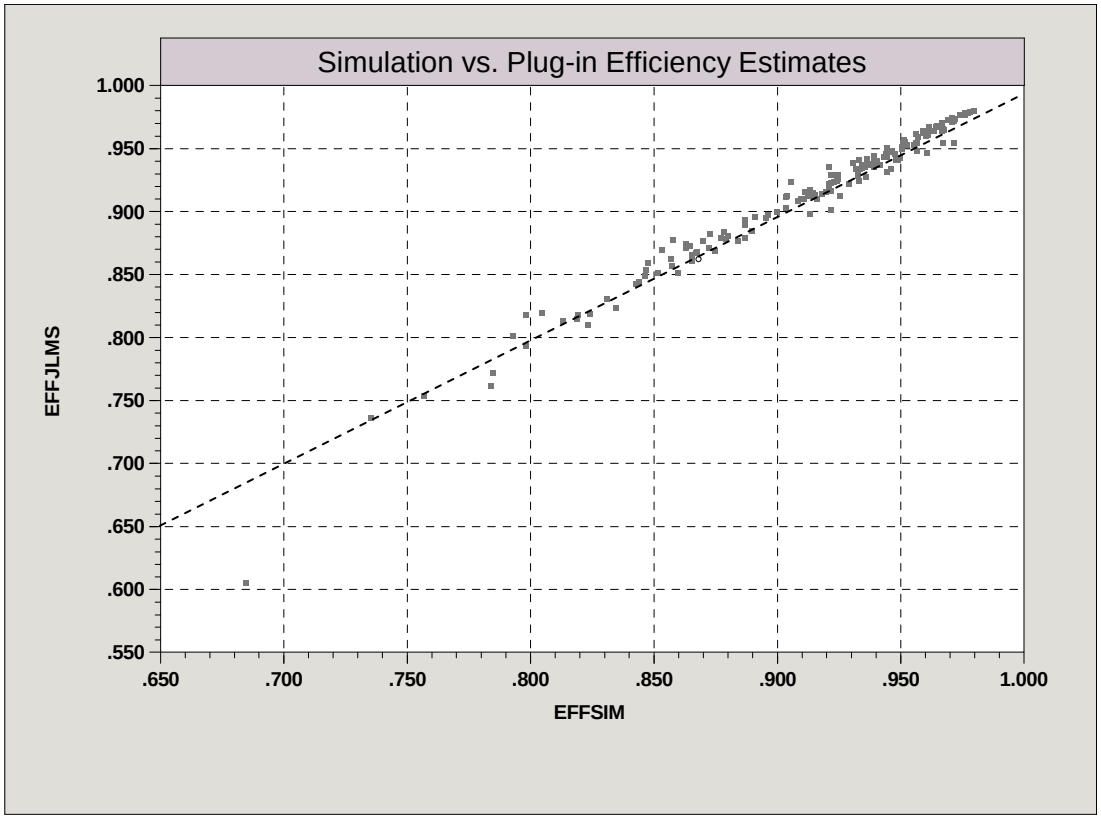

Figure 3 Alternative Estimators of Efficiency Scores

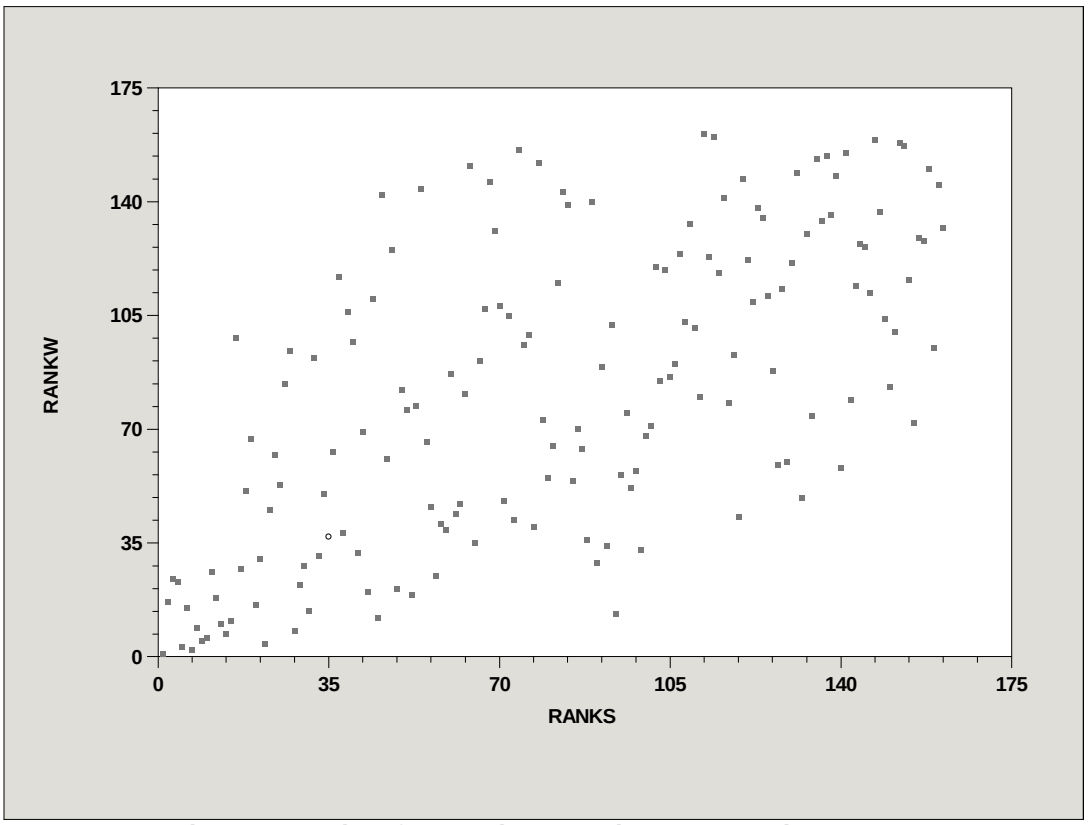

Figure 4 Ranks of Countries Based on WHO and Simulation Efficiency Estimates

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, 2001, 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 Unit, 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 Unit, Manuscript, UK, 2002b.

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

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., "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.

Greene, W. and S. Misra, "Simulated Maximum Likelihood Estimation of the Stochastic Frontier Model," Manuscript, Department of Marketing, University of Rochester, 2004.

Heckman J., "Discrete, Qualitative and Limited Dependent Variables" *Annals of Economic and Social Measurement*, 4, 5, 1976, pp. 475-492.

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

Kumbhakar, S., M. Tsionas and T. Sipilainen, "Joint Estimation of Technology Choice and Technical Efficiency: An Application to Organic and Conventional Dairy Farming," *Journal of Productivity Analysis*, 31, 3, 2009, pp. 151-162.

Lai, H., S. Polachek and H. Wang, "Estimation of a Stochastic Frontier Model with a Sample Selection Problem," Working Paper, Department of Economics, National Chung Cheng University, Taiwan, 2009.

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.

Rahman, S., A. Wiboonpongse, S. Sriboonchitta and Y. Chaovanapoonphol, 2009, "Production Efficiency of Jasmine Rice Producers in Northern and North-eastern Thailand, *Journal of Agricultural Economics*, online, pp. 1-17 (forthcoming).

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., "Modeling 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. 1994. "Dummy Endogenous Variables and Endogenous Switching in Exponential Conditional Mean Regression Models," Manuscript, Department of Economics, Penn State University.

Terza, J., "FIML, Method of Moments and Two Stage Method of Moments Estimators for Nonlinear Regression Models with Endogenous Switching and Sample Selection," Working Paper, Department of Economics, Penn State University, 1996.

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

Terza, J.V. "Parametric Nonlinear Regression with Endogenous Switching," *Econometric Reviews*, 28, 2009, pp. 555-580.

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.

Weinstein, M., 1964, `The Sum of Values from a Normal and a Truncated Normal Distribution,' *Technometrics*, 6, pp. 104-105, 469-470.

Winkelmann, R. "Count Data Models with Selectivity," *Econometric Reviews*, 4, 17, 1998, pp. 339-359.

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.

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: C13; C15; C21

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.