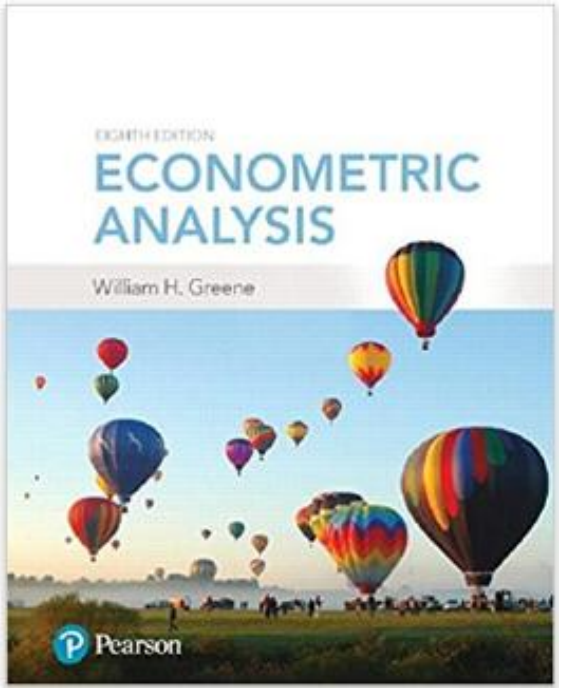
Econometrics I

Professor William Greene
Stern School of Business
Department of Economics



Econometrics I

Part 19 – Sample Selection
Two Step Estimation

★☆☆☆☆ **This Book is Seriously Flawed**, November 19, 2012

By **Dennis Hanseman** (Cincinnati, OH United States) - See all my reviews

REAL NAME

Amazon Verified Purchase (What's this?)

This review is from: Applied Econometrics for Health Economists: A Practical Guide (Paperback)

After paying over $30 for a 115-page book, I was shocked to find that it was seriously flawed. All of the analysis in the book is based on data from the British "Health and Lifestyle Survey". As Jones points out in Chapter 2, this Survey employed a complex sample design that incorporated stratification, clustering, and -- presumably -- unequal probabilities of selection. Rather than taking the design characteristics into account, Jones analyzes this data as if it came from a simple random sample. As a result, his estimates are likely to be biased, with overstated significance levels.

Dueling Selection Biases – From two emails, same day.

- "I am trying to find methods which can deal with **data that is non-randomised and suffers from selection bias**."
- "I explain the probability of answering questions using, among other independent variables, a variable which measures knowledge breadth. Knowledge breadth can be constructed only for those individuals that fill in a skill description in the company intranet. **This is where the selection bias comes from.**

Samples and Populations

- Consistent estimation
  - The sample is randomly drawn from the population
  - Sample statistics converge to their population counterparts
- A presumption: The ‘population’ is the population of interest.
- Implication: If the sample is randomly drawn from a specific subpopulation, statistics converge to the characteristics of that subpopulation

Nonrandom Sampling

- Simple nonrandom samples: Average incomes of airport travelers → mean income in the population as a whole?
- Survivorship: Time series of returns on business performance. Mutual fund performance. (Past performance is no guarantee of future success.)
- Attrition: Drug trials. Effect of erythropoetin on quality of life survey.
- Self-selection:
  - Labor supply models
  - Shere Hite's (1976) "The Hite Report" 'survey' of sexual habits of Americans. "While her books are ground-breaking and important, they are based on flawed statistical methods and one must view their results with skepticism."

The Crucial Element

- Selection on the unobservables
  - Selection into the sample is based on both observables and unobservables.
  - All the observables are accounted for.
  - Unobservables in the selection rule also appear in the model of interest (or are correlated with unobservables in the model of interest).
- “Selection Bias” = the bias due to not accounting for the unobservables that link the equations.

Heckman's Canonical Model

A behavioral model:

Offered wage = $o^* = \beta'x+v$ (x = age,experience,educ...)

Reservation wage = $r^* = \delta'z + u$ (z = age, kids, family stuff)

Labor force participation:

$$LFP = 1 \text{ if } o^* \geq r^*, 0 \text{ otherwise}$$

$$\text{Prob}(LFP=1) = \Phi\left[(\beta'x-\delta'z)/\sqrt{\sigma_v^2 + \sigma_u^2}\right]$$

Desired Hours = $H^* = \gamma'\mathbf{w} + \varepsilon$

Actual Hours = H^* if LFP = 1

unobserved if LFP = 0

ε and u are correlated. ε and v might be correlated.

What is $E[H^* \mid w, LFP = 1]$? Not $\gamma'\mathbf{w}$.

