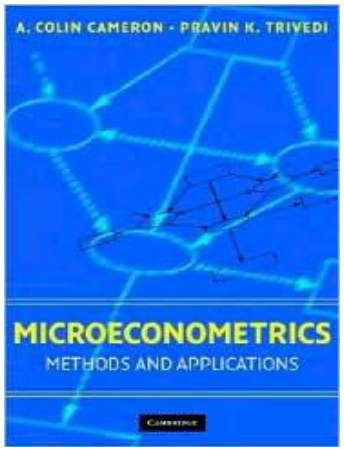
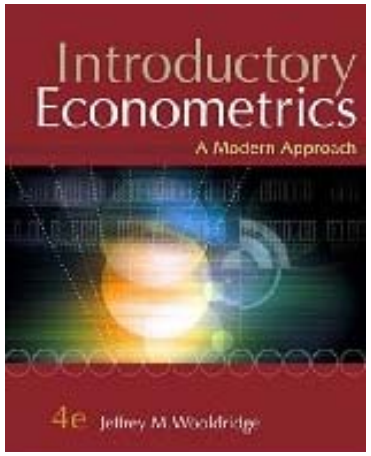
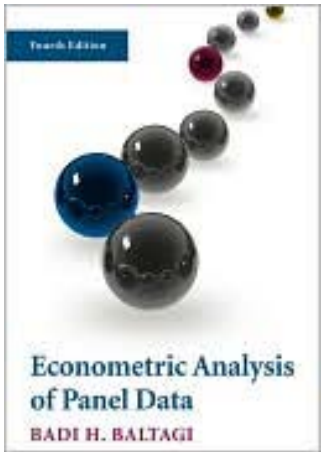
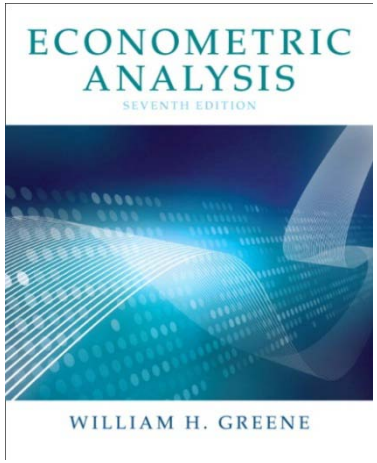
Topics in Microeconometrics

William Greene
Department of Economics
Stern School of Business

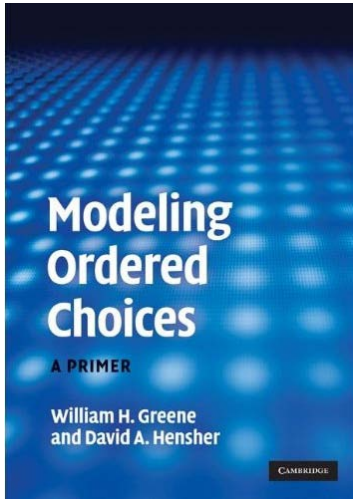


