

# **Methods for Dealing with Endogeneity and Selection Bias in Empirical Health Policy Analysis**

**Joseph V. Terza\***

**Department of Epidemiology and Health Policy Research**

**Institute for Child Health Policy**

**Department of Economics**

**University of Florida**

**Gainesville, FL 32610-0177**

**[jvt@ichp.ufl.edu](mailto:jvt@ichp.ufl.edu)**

**(June, 2007)**

# Motivation

**At issue here is the statistical estimation of the expected value of the outcome (y) for an exogenously imposed (counterfactual) value of a variable of interest ( $x_e$ ).**

**$y_{x_e^*}$   $\equiv$  the random variable representing the value of the outcome under the imposed  $x_e = x_e^*$  counterfactual scenario**

**The object of interest then is**

**$E\left[y_{x_e^*}\right]$  = the average value of the outcome for the counterfactual scenario in which  $x_e = x_e^*$  is imposed on everyone in the population.**

# Motivation

**This is the policy analytic device that will be implemented in measuring:**

**– Treatment effects ( $x_e$  is binary)**

$$\mathbf{E}[y_1] - \mathbf{E}[y_0]$$

**– Incremental effects ( $x_e$  is not binary)**

$$\mathbf{E}[y_{x_{e2}}] - \mathbf{E}[y_{x_{e1}}]$$

**– Marginal effects ( $x_e$  is continuous)**

$$\frac{\partial \mathbf{E}[y_{x_e^*}]}{\partial x_e}$$

**Example:  $y$  = number of yearly visits to the doc,  $x_e$  = copay**

Person	$y_{15} (x_e^* = \$15)$
<b>(<math>x_e = \\$15</math>)</b>	<b>A (<math>y   x_e = \\$15</math>)</b>
1	2
2	5
3	3
4	1
5	0
6	2
7	1
8	3
<b>(<math>x_e \neq \\$15</math>)</b>	<b>B (<math>y   x_e \neq \\$15</math>)</b>
9	4
10	5
11	3
12	2
13	5
14	6

# Factual and Counterfactual Data

**Cell A contains the factual population data.**

**Cell B contains the counterfactual population data.**

**We seek to characterize the case in which we can obtain**

$$\mathbf{E}\left[\mathbf{y}_{x_c^*}\right]$$

**From the factual population data.**

## Example (Cont'd)

In this case we could obtain

$$\mathbf{E}\left[\mathbf{y}_{\mathbf{x}_e^*}\right] = \mathbf{E}\left[\mathbf{y}_{15}\right]$$

from the factual data (cell A) if the average value of  $y$  in cell A, i.e.

$$\mathbf{E}\left[\mathbf{y} \mid \mathbf{x}_e = 15\right]$$

were equal to the average value of  $y$  over both cells A and B, i.e.

$$\mathbf{E}\left[\mathbf{y}_{15}\right]$$

Clearly this is not the case

$$\mathbf{E}\left[\mathbf{y} \mid \mathbf{x}_e = 15\right] = 2.125 \quad \text{and} \quad \mathbf{E}\left[\mathbf{y}_{15}\right] = 3$$

# Using the Factual Data to Obtain $\mathbf{E}\left[\mathbf{y}_{\mathbf{x}_e^*}\right]$

Under what conditions will the factual data allow us to obtain

$$\mathbf{E}\left[\mathbf{y}_{\mathbf{x}_e^*}\right]$$

Note that for any vector of factual random variables ( $\mathbf{v}$ ) we can write

$$\mathbf{E}\left[\mathbf{y}_{\mathbf{x}_e^*}\right] = \mathbf{E}\left[\mathbf{E}\left[\mathbf{y}_{\mathbf{x}_e^*} \mid \mathbf{v}\right]\right] \tag{1}$$

Clearly then,  $\mathbf{E}\left[\mathbf{y}_{\mathbf{x}_e^*}\right]$  can be derived from the factual data if  $\mathbf{v}$  is such that

$$\mathbf{E}\left[\mathbf{y}_{\mathbf{x}_e^*} \mid \mathbf{v}\right] = \mathbf{E}\left[\mathbf{y} \mid \mathbf{x}_e = \mathbf{x}_e^*, \mathbf{v}\right] \tag{2}$$

## Using the Factual Data (Cont'd)

Now let

$$\mathbf{v} = [\mathbf{x}_o \quad \mathbf{x}_u]$$

where

$\mathbf{x}_o$  is the vector of all observable confounders

$\mathbf{x}_u$  is a vector of all unobservable confounders

Under this definition of  $\mathbf{v}$  equality (2) holds.