Standard Sample Selection Model

$d_i^* = \alpha'\mathbf{z}_i + u_i$

$d_i = 1(d_i^* > 0)$

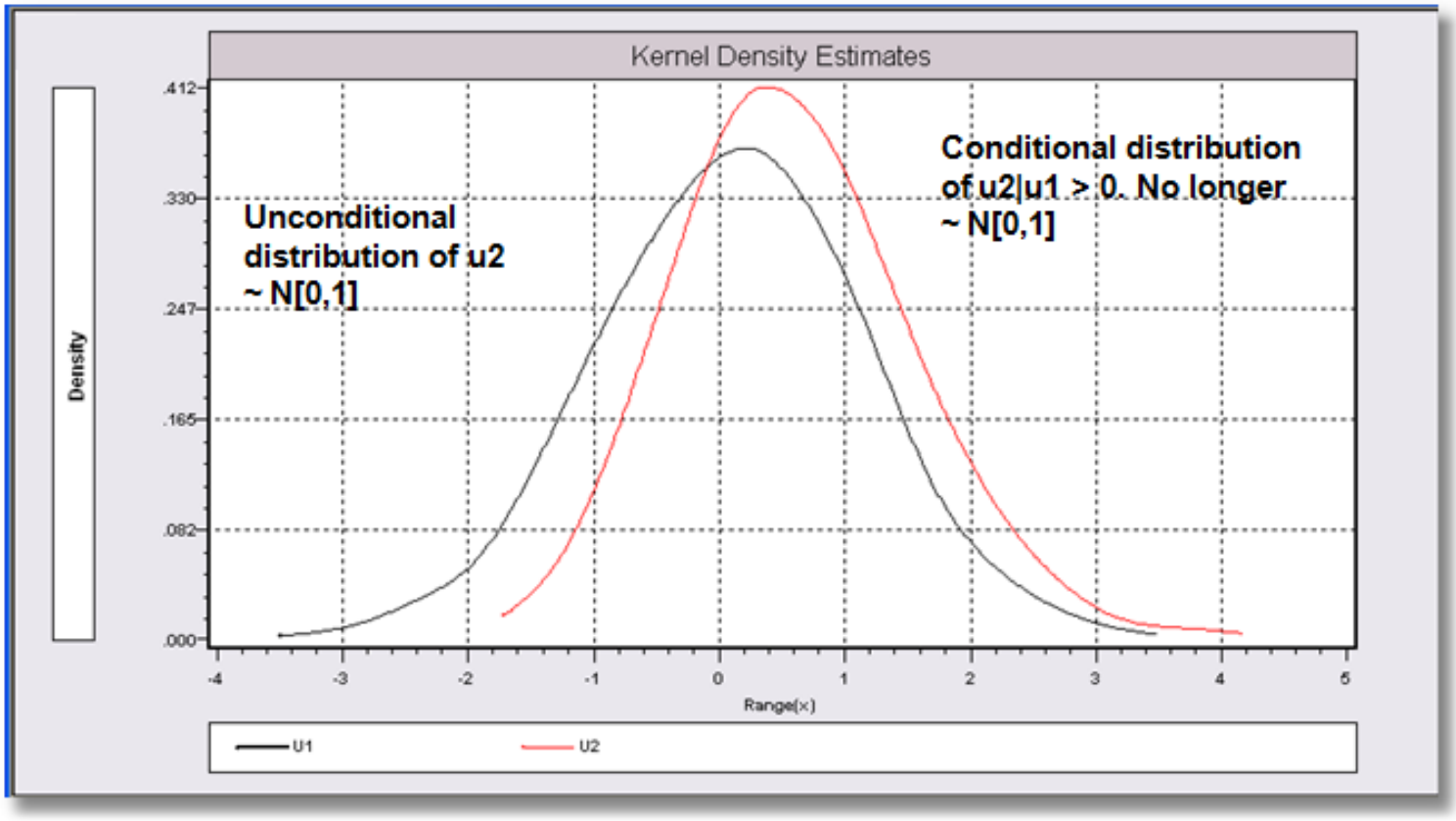
$y_i^* = \beta'\mathbf{x}_i + \varepsilon_i$

$y_i = y_i^*$ when $d_i = 1$, unobserved otherwise

$(u_i, v_i) \sim$ Bivariate Normal$[(0,0),(1,\rho\sigma,\sigma^2)]$

$$
\begin{aligned}
E[y_i \mid y_i \text{ is observed}] &= E[y_i \mid d_i = 1] \\
&= \beta'\mathbf{x}_i + E[\varepsilon_i \mid d_i = 1] \\
&= \beta'\mathbf{x}_i + E[\varepsilon_i \mid u_i > -\alpha' z_i] \\
&= \beta'\mathbf{x}_i + (\rho\sigma)\frac{\phi(\alpha'\mathbf{z}_i)}{\Phi(\alpha'\mathbf{z}_i)} \\
&= \beta'\mathbf{x}_i + \theta\lambda_i
\end{aligned}
$$

Incidental Truncation
u1,u2~N[(0,0),(1,.71,1)



Selection as a Specification Error

- $E[y_i|\mathbf{x}_i, y_i \text{ observed}] = \boldsymbol{\beta}'\mathbf{x}_i + \theta\lambda_i$
- Regression of y_i on $\mathbf{x}_i$ omits λ_i.
  - λ_i will generally be correlated with $\mathbf{x}_i$ if $\mathbf{z}_i$ is.
  - $\mathbf{z}_i$ and $\mathbf{x}_i$ often have variables in common.
  - There is no specification error if $\theta = 0 \Leftrightarrow \rho = 0$
- "Selection Bias" is plim $(\mathbf{b} - \boldsymbol{\beta})$
- What is "selection bias…"

Control Function

Labor Force Participation

$d^* = \boldsymbol{\alpha}'\mathbf{z} + u$

What is u? Unmeasured factors that motivate LFP, u = (m,a)

Desired Hours

$H^* = \boldsymbol{\beta}'\mathbf{x} + \varepsilon$

What is ε? Unmeasured factors that motivate H*, ε = (m,c)

$\varepsilon = \rho u + w$ ε and u share factors, m.

$H^* = \boldsymbol{\beta}'\mathbf{x} + \rho u + w$

Regression of H* on **x** omits u. λ is the prediction of u.

Note, the problem goes away if $\rho = 0$.