## Using the Factual Data (Cont'd)

The idea here is that as we condition on more and more confounders

$$\mathbf{E}\left[\mathbf{y}_{\mathbf{x}_e^*} \mid \mathbf{v}\right] \quad \text{and} \quad \mathbf{E}\left[\mathbf{y} \mid \mathbf{x}_e = \mathbf{x}_e^*, \mathbf{v}\right]$$

are brought closer to equality.

**Following up with the example, suppose we condition on income (\$50K)**

<b>Person</b>	<b><math>y_{15} (x_e^* = \\$15)</math></b>
<b><math>(x_e = \\$15, \\$50K)</math></b>	<b>A (<math>y \mid x_e = \\$15, \\$50K</math>)</b>
1	2
2	5
3	3
4	1
5	0
8	3
<b><math>(x_e \neq \\$15, \\$50K)</math></b>	<b>B (<math>y \mid x_e \neq \\$15, \\$50K</math>)</b>
9	4
10	5
11	3
12	2

## Example (Cont'd)

From the figure we can see that

$$E[y \mid x_e=15, \$50K] = 2.33$$

and

$$E[y_{15} \mid \$50K] = 2.8$$

# Estimation

Ultimately we condition on all confounders, both observable ( $\mathbf{x}_o$ ) and unobservable ( $\mathbf{x}_u$ ), and obtain exact equality between

$$E[y_{\mathbf{x}_e^*} | \mathbf{x}_o, \mathbf{x}_u]$$

and

$$E[y | \mathbf{x}_e = \mathbf{x}_e^*, \mathbf{x}_o, \mathbf{x}_u]$$

Therefore, estimation of the desired counterfactual entity hinges on our ability to estimate

$$E[y | \mathbf{x}_e, \mathbf{x}_o, \mathbf{x}_u]$$

# Estimation

The sample analog estimator of

$$E[y_{x_e^*}] = E[E[y | x_e = x_e^*, x_o, x_u]]$$

then is

$$\frac{1}{n} \sum_{i=1}^n \hat{E}[y | x_e^*, x_{oi}, x_{ui}]$$

This estimator, as it stands is infeasible b/c  $x_u$  is not observable.

We focus here on methods for estimating  $E[y | x_e, x_o, x_u]$  that include an estimator for  $x_u$ .

# The Linear Model

In this case we assume

$$E[y | \mathbf{x}_e, \mathbf{x}_o, \mathbf{x}_u] = \mathbf{x}_e \boldsymbol{\beta}_e + \mathbf{x}_o \boldsymbol{\beta}_o + \mathbf{x}_u \boldsymbol{\beta}_u \quad (1)$$

where

$\mathbf{x}_e \equiv$  the variable of interest (the endogenous variable)

$\mathbf{x}_o \equiv$  observable confounders

$\mathbf{x}_u \equiv$  unobservable confounders.

## The Linear Model (Cont'd)

The corresponding (idealized) regression model is

$$y = \mathbf{x}_e \boldsymbol{\beta}_e + \mathbf{x}_o \boldsymbol{\beta}_o + \mathbf{x}_u \boldsymbol{\beta}_u + e \quad (2)$$

where

$$e = y - (\mathbf{x}_e \boldsymbol{\beta}_e + \mathbf{x}_o \boldsymbol{\beta}_o + \mathbf{x}_u \boldsymbol{\beta}_u)$$

$$E[e | \mathbf{x}_e, \mathbf{x}_o, \mathbf{x}_u] = 0$$

**This parametric framework is designed to take explicit account of  $\mathbf{x}_u$  with the objective of avoiding the endogeneity bias inherent in models that ignore potential unobservable confounders (sometimes called *omitted variable bias*).**

# Auxiliary Regression

Without loss of generality, we assume that  $x_u$  is a scalar that comprises all unobserved variables that confound estimation of the influence of  $x_e$  on  $y$  (i.e. all other variables that influence  $y$  but are correlated with  $x_e$ )

We also define the auxiliary regression as

$$x_e = \mathbf{w}\alpha + x_u \quad (3)$$

where

$$\mathbf{w} = [x_0 \quad \mathbf{w}^+]$$

$x_0$  is a  $1 \times K$  vector, and  $\mathbf{w}^+ = [w_1^+, w_2^+, \dots, w_{S^+}^+]$  is a  $1 \times S^+$  vector of instrumental variables.

# Instrumental Variables (IV)

*Instrumental variables* satisfy the following three conditions:

1) they are not correlated with  $x_u$ , specifically

$$E[x_u | w] = 0 \rightarrow \text{cov}(x_u, w) = 0 \quad (4)$$

2) They are sufficiently correlated with  $x_e$ ;

in other words,  $\text{cov}(x_e, w) \neq 0$  (sufficiently)

in other words, they are not weak instruments

3) aside from  $x_o$  and  $x_e$ , they are not correlated with  $y$

$$E[y | x_e, x_o, x_u] = E[y | x_e, w, x_u] = x_e \beta_e + x_o \beta_o + x_u \beta_u$$

For the purpose of identification it must be true that  $S+ \geq 1$ .

## IV Estimation – Two-Stage Predictor Substitution (2SPS)

### *First Stage*

Estimate  $\alpha$  by applying ordinary least squares (OLS) to the auxiliary regression (3).

OLS is consistent for  $\alpha$  by the first IV condition –  $E[x_u | w] = 0$

Then compute the “predictor” of  $x_e$  as

$$\hat{x}_e = w\hat{\alpha} \tag{5}$$

where  $\hat{\alpha}$  denotes the first stage estimate of  $\alpha$ .

## IV Estimation – 2SPS (Cont'd)

### *Second Stage*

Estimate  $\beta^{*'} = [\beta_e' \quad \beta_o']$  by applying OLS to

$$y = \hat{\mathbf{x}}_e \boldsymbol{\beta}_e + \mathbf{x}_o \boldsymbol{\beta}_o + e^{2SPS} \quad (6)$$

where  $e^{2SPS}$  denotes the regression error term.

## IV Estimation – Two-Stage Residual Inclusion (2SRI)

*First Stage:* Same as 2SPS.

*Second Stage:*

Apply OLS to

$$y = \mathbf{x}_e \boldsymbol{\beta}_e + \mathbf{x}_0 \boldsymbol{\beta}_0 + \hat{\mathbf{x}}_u \boldsymbol{\beta}_u + e^{2SRI} \quad (7)$$

where  $e_{2SRI}$  is the regression error term

$$\hat{\mathbf{x}}_u = \mathbf{x}_e - \mathbf{w} \hat{\alpha}$$

where  $\hat{\alpha}$  denotes the first stage estimate of  $\alpha$ .

## Notes on 2SPS and 2SRI

– 2SPS  $\equiv$  2SRI and the methods are consistent.

**This is not the case in the generic nonlinear context.**

-- 2SPS is most commonly referred to as Two-Stage Least Squares (2SLS)

– 2SRI is not new.

**First proposed by Hausman (*Econometrica*, 1978) as a means of directly testing for endogeneity in linear models.**

**The estimated coefficient of  $\beta_u$  can be used to test for endogeneity.**

**Exogeneity null is  $H_0: \beta_u = 0$ .**

# Characterization and Estimation of $E[y_{x_e^*}]$

In this case the object of interest is

$$E[y_{x_e^*}] = E[E[y | x_e = x_e^*, x_o, x_u]] = E[x_e^* \beta_e + x_o \beta_o + x_u \beta_u]$$

and the sample analog estimator is

$$\frac{1}{n} \sum_{i=1}^n \hat{E}[y | x_e^*, x_{oi}, x_{ui}] = x_e^* \hat{\beta}_e + \bar{x}_o \hat{\beta}_o + \bar{x}_u \hat{\beta}_u$$

Therefore in analyzing policy on the estimate of  $\beta_e$  is required.

# Policy Analysis

– Treatment effects ( $x_e$  is binary)

$$\mathbf{E}[y_1] - \mathbf{E}[y_0] = \beta_e$$

– Incremental effects ( $x_e$  is not binary)

$$\mathbf{E}[y_{x_{e2}}] - \mathbf{E}[y_{x_{e1}}] = (x_{e2} - x_{e1})\beta_e$$

– Marginal effects ( $x_e$  is continuous)

$$\frac{\partial \mathbf{E}[y_{x_e^*}]}{\partial x_e} = \beta_e$$

# Nonlinear Estimation with Endogenous Regressors

Now extend the linear model in the following way:

**Conditional Mean Function**

$$E[y | \mathbf{x}_e, \mathbf{x}_o, \mathbf{x}_u] = M(\mathbf{x}_e\boldsymbol{\beta}_e + \mathbf{x}_o\boldsymbol{\beta}_o + \mathbf{x}_u\boldsymbol{\beta}_u) \quad (8)$$

Where  $M(\cdot)$  is a known nonlinear function.