Estimation of the Selection Model

- Two step least squares
  - Inefficient
  - Simple – exists in current software
  - Simple to understand and widely used
- Full information maximum likelihood
  - Efficient
  - Simple – exists in current software
  - Not so simple to understand – widely misunderstood

Estimation

Heckman's two step procedure

- (1) Estimate the probit model and compute λ_i for each observation using the estimated parameters.
- (2) a. Linearly regress y_i on $\mathbf{x}_i$ and λ_i using the observed data

  b. Correct the estimated asymptotic covariance matrix for the use of the estimated λ_i. (An application of Murphy and Topel (1984) – Heckman was 1979) See text, pp. 953-955.

Variance of a Heckman's Two Step Estimator

The parameters in $\boldsymbol{\gamma}$ do have to be estimated using the probit equation. Rewrite (19-24) as

$$(y_i | z_i = 1, \mathbf{x}_i, \mathbf{w}_i) = \mathbf{x}_i'\boldsymbol{\beta} + \beta_\lambda \hat{\lambda}_i + v_i - \beta_\lambda(\hat{\lambda}_i - \lambda_i).$$

In this form, we see that in the preceding expression we have ignored both an additional source of variation in the compound disturbance and correlation across observations; the same estimate of $\boldsymbol{\gamma}$ is used to compute $\hat{\lambda}_i$ for every observation. Heckman has shown that the earlier covariance matrix can be appropriately corrected by adding a term inside the brackets,

$$\mathbf{Q} = \hat{\rho}^2(\mathbf{X}_*'\hat{\boldsymbol{\Delta}}\mathbf{W})\text{Est.Asy.Var}[\hat{\boldsymbol{\gamma}}](\mathbf{W}'\hat{\boldsymbol{\Delta}}\mathbf{X}_*) = \hat{\rho}^2\hat{\mathbf{F}}\hat{\mathbf{V}}\hat{\mathbf{F}}',$$

where $\hat{\mathbf{V}} = \text{Est.Asy.Var}[\hat{\boldsymbol{\gamma}}]$, the estimator of the asymptotic covariance of the probit coefficients. Any of the estimators in (17-22) to (17-24) may be used to compute $\hat{\mathbf{V}}$. The complete expression is

$$\text{Est.Asy.Var}[\mathbf{b}, b_\lambda] = \hat{\sigma}_\varepsilon^2[\mathbf{X}_*'\mathbf{X}_*]^{-1}[\mathbf{X}_*'(\mathbf{I} - \hat{\rho}^2\hat{\Delta})\mathbf{X}_* + \mathbf{Q}][\mathbf{X}_*'\mathbf{X}_*]^{-1}.$$

This is the estimator that is embedded in contemporary software such as *Stata*. We note three useful further aspects of the two-step estimator:

1. This is an application of the two-step procedures we developed in Section 8.4.1 and 14.7 and that were formalized by Murphy and Topel (1985).[40]

Application – Labor Supply

```
MROZ labor supply data. Cross section, 753 observations
Use LFP for binary choice, KIDS for count models.
LFP      = labor force participation, 0 if no, 1 if yes.
WHRS     = wife's hours worked. 0 if LFP=0
KL6      = number of kids less than 6
K618     = kids 6 to 18
WA       = wife's age
WE       = wife's education
WW       = wife's wage, 0 if LFP=0.
RPWG     = Wife's reported wage at the time of the interview
HHRS     = husband's hours
HA       = husband's age
HE       = husband's education
HW       = husband's wage
FAMINC   = family income
MTR      = marginal tax rate
WMED     = wife's mother's education
WFED     = wife's father's education
UN       = unemployment rate in  county of residence
CIT      = dummy for urban residence
AX       = actual years of wife's previous labor market experience
AGE      = Age
AGESQ    = Age squared
EARNINGS= WW * WHRS
LOGE     = Log of EARNINGS
KIDS     = 1 if kids < 18 in the home.
```

Labor Supply Model

```
NAMELIST  ; Z = One,KL6,K618,WA,WE,HA,HE $
NAMELIST  ; X = One,KL6,K618,Age,Agesq,WE,Faminc $
PROBIT     ; Lhs = LFP ; Rhs = Z ; Hold(IMR=Lambda) $
SELECT     ; Lhs = WHRS ; Rhs = X $
REGRESS  ; If [ LFP = 1] ; Lhs = WHRS ; Rhs = X $
REGRESS  ; If [ LFP = 1] ; Lhs = WHRS ; Rhs = X,Lambda $
REGRESS  ; If [ LFP = 1] ; Lhs = WHRS ; Rhs = X,Lambda
           ; Cluster = 1 $
```

Participation Equation

```
---------------------------------------------------------------------------
Binomial Probit Model
Dependent variable                    LFP
Log likelihood function       -461.37865
Restricted log likelihood     -514.87320
Chi squared [  6](P= .000)     106.98911
Significance level                .00000
McFadden Pseudo R-squared      .1038985
Estimation based on N =    753, K =   7
Inf.Cr.AIC  =    936.8 AIC/N =    1.244
Results retained for SELECTION model.
--------+------------------------------------------------------------------
        |                    Standard             Prob.      95% Confidence
    LFP |  Coefficient         Error       z     |z|>Z*         Interval
--------+------------------------------------------------------------------
        |Index function for probability.....................................
Constant|    1.00265**        .49994     2.01   .0449       .02277   1.98252
    KL6 |    -.90400***       .11434    -7.91   .0000    -1.12811   -.67989
   K618 |    -.05453          .04021    -1.36   .1751     -.13334    .02428
     WA |    -.02602*         .01333    -1.95   .0508     -.05214    .00009
     WE |     .16039***       .02774     5.78   .0000      .10603    .21475
     HA |    -.01643          .01329    -1.24   .2165     -.04248    .00962
     HE |    -.05191**        .02040    -2.54   .0110     -.09190   -.01192
--------+------------------------------------------------------------------
```