**Common examples:**

$M(\cdot) = \exp(\cdot)$  [count data models, duration models]

$M(\cdot) = \text{logistic}(\cdot)$  [logistic regression models]

$M(\cdot) = \text{stnormal}(\cdot)$  [probit models]

$M(\cdot) = \text{ibc}(\cdot)$  [inverse box cox models]

Where  $\text{logistic}(\cdot)$  and  $\text{stnormal}(\cdot)$  are the logistic and standard normal distribution functions, respectively.

## Nonlinear Models (Cont'd)

The corresponding (idealized) regression model is

$$y = \mathbf{M}(\mathbf{x}_e\boldsymbol{\beta}_e + \mathbf{x}_o\boldsymbol{\beta}_o + \mathbf{x}_u\boldsymbol{\beta}_u) + e \quad (9)$$

The auxiliary regression

$$\mathbf{x}_e = \mathbf{r}(\mathbf{w}\boldsymbol{\alpha}) + \mathbf{x}_u \quad (10)$$

where  $\mathbf{r}(\ )$  is a known (possibly nonlinear) function.

# Nonlinear Models – 2SPS

## *First Stage*

Estimate  $\alpha$  by applying nonlinear least squares (NLS) [or some other appropriate nonlinear estimation method – e.g. MLE] to the auxiliary regression (10).

NLS is consistent for  $\alpha$  by the first IV condition –  $E[\mathbf{x}_u | \mathbf{w}] = \mathbf{0}$

Then compute the “predictor” of  $x_e$  as

$$\hat{\mathbf{x}}_e = \mathbf{r}(\mathbf{w}\hat{\alpha}) \tag{11}$$

where  $\hat{\alpha}$  denotes the first stage estimate of  $\alpha$ .

## Nonlinear Models – 2SPS (Cont'd)

### *Second Stage*

Estimate  $\beta^{*'}$  =  $[\beta_e' \quad \beta_o']$  by applying NLS to

$$y = M(\hat{x}_e \beta_e + x_o \beta_o) + e^{2SPS*} \quad (12)$$

where  $e^{2SPS*}$  denotes the regression error term.

## **Nonlinear Models – 2SPS (Cont'd)**

***Applications of the 2SPS method in nonlinear health econometric contexts can be found in:***

**Burgess, S., Gregg, P., Propper, C., Wasgbrook, E., and ALSPAC Study Team (2002) Working Paper.**

**\*Bollen, K.A., Guilkey, D.K., and Mroz, T.A. (1995) *Demography***

**Cawley, J. (2000) *Health Services Research*.**

**Ettner, S.L., Hermann, R.C., and Tang, H. (1999) *Health Services Research***

**Fox, M. (2002) *Health Economics***

**French, M.T., Roebuck, M.C., McGeary, K.A., Chitwood, D.D., and McCoy, C.B. (2001) *Substance Use & Misuse*,**

**French, M.T., Roebuck, M.C., and Alexandre, P.K. (2001) *Southern Economic Journal***

## **Nonlinear Models – 2SPS (Cont'd)**

**Holmes, A.M., and Deb, P. (1998) *Health Services Research***

**Howard, D. (2000) *Health Services Research***

**Lu, M., and McGuire, T.G. (2002) *Journal of Human Resources***

**Meer, J., and Rosen, H.S. (2003) Working Paper - National Bureau of  
Economic Research, #9812.**

**\*Mroz, T.A., Bollen, K.A., Speizer, I.S., and Mancini, D.J. (1999)  
*Demography,***

**\*Norton, E.C., Lindrooth, R.C., and Ennett, S.T. (1998) *Health Economics***

**Register, C.A., and Williams, D.R. (1992) *Industrial and Labor Relations  
Review***

**Savage, E., and Wright, D.J. (2003) *Journal of Health Economics***

**Sen, B. (2002) *Journal of Health Economics***

## **Nonlinear Models – 2SPS (Cont'd)**

**It can be shown that in the class of models on which we are focusing, 2SPS is not, in general, consistent.**

**See**

**Terza, J.V., Basu, A. and Rathouz, P.J. (2007): “Addressing Endogeneity in Nonlinear Parametric Models: A Guide for Empirical Research in Health Economics” Under review.**

## **Nonlinear Models – 2SPS (Cont'd)**

**Exceptions:**

**1) 2SPS is consistent when outcome equation is linear.**

**This fact has been recognized in the literature.**

**Heckman (*Econometrica*, 1978), Maddala (*Textbook*, 1983), and Dubin and McFadden (*Econometrica*, 1984) all suggest consistent 2SPS estimators for linear models with endogenous binary variables.**

**In the first stage, nonlinear qualitative dependent variable models are estimated as auxiliary equations (e.g. probits).**

**In the second stage, the first-stage predicted probabilities replace the endogenous dummies in OLS estimation of the linear outcome equation.**

## **Nonlinear Models – 2SPS (Cont'd)**

### **Exceptions (Cont'd):**

**– 2SPS Consistent if the following three conditions hold:**

- 1) the outcome of interest ( $y$ ) is limited in range or qualitative (e.g. Tobit- or probit- type models)**
- 2) the endogenous regressor ( $x_e$ ) is continuous and has unrestricted support**
- 3) the auxiliary equation is linear as in (3)**
- 4) conditional on all observable confounders,  $y$  and  $x_e$  are joint normally distributed.**

## **Nonlinear Models – 2SPS (Cont'd)**

**Exceptions (Cont'd):**

### **Tobit and Probit Models**

**Nelson and Olson (*International Economic Review*, 1978)**

**Lee (*Econometrica*, 1979; *Textbook w/ Manski*, 1981)**

### **Lee's Estimator Implemented by:**

**Bollen et al. (*Demography*, 1995)**

**Norton et al. (*Health Economics*, 1998)**

**Mroz et al. (*Demography*, 1999)**

**NOTE THAT AMONG THE STUDIES LISTED ON SLIDES 14 AND 15, CONSISTENCY CAN ONLY BE ARGUED FOR THESE THREE.**

# Nonlinear Models – 2SRI

*First Stage:* Same as 2SPS.

*Second Stage:*

Apply NLS to

$$y = M(\mathbf{x}_e \boldsymbol{\beta}_e + \mathbf{x}_o \boldsymbol{\beta}_o + \hat{\mathbf{x}}_u \boldsymbol{\beta}_u) + e^{2SRI*} \quad (13)$$

where  $e^{2SRI*}$  is the regression error term

$$\hat{\mathbf{x}}_u = \mathbf{x}_e - r(w\hat{\alpha})$$

and  $\hat{\alpha}$  denotes the first stage estimate of  $\alpha$ .

## **Nonlinear Models – 2SRI (Cont'd)**

**It can be shown that in the class of models on which we are focusing, 2SRI is consistent.**

**See**

**Terza, J.V., Basu, A. and Rathouz, P.J. (2007): “Addressing Endogeneity in Nonlinear Parametric Models: A Guide for Empirical Research in Health Economics” Under review.**

## **Nonlinear Models – 2SRI (Cont'd)**

**– Examples of the use of the 2SRI method in health economics can be found in:**

**Baser et al. (*Health Services and Outcomes Research Methodology*, 2004)**

**DeSimone (*Journal of Labor Economics*, 2002)**

**Gibson et al. (*American Journal of Managed Care*, 2006)**

**Norton and Van Houtven (*Southern Economic Journal*, 2006)**

**Shea, Terza, et al. (*Health Services Research*, 2007)**

**Bollen et al. (*Demography*, 1995)**

**– Other applications of 2SRI methods (not in health economics or health services research) are:**

**Burnett (*Journal of Economic Education*, 1997)**

**Alvarez and Glasgow (*Political Analysis*, 1999)**

**McGarrity and Sutter (*Southern Economic Journal*, 2000) .**

# Consistency Properties of the Estimators

- **As we said earlier: In the linear case  $2SLS \equiv 2SPS \equiv 2SRI$ , moreover,  $2SLS$  is consistent.**
- **Therefore  $2SPS$  and  $2SRI$  are both consistent in the linear case**
- **The identities do not hold for the generic nonlinear case**
- **As we have shown  $2SRI$  is consistent but  $2SPS$  is not.**

# Alternative Estimators

– Approximate the model as linear and apply the conventional instrumental variables method

Unfortunately, this can lead to serious bias:

See

**Terza, J., Bradford, W.D. and Dismuke, C.E (2006): “The Use of Linear Instrumental Variables Methods in Health Economic Research: A Cautionary Note,” Working Paper.**

## Alternative Estimators (Cont'd)

– Use the Generalized Method of Moments (GMM)

Works for additive models of the form

$$E[y | \mathbf{x}_e, \mathbf{x}_o, \mathbf{x}_u] = \mathbf{M}' (\mathbf{x}_e \boldsymbol{\beta}_e + \mathbf{x}_o \boldsymbol{\beta}_o) + \mathbf{x}_u \boldsymbol{\beta}_u$$

but generally not feasible for “symmetric” models like the ones we are focusing on here

See

Terza, J. (2006): “Estimation of Policy Effects Using Parametric Nonlinear Models: A Contextual Critique of the Generalized Method of Moments,” *Health Services and Outcomes Research Methodology*, 6, 177-198.

Only exception is  $M(\cdot) \equiv \exp(\cdot)$

Mullahy, J. (1997): “Instrumental-Variable Estimation of Count Data Models: Applications to Models of Cigarette Smoking Behavior,” *Review of Economics and Statistics*, 79, 586-593.

# Characterization and Estimation of $E[y_{x_e^*}]$

In this case the object of interest is

$$E[y_{x_e^*}] = E[E[y | x_e = x_e^*, x_o, x_u]] = E[M(x_e^* \beta_e + x_o \beta_o + x_u \beta_u)]$$

and the sample analog estimator is

$$\frac{1}{n} \sum_{i=1}^n \hat{E}[y | x_e^*, x_{oi}, x_{ui}] = \frac{1}{n} \sum_{i=1}^n M(x_e^* \hat{\beta}_e + x_{oi} \hat{\beta}_o + \hat{x}_{ui} \hat{\beta}_u)$$

# Policy Analysis

– **Treatment effects ( $x_e$  is binary)**

$$\mathbf{E}[y_1] - \mathbf{E}[y_0] = \mathbf{E}[\mathbf{M}(\beta_e + x_o\beta_o + x_u\beta_u) - \mathbf{M}(x_o\beta_o + x_u\beta_u)]$$

– **Incremental effects ( $x_e$  is not binary)**

$$\mathbf{E}[y_{x_{e2}}] - \mathbf{E}[y_{x_{e1}}] = \mathbf{E}[\mathbf{M}(x_{e2}\beta_e + x_o\beta_o + x_u\beta_u) - \mathbf{M}(x_{e1}\beta_e + x_o\beta_o + x_u\beta_u)]$$

– **Marginal effects ( $x_e$  is continuous)**

$$\mathbf{E}\left[\frac{\partial \mathbf{M}(x_e\beta_e + x_o\beta_o + x_u\beta_u)}{\partial x_e}\right]_{x_e=x_e^*} = \mathbf{E}[\beta_e \mathbf{M}'(x_e^*\beta_e + x_o\beta_o + x_u\beta_u)]$$

# Sample Analog Estimators

– Treatment effects ( $x_e$  is binary)

$$\sum_{i=1}^n \frac{1}{n} \left[ M(\hat{\beta}_e + x_{oi} \hat{\beta}_o + \hat{x}_{ui} \hat{\beta}_u) - M(x_{oi} \hat{\beta}_o + \hat{x}_{ui} \hat{\beta}_u) \right]$$

– Incremental effects ( $x_e$  is not binary)

$$\sum_{i=1}^n \frac{1}{n} \left[ M(x_{e2} \hat{\beta}_e + x_{oi} \hat{\beta}_o + \hat{x}_{ui} \hat{\beta}_u) - M(x_{e1} \hat{\beta}_e + x_{oi} \hat{\beta}_o + \hat{x}_{ui} \hat{\beta}_u) \right]$$

– Marginal effects ( $x_e$  is continuous)

$$\sum_{i=1}^n \frac{1}{n} \left[ \hat{\beta}_e M'(x_e^* \hat{\beta}_e + x_{oi} \hat{\beta}_o + \hat{x}_{ui} \hat{\beta}_u) \right]$$

# Summary

- **Want to correct for endogenous regressors in a generic nonlinear regression model**

$$E[y | \mathbf{x}_e, \mathbf{x}_o, \mathbf{x}_u] = M(\mathbf{x}_e\boldsymbol{\beta}_e + \mathbf{x}_o\boldsymbol{\beta}_o + \mathbf{x}_u\boldsymbol{\beta}_u)$$