Hours Equation

```
---------------------------------------------------------------------
Sample Selection Model
Probit selection equation based on LFP
Selection rule is: Observations with LFP      =  1
Results of selection:
                  Data points     Sum of weights
Data set                753              753.0
Selected sample         428              428.0
---------------------------------------------------------------------

---------------------------------------------------------------------------
Sample Selection Model.............................
Two step      least squares regression ............
LHS=WHRS      Mean                   =     1302.92991
              Standard deviation     =      776.27438
              Number of observs.     =            428
Model size    Parameters             =              8
              Degrees of freedom     =            420
Residuals     Sum of squares         =    .226721E+09
              Standard error of e    =      734.71953
Fit           R-squared              =          .10210
              Adjusted R-squared     =          .08713
Model test    F[  7,   420] (prob)   =     6.8(.0000)
Not using OLS or no constant. Rsqrd & F may be < 0
Standard error corrected for selection    939.01825
Correlation of disturbance in regression
and Selection Criterion (Rho)     =        -.84541
--------+------------------------------------------------------------------
        |                 Standard            Prob.      95% Confidence
    WHRS|  Coefficient       Error       z    |z|>Z*          Interval
--------+------------------------------------------------------------------
Constant|    2442.27**     1202.111      2.03   .0422        86.17   4798.36
     KL6|     115.110      282.0086       .41   .6831     -437.617   667.836
    K618|   -101.721***     38.28339    -2.66   .0079     -176.755   -26.687
     AGE|     14.6359      53.19166       .28   .7832     -89.6178  118.8897
   AGESQ|     -.10079        .61856      -.16   .8706     -1.31315   1.11157
      WE|   -102.203***     39.40963    -2.59   .0095     -179.445   -24.962
  FAMINC|      .01379***      .00345     4.00   .0001       .00703    .02056
  LAMBDA|   -793.857       494.5410    -1.61   .1084    -1763.140   175.426
--------+------------------------------------------------------------------
```

Selection “Bias”

WHRS	Coefficient	Standard Error	t	Prob. |t|>T*	95% Confidence Interval	
Constant	1812.13	1144.333	1.58	.1140	-430.73	4054.98
KL6	-299.128***	100.0331	-2.99	.0030	-495.189	-103.067
K618	-126.400***	30.87285	-4.09	.0001	-186.909	-65.890
AGE	11.2795	53.84421	.21	.8342	-94.2532	116.8122
AGESQ	-.26104	.62633	-.42	.6771	-1.48862	.96655
WE	-47.3272***	17.29681	-2.74	.0065	-81.2283	-13.4260
FAMINC	.01262***	.00339	3.72	.0002	.00598	.01926

WHRS	Coefficient	Standard Error	t	Prob. |t|>T*	95% Confidence Interval	
Constant	2442.27**	1194.817	2.04	.0416	100.47	4784.06
KL6	115.110	252.7874	.46	.6491	-380.345	610.564
K618	-101.721***	33.75941	-3.01	.0027	-167.888	-35.554
AGE	14.6359	53.73825	.27	.7855	-90.6891	119.9610
AGESQ	-.10079	.63114	-.16	.8732	-1.33780	1.13623
WE	-102.203***	35.27561	-2.90	.0040	-171.342	-33.064
FAMINC	.01379***	.00344	4.01	.0001	.00704	.02054
LAMBDA	-793.857*	445.1168	-1.78	.0752	-1666.270	78.556

Heckman's corrected standard errors

WHRS	Coefficient	Standard Error	z	Prob. |z|>Z*	95% Confidence Interval	
Constant	2442.27**	1202.111	2.03	.0422	86.17	4798.36
KL6	115.110	282.0086	.41	.6831	-437.617	667.836
K618	-101.721***	38.28339	-2.66	.0079	-176.755	-26.687
AGE	14.6359	53.19166	.28	.7832	-89.6178	118.8897
AGESQ	-.10079	.61856	-.16	.8706	-1.31315	1.11157
WE	-102.203***	39.40963	-2.59	.0095	-179.445	-24.962
FAMINC	.01379***	.00345	4.00	.0001	.00703	.02056
LAMBDA	-793.857	494.5410	-1.61	.1084	-1763.140	175.426

Uncorrected standard errors - OLS

WHRS	Coefficient	Standard Error	t	Prob. |t|>T*	95% Confidence Interval	
Constant	2442.27**	1194.817	2.04	.0416	100.47	4784.06
KL6	115.110	252.7874	.46	.6491	-380.345	610.564
K618	-101.721***	33.75941	-3.01	.0027	-167.888	-35.554
AGE	14.6359	53.73825	.27	.7855	-90.6891	119.9610
AGESQ	-.10079	.63114	-.16	.8732	-1.33780	1.13623
WE	-102.203***	35.27561	-2.90	.0040	-171.342	-33.064
FAMINC	.01379***	.00344	4.01	.0001	.00704	.02054
LAMBDA	-793.857*	445.1168	-1.78	.0752	-1666.270	78.556

Heteroscedasticity robust standard errors (cluster = 1)

WHRS	Coefficient	Clustered Std.Error	t	Prob. |t|>T*	95% Confidence Interval	
Constant	2442.27**	1232.878	1.98	.0482	25.87	4858.66
KL6	115.110	302.2937	.38	.7036	-477.375	707.594
K618	-101.721***	33.77584	-3.01	.0028	-167.920	-35.521
AGE	14.6359	57.43322	.25	.7990	-97.9311	127.2030
AGESQ	-.10079	.67949	-.15	.8822	-1.43257	1.23100
WE	-102.203***	36.79001	-2.78	.0057	-174.310	-30.096
FAMINC	.01379***	.00409	3.37	.0008	.00578	.02181
LAMBDA	-793.857*	470.2006	-1.69	.0921	-1715.433	127.719

Maximum Likelihood Estimation

$$\log L = \sum_{d=1} \log\left[\frac{\exp\left(-\frac{1}{2}(\varepsilon_i/\sigma)^2\right)}{\sigma\sqrt{2\pi}}\Phi\left(\frac{\rho(\varepsilon_i/\sigma)+\boldsymbol{\alpha}'\mathbf{z}_i}{\sqrt{1-\rho^2}}\right)\right]$$

$$+ \sum_{d=0} \log\left[1-\Phi(\boldsymbol{\alpha}'\mathbf{z}_i)\right]$$

Reparameterize this: let $q_i = \boldsymbol{\alpha}'\mathbf{z}_i$

(1) $\theta = 1/\sigma$

(2) $\gamma = \beta/\sigma$ (Olsen transformation)

(3) $\tau = \rho/\sqrt{1-\rho^2}$

(4) Constrain ρ to be in (-1,1) by using

$$\psi = \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right) = \text{atanh}\rho, \text{so } \rho = \text{atanh}^{-1}(\psi) = \frac{\exp(2\psi)-1}{\exp(2\psi)+1}$$

$$\log L = \sum_{d=0}\log\Phi(-q_i) + \sum_{d=1}\begin{matrix}\log\theta - \frac{1}{2}\log 2\pi - \frac{1}{2}(\theta y_i - \boldsymbol{\gamma}'\mathbf{x}_i)^2 \\ +\log\Phi[\tau(\theta y_i - \boldsymbol{\gamma}'\mathbf{x}) + q_i\sqrt{1+\tau^2}]\end{matrix}$$

MLE

```
ML Estimates of Selection Model
Dependent variable                    WHRS
Log likelihood function       -3894.47094
Estimation based on N =     753, K =  16
Inf.Cr.AIC  =   7820.9 AIC/N =   10.386
```

WHRS	Coefficient	Standard Error	z	Prob. |z|>Z*	95% Confidence Interval	
	Selection (probit) equation for LFP.....					
Constant	1.01351**	.51518	1.97	.0492	.00376	2.02325
KL6	-.90130***	.11218	-8.03	.0000	-1.12116	-.68143
K618	-.05292	.03999	-1.32	.1857	-.13130	.02545
WA	-.02492*	.01382	-1.80	.0714	-.05201	.00217
WE	.16396***	.02783	5.89	.0000	.10942	.21850
HA	-.01763	.01379	-1.28	.2011	-.04467	.00940
HE	-.05597***	.02020	-2.77	.0056	-.09555	-.01638
	Corrected regression, Regime 1.....					
Constant	1946.85*	1167.225	1.67	.0953	-340.87	4234.56
KL6	-209.025	221.3726	-.94	.3451	-642.907	224.857
K618	-120.969***	35.45852	-3.41	.0006	-190.467	-51.472
AGE	12.0376	51.99025	.23	.8169	-89.8615	113.9366
AGESQ	-.22652	.59914	-.38	.7054	-1.40082	.94777
WE	-59.2166*	33.33120	-1.78	.0756	-124.5446	6.1113
FAMINC	.01289***	.00332	3.88	.0001	.00639	.01940
SIGMA(1)	748.132***	59.64630	12.54	.0000	631.227	865.036
RHO(1,2)	-.22965	.49962	-.46	.6458	-1.20890	.74959

Two Step Estimates

Coefficient	Standard Error
Index function for probabil	
1.00265**	.49994
-.90400***	.11434
-.05453	.04021
-.02602*	.01333
.16039***	.02774
-.01643	.01329
-.05191**	.02040
2442.27**	1202.111
115.110	282.0086
-101.721***	38.28339
14.6359	53.19166
-.10079	.61856
-102.203***	39.40963
.01379***	.00345
-793.857	494.5410

```
Standard error corrected for selection   939.01825
Correlation of disturbance in regression
and Selection Criterion (Rho)     =       -.84541
```

How to Handle Selectivity

- The ‘Mills Ratio’ approach – just add a ‘lambda’ to whatever model is being estimated?
  - The Heckman model applies to a probit model with a linear regression.
  - The conditional mean in a nonlinear model is not something “+lambda”
- The model can sometimes be built up from first principles

Received Sunday, April 27, 2014

I have a paper regarding strategic alliances between firms, and their impact on firm risk. While observing how a firm's strategic alliance formation impacts its risk, I need to correct for two types of selection biases. The reviews at Journal of Marketing asked us to correct for the propensity of firms to enter into alliances, and also the propensity to select a specific partner, before we examine how the partnership itself impacts risk.

Our approach involved conducting a probit of alliance formation propensity, take the inverse mills and include it in the second selection equation which is also a probit of partner selection. Then, we include inverse mills from the second selection into the main model. The review team states that this is not correct, and we need an MLE estimation in order to correctly model the set of three equations. The Associate Editor's point is given below. Can you please provide any guidance on whether this is a valid criticism of our approach. Is there a procedure in LIMDEP that can handle this set of three equations with two selection probit models?

AE's comment:
"Please note that the procedure of using an inverse mills ratio is only consistent when the main equation where the ratio is being used is linear. In non-linear cases (like the second probit used by the authors), this is not correct. Please see any standard econometric treatment like Greene or Wooldridge. A MLE estimator is needed which will be far from trivial to specify and estimate given error correlations between all three equations."