- **Five approaches available:**
- **Full Information Maximum Likelihood Estimation: Requires strong assumptions.**
- **2SPS: Very popular but not consistent**
- **2SRI: Consistent but not very popular**
- **Linearized Conventional IV: Inconsistent in inherently nonlinear models**
- **GMM: Not feasible for nearly all specifications of  $M(\cdot)$**

# Correcting for Sample Selection in Linear Models

**Heckman's Model (Heckman, J. (1976): "The Common Structure of Statistical Models of Truncation Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models," *Annals of Economic and Social Measurement*, 5, 475-492.**

**Two equations:**

**Outcome equation:**

$$y = x_0 \delta_0 + \varepsilon_1$$

**Sample selection criterion (y is only observable if d = 1):**

$$d = I(w\alpha + \varepsilon_2 > 0)$$

**where  $(\varepsilon_1, \varepsilon_2 | w)$  is bivariate normally distributed with mean vector 0 and covariance matrix**

$$\begin{pmatrix} \sigma^2 & \rho \\ \rho & 1 \end{pmatrix}$$

**and  $I(A)$  takes the value 1 if condition A holds, 0 otherwise.**

## Correcting for Sample Selection in Linear Models (Cont'd)

Heckman (1976) showed that:

$$E[y | \mathbf{w}, \mathbf{d} = 1] = \mathbf{x}_0 \boldsymbol{\delta}_0 + \sigma \rho \lambda \quad (14)$$

where

$$\lambda = \frac{\varphi(\mathbf{w}\boldsymbol{\alpha})}{1 - \Phi(\mathbf{w}\boldsymbol{\alpha})}$$

## Correcting for Sample Selection in Linear Models (Cont'd)

**Two-Stage Estimator:**

*First Stage*

Estimate  $\alpha$  via probit analysis (using the full sample)

*Second Stage*

Estimate  $\delta$  and  $\theta = \sigma\rho$  by applying OLS to the following equation (using the subsample for whom  $d=1$ )

$$y = x_0 \delta_0 + \theta \hat{\lambda} + v$$

where

$$\lambda = \frac{\varphi(w\hat{\alpha})}{1 - \Phi(w\hat{\alpha})}$$

$\hat{\alpha}$  is the first-stage estimate of  $\alpha$ , and  $v$  is the random error term.

## Correcting for Sample Selection in Linear Models (Cont'd)

**Terza, J.V. (2007): "Parametric Nonlinear Regression with Endogenous Switching," Under review.**

**Shows that the same regression formulation can be derived without the joint normality assumption. The fundamentals of the model are:**

**Outcome equation:**

$$E[y | \mathbf{x}_o, \mathbf{x}_u] = \mathbf{x}_o\boldsymbol{\beta}_o + \mathbf{x}_u\boldsymbol{\beta}_u \quad (15)$$

**Sample selection criterion (y is only observable if  $d = 1$ ):**

$$d = I(w\alpha + \mathbf{x}_u > 0) \quad (16)$$

**where  $(\mathbf{x}_u | w)$  is standard normally distributed.**

## Correcting for Sample Selection in Linear Models (Cont'd)

Terza (2007) shows that (15) and (16) imply:

$$E[y | \mathbf{w}, \mathbf{d} = 1] = \mathbf{x}_0 \boldsymbol{\beta}_0 + \lambda \boldsymbol{\beta}_u$$

which can be consistently estimated using Heckman's two-stage method.

Note that joint normality is replaced by assumption (15).

## Correcting for Sample Selection in Nonlinear Models

Now consider the generic nonlinear case in which

Outcome equation:

$$E[y | \mathbf{x}_o, \mathbf{x}_u] = \mathbf{J}(\mathbf{x}_o\boldsymbol{\beta}_o + \mathbf{x}_u\boldsymbol{\beta}_u) \quad (17)$$

where the selection criterion is defined as in (16) where  $(\mathbf{x}_u | \mathbf{w})$  has known distribution function (not necessarily standard normal).

In this case Terza (2007) shows that

$$E[y | \mathbf{w}, \mathbf{d} = 1] = \frac{\int_{-\infty}^{\infty} \mathbf{J}(\mathbf{x}_o\boldsymbol{\beta}_o + \mathbf{x}_u\boldsymbol{\beta}_u)g(\mathbf{x}_u)d\mathbf{x}_u}{G(-\mathbf{w}\boldsymbol{\alpha})} \quad (18)$$

$g(\cdot)$  and  $G(\cdot)$  are the density and distribution functions of  $(\mathbf{x}_u | \mathbf{w})$ , respectively.