A Bivariate Probit Model

Labor Force Participation Equation

$d^* = \alpha'z + u$

$d = 1(d^* > 0)$

Full Time or Part Time?

$f^* = \beta'x + \varepsilon$

$f = 1(f^* > 0)$

Probability Model:

Nonparticipant: $\text{Prob}[d=0] = \Phi(-\alpha'z)$

Participant and Full Time

$\text{Prob}[f=1,d=1] = \text{Prob}[f=1|d=1]\text{Prob}[d=1]$

$= \text{Bivariate Normal}(\beta'x, \alpha'z, \rho)$

Participant and Part Time

$\text{Prob}[f=0,d=1] = \text{Prob}[f=0|d=1]\text{Prob}[d=1]$

$= \text{Bivariate Normal}(\beta'x, -\alpha'z, -\rho)$

FT/PT Selection Model

```
+---------------------------------------------+
| FIML Estimates of Bivariate Probit Model    |
| Dependent variable                FULLFP    |
| Weighting variable                  None    |
| Number of observations               753    |
| Log likelihood function        -723.9798    |
| Number of parameters                  16    |
| Selection model based on LFP                |
+---------------------------------------------+
```

Full Time = Hours > 1000

Variable	Coefficient	Standard Error	b/St.Er.	P[|Z|>z]	Mean of X
	Index equation for FULLTIME				
Constant	.94532822	1.61674948	.585	.5587	
WW	-.02764944	.01941006	-1.424	.1543	4.17768154
KL6	.04098432	.26250878	.156	.8759	.14018692
K618	-.13640024	.05930081	-2.300	.0214	1.35046729
AGE	.03543435	.07530788	.471	.6380	41.9719626
AGESQ	-.00043848	.00088406	-.496	.6199	1821.12150
WE	-.08622974	.02808185	-3.071	.0021	12.6588785
FAMINC	.210971D-04	.503746D-05	4.188	.0000	24130.4229
	Index equation for LFP				
Constant	.98337341	.50679582	1.940	.0523	
KL6	-.88485756	.11251971	-7.864	.0000	.23771580
K618	-.04101187	.04020437	-1.020	.3077	1.35325365
WA	-.02462108	.01308154	-1.882	.0598	42.5378486
WE	.16636047	.02738447	6.075	.0000	12.2868526
HA	-.01652335	.01287662	-1.283	.1994	45.1208499
HE	-.06276470	.01912877	-3.281	.0010	12.4913679
	Disturbance correlation				
RHO(1,2)	-.84102682	.25122229	-3.348	.0008	

Building a Likelihood for a Poisson Regression Model with Selection

Poisson Probability Functions

$$P(y_i \mid x_i) = \exp(-\lambda_i)\lambda_i^{y} / y_i!$$

Covariates and Unobserved Heterogeneity

$$\lambda(\mathbf{x}_i, \varepsilon_i) = \exp(\mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i)$$

Conditional Contribution to the Log Likelihood

$$\log L_i \mid \varepsilon_i = -\lambda(\mathbf{x}_i, \varepsilon_i) + y_i \log \lambda(\mathbf{x}_i, \varepsilon_i) - \log y_i!$$

Probit Selection Mechanism

$$d_i^* = \mathbf{z}_i'\boldsymbol{\gamma} + u_i, \; d_i = 1[d_i^* > 0]$$

$$[\varepsilon_i, u_i] \sim \text{BVN}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{pmatrix}\right]$$

$y_i, \mathbf{x}_i$ observed only when $d_i = 1$.

Building the Likelihood

The Conditional Probit Probability

$$u_i \mid \varepsilon_i \sim N[(\rho/\sigma)\varepsilon_i, (1-\rho^2)]$$

$$\text{Prob}[d_i = 1 \mid \mathbf{z_i}, \varepsilon_i] = \Phi\left[\frac{\mathbf{z}_i'\gamma + (\rho/\sigma)\varepsilon_i}{\sqrt{1-\rho^2}}\right]$$

$$\text{Prob}[d_i = 0 \mid \mathbf{z_i}, \varepsilon_i] = \Phi\left[\frac{\mathbf{-z}_i'\gamma - (\rho/\sigma)\varepsilon_i}{\sqrt{1-\rho^2}}\right]$$

Conditional Contribution to Likelihood

$$L_i(y_i, d_i = 1) \mid \varepsilon_i, = [f(y_i \mid \mathbf{x_i}, \varepsilon_i, d_i = 1)\text{Prob}[d_i = 1 \mid \mathbf{z_i}, \varepsilon_i]$$

$$L_i(d_i = 0) = \text{Prob}[d_i = 0 \mid \mathbf{z_i}, \varepsilon_i]$$

Dear Professor Greene,

I am very sorry to bother you considering this is my first time emailing you. I am **********, lecturer in Finance at &&&&&&&&&& University (Scotland).

I am doing a project investigating the impact of hedge fund manager's co-investment on the survival probability of the fund. As fund managers' co-investment decision is self-selection which might cause endogeneity issue, I jointly estimate the co-investment decision (Probit model) and the survival probability (Hazard model) to account for endogeneity of co-investment decision. I received one comment saying that I should use Heckman's two procedure to correct for endogeneity. My understanding is the Heckman's approach applies to a Probit and a LINEAR model. Since hazard model is nonlinear, simply adding inverse Mill's ration in the hazard model is wrong.

What I am asking is if my understanding of this is correct? If so, why can we not simply add Mill's ratio in a nonlinear model?

Conditional Likelihood

Conditional Density (not the log)

$$f(y_i, d_i = 1 \mid \varepsilon_i) = [f(y_i \mid \varepsilon_i, d_i = 1)]\text{Prob}[d_i = 1 \mid \varepsilon_i]$$

$$f(y_i, d_i = 0 \mid \varepsilon_i) = \text{Prob}[d_i = 0 \mid \varepsilon_i]$$

Unconditional Densities

$$f(y_i, d_i = 1) = \int_{-\infty}^{\infty} [f(y_i \mid \varepsilon_i, d_i = 1)]\text{Prob}[d_i = 1 \mid \varepsilon_i] \frac{1}{\sigma}\phi\left(\frac{\varepsilon}{\sigma}\right) d\varepsilon$$

$$f(y_i, d_i = 0) = \int_{-\infty}^{\infty} \text{Prob}[d_i = 0 \mid \varepsilon_i] \frac{1}{\sigma}\phi\left(\frac{\varepsilon}{\sigma}\right) d\varepsilon$$

Log Likelihoods

$$\log L_i = \log f(y_i, d_i)$$

Poisson Model with Selection

- Strategy:
  - Hermite quadrature or maximum simulated likelihood.
  - Not by throwing a ‘lambda’ into the unconditional likelihood
- Could this be done without joint normality?
  - How robust is the model?
  - Is there any other approach available?
  - Not easily. The subject of ongoing research

Nonnormality Issue

- How robust is the Heckman model to nonnormality of the unobserved effects?
- Are there other techniques
  - Parametric: Copula methods
  - Semiparametric: Klein/Spady and Series methods
- Other forms of the selection equation – e.g., multinomial logit
- Other forms of the primary model: e.g., as above.

Application: Health Care Usage

German Health Care Usage Data, 7,293 Individuals, Varying Numbers of Periods

This is an unbalanced panel with 7,293 individuals. There are altogether 27,326 observations. The number of observations ranges from 1 to 7.

(Frequencies are: 1=1525, 2=2158, 3=825, 4=926, 5=1051, 6=1000, 7=987).

(Downloaded from the JAE Archive)

Variables in the file are

DOCTOR	=	**1(Number of doctor visits > 0)**
HOSPITAL	=	**1(Number of hospital visits > 0)**
HSAT	=	**health satisfaction, coded 0 (low) - 10 (high)**
DOCVIS	=	**number of doctor visits in last three months**
HOSPVIS	=	**number of hospital visits in last calendar year**
PUBLIC	=	**insured in public health insurance = 1; otherwise = 0**
ADDON	=	**insured by add-on insurance = 1; otherswise = 0**
HHNINC	=	**household nominal monthly net income in German marks / 10000**. (4 observations with income=0 were dropped)
HHKIDS	=	**children under age 16 in the household = 1; otherwise = 0**
EDUC	=	**years of schooling**
AGE	=	**age in years**
MARRIED	=	**marital status**

```
SAMPLE   ; All $
NAMELIST ; z = one,age,educ,married,hhkids,hhninc $
NAMELIST ; x = one,age,hsat $
PROBIT   ; Lhs = public ; Rhs = z ; Hold $
SELECT   ; Lhs = docvis ; Rhs = x ; mle$
POISSON  ; Lhs = docvis ; Rhs = x ; Selection ; MLE $
BIVARIATE; Lhs = doctor,public ; Rh1=x ; Rh2=z ; Selection $
```

```
Binomial Probit Model
Dependent variable                PUBLIC
Log likelihood function     -8320.32399
Restricted log likelihood   -9711.25153
Chi squared [   5 d.f.]      2781.85508
Significance level               .00000
McFadden Pseudo R-squared      .1432285
Estimation based on N =  27326, K =   6
Inf.Cr.AIC  =  16652.6 AIC/N =     .609
Results retained for SELECTION model.
```

PUBLIC	Coefficient	Standard Error	z	Prob. |z|>Z*	95% Confidence Interval	
	Index function for probability					
Constant	3.63081***	.07341	49.46	.0000	3.48693	3.77469
AGE	.00115	.00111	1.03	.3011	-.00103	.00333
EDUC	-.17193***	.00406	-42.30	.0000	-.17990	-.16397
MARRIED	-.02762	.02903	-.95	.3413	-.08452	.02927
HHKIDS	-.06940***	.02503	-2.77	.0056	-.11845	-.02035
HHNINC	-.97958***	.05581	-17.55	.0000	-1.08895	-.87020

```
+---------------------------------------------------------+
| Sample Selection Model                                  |
| Probit selection equation based on PUBLIC               |
| Selection rule is: Observations with PUBLIC   =  1      |
| Results of selection:                                   |
|                  Data points     Sum of weights         |
| Data set            27326            27326.0            |
| Selected sample     24203            24203.0            |
+---------------------------------------------------------+
```

```
Sample Selection Model.....................
Two step     least squares regression .....
LHS=DOCVIS   Mean                   =        3.31207
             Standard deviation     =        5.88224
             Number of observs.     =          24203
Model size   Parameters             =              4
             Degrees of freedom     =          24199
Residuals    Sum of squares         =        719470.
             Standard error of e    =        5.45265
Fit          R-squared              =         .14069
             Adjusted R-squared     =         .14059
Model test   F[  3, 24199] (prob) =  1320.7(.0000)
Not using OLS or no constant. Rsqrd & F may be < 0
Standard error corrected for selection     5.47621
Correlation of disturbance in regression
and Selection Criterion (Rho)      =       -.18222
```