# Correcting for Sample Selection in Nonlinear Models

**Two-Stage Estimator:**

*First Stage*

Estimate  $\alpha$  via appropriate MLE corresponding to  $G(\cdot)$  [using the full sample]

*Second Stage*

Estimate  $\delta$  and  $\theta = \sigma\rho$  by applying NLS to the following equation (using the subsample for whom  $d=1$ )

$$y = \frac{\int_{-\infty}^{\infty} \mathbf{J}(\mathbf{x}_o\beta_o + \mathbf{x}_u\beta_u)g(\mathbf{x}_u)d\mathbf{x}_u}{G(-w\hat{\alpha})} + \zeta \quad (19)$$

where  $\hat{\alpha}$  is the first-stage estimate of  $\alpha$ .

## The Exponential Case [ $J(\cdot) = \exp(\cdot)$ ]

As an variant of the model considered by:

**Terza, J.V. (1998): "Estimating Count Data Models with Endogenous Switching: Sample Selection and Endogenous Treatment Effects," *Journal of Econometrics*, 84,129-154.**

**Terza (2007) shows that when  $J(\cdot) = \exp(\cdot)$  in the above model, and  $(x_u | w)$  is standard normally distributed, (18) becomes**

$$E[y | w, d = 1] = \exp(x_0 \beta_0^+) \frac{\Phi(w\alpha + \beta_u)}{\Phi(w\alpha)}$$

**where  $\beta_0^+$  is the same as  $\beta_0$  except for a shift in the constant term.**

## The Exponential Case [ $J(\cdot) = \exp(\cdot)$ ] (Cont'd)

**Two-Stage Estimator:**

*First Stage*

Estimate  $\alpha$  via probit analysis (using the full sample)

*Second Stage*

Estimate  $\beta_o^+$  and  $\beta_u$  by applying NLS to the following equation (using the subsample for whom  $d=1$ )

$$y = \exp(x_o \beta_o^+) \frac{\Phi(w\hat{\alpha} + \beta_u)}{\Phi(w\hat{\alpha})} + \kappa \quad (20)$$

where  $\hat{\alpha}$  is the first-stage estimate of  $\alpha$ ; and  $\kappa$  is the random error term.

A variant of this method is programmed in Stata<sup>®</sup>

See

Miranda, A. (2004): "FIML Estimation of an Endogenous Switching Model for Count Data," *Stata Journal*, 4, pp. 40–49

## **Correcting for Sample Selection in Nonlinear Models (Cont'd)**

**The generic two-stage estimator culminating in (19), and variants thereof, have been applied in a number of empirical contexts:**

**Terza, J.V. (2002), “Alcohol Abuse and Employment: A Second Look”  
*Journal of Applied Econometrics*, 17, (2002), 393-404.**

**Treglia, M, Neslusan, C.A., Dunn R.L. (1999): “Fluoxetine and Dothiepin  
Therapy in Primary Care and Health Resource Utilization:  
Evidence from the United Kingdom,” *International Journal of  
Psychiatry in Clinical Practice*, 3, 23-30.**

**Kenkel, D.S., and Terza , J.V. (2001): "The Effect of Physician Advice on  
Alcohol Consumption: Count Regression with an Endogenous  
Treatment Effect," *Journal of Applied Econometrics*,16, (2001), 165-  
184.**

## **Applications (Cont'd)**

**Koc, C. (2005): “Health-Specific Moral Hazard Effects,” *Southern Economic Journal*, 72, 98-118.**

**McGeary, K.A., and M. T. French (2000): "Illicit Drug Use and Emergency Room Utilization," *Health Services Research*, 35, 153-169.**

**Neslusan CA, Hylan TR, Dunn RL, Donoghue J. (1999): “Controlling for Systematic Selection in Retrospective Analyses: An Application to Fluoxetine and Sertraline Prescribing in the United Kingdom,” *Value in Health*, 2, 435-445.**

**Pryor, C., and Terza, J. (2002): "Are Going Concern Audit Opinions a Self-Fulfilling Prophecy?" *Advances in Quantitative Analysis of Finance and Accounting*, 10, 89-116.**