DOCVIS	Coefficient	Standard Error	z	Prob. |z|>Z*	95% Confidence Interval	
Constant	8.22501***	.20329	40.46	.0000	7.82657	8.62345
AGE	.02939***	.00319	9.22	.0000	.02314	.03563
HSAT	-.89510***	.01564	-57.23	.0000	-.92575	-.86445
LAMBDA	-.99788***	.23717	-4.21	.0000	-1.46273	-.53303

```
Sample Selection Model.....................
Two step     least squares regression .....
Standard error corrected for selection     5.47621
Correlation of disturbance in regression
and Selection Criterion (Rho)     =        -.18222
```

DOCVIS	Coefficient	Standard Error	z	Prob. |z|>Z*	95% Confidence Interval	
Constant	8.22501***	.20329	40.46	.0000	7.82657	8.62345
AGE	.02939***	.00319	9.22	.0000	.02314	.03563
HSAT	-.89510***	.01564	-57.23	.0000	-.92575	-.86445
LAMBDA	-.99788***	.23717	-4.21	.0000	-1.46273	-.53303

```
ML Estimates of Selection Model
Dependent variable                DOCVIS
Log likelihood function    -83716.13744
Estimation based on N =  27326, K =  11
Inf.Cr.AIC  = 167454.3 AIC/N =    6.128
Model estimated: Aug 01, 2012, 23:04:48
```

DOCVIS	Coefficient	Standard Error	z	Prob. |z|>Z*	95% Confidence Interval	
	Selection (probit) equation for PUBLIC					
Constant	3.63881***	.07361	49.44	.0000	3.49455	3.78308
AGE	.00117	.00115	1.02	.3073	-.00108	.00342
EDUC	-.17218***	.00412	-41.83	.0000	-.18024	-.16411
MARRIED	-.02901	.02925	-.99	.3212	-.08633	.02831
HHKIDS	-.07355***	.02512	-2.93	.0034	-.12279	-.02431
HHNINC	-.98832***	.05164	-19.14	.0000	-1.08953	-.88712
	Corrected regression, Regime 1					
Constant	8.10465***	.19392	41.79	.0000	7.72457	8.48473
AGE	.03040***	.00325	9.35	.0000	.02403	.03678
HSAT	-.89792***	.01389	-64.65	.0000	-.92514	-.87070
SIGMA(1)	5.45984***	.01019	535.78	.0000	5.43987	5.47982
RHO(1,2)	-.08970*	.04810	-1.86	.0622	-.18398	.00458

```
Poisson  Model with Sample Selection.
Dependent variable                DOCVIS
Log likelihood function     -60829.53023
Restricted log likelihood  -205953.60785
Chi squared [   2 d.f.]     290248.15524
Significance level              .00000
McFadden Pseudo R-squared       .7046445
Estimation based on N =  27326, K =  11
Inf.Cr.AIC  = 121681.1 AIC/N =    4.453
Restr. Log-L is Poisson+Probit (indep).
LogL for initial probit =    -8320.3674
LogL for initial Poisson=  -197633.2404
Means for Psn/Neg.Bin. use selected data.
Means for Probit based on all observations.
```

DOCVIS	Coefficient	Standard Error	z	Prob. |z|>Z*	95% Confidence Interval	
	Parameters of Poisson/Neg. Binomial Probability					
Constant	1.55949***	.04592	33.96	.0000	1.46948	1.64949
AGE	.01053***	.00078	13.50	.0000	.00900	.01206
HSAT	-.25228***	.00339	-74.50	.0000	-.25892	-.24564
	Parameters of Probit Selection Model					
Constant	3.55923***	.07381	48.22	.0000	3.41456	3.70390
AGE	.00175	.00115	1.52	.1285	-.00051	.00401
EDUC	-.16992***	.00412	-41.21	.0000	-.17800	-.16184
MARRIED	-.02874	.02937	-.98	.3277	-.08631	.02882
HHKIDS	-.06369**	.02518	-2.53	.0114	-.11305	-.01434
HHNINC	-.93895***	.05262	-17.84	.0000	-1.04208	-.83582
	Standard Deviation of Heterogeneity					
Sigma	1.16922***	.00768	152.19	.0000	1.15416	1.18427
	Correlation of Heterogeneity & Selection					
Rho	-.02853	.05534	-.52	.6062	-.13698	.07993

```
FIML Estimates of Bivariate Probit Model
Dependent variable                DOCPUB
Log likelihood function    -22945.59406
Estimation based on N =  27326, K =  10
Inf.Cr.AIC  =  45911.2 AIC/N =    1.680
Selection model based on PUBLIC
Selected obs. 24203, Nonselected:  3123
```

DOCTOR / PUBLIC	Coefficient	Standard Error	z	Prob. |z|>Z*	95% Confidence Interval	
	Index equation for DOCTOR					
Constant	1.18664***	.05011	23.68	.0000	1.08843	1.28485
AGE	.00892***	.00079	11.29	.0000	.00737	.01047
HSAT	-.17226***	.00416	-41.43	.0000	-.18041	-.16411
	Index equation for PUBLIC					
Constant	3.63944***	.07354	49.49	.0000	3.49530	3.78357
AGE	.00098	.00115	.85	.3949	-.00127	.00323
EDUC	-.17210***	.00411	-41.86	.0000	-.18016	-.16405
MARRIED	-.02446	.02927	-.84	.4033	-.08182	.03290
HHKIDS	-.07540***	.02511	-3.00	.0027	-.12462	-.02619
HHNINC	-.97675***	.05162	-18.92	.0000	-1.07792	-.87558
	Disturbance correlation					
RHO(1,2)	-.13237**	.05753	-2.30	.0214	-.24513	-.01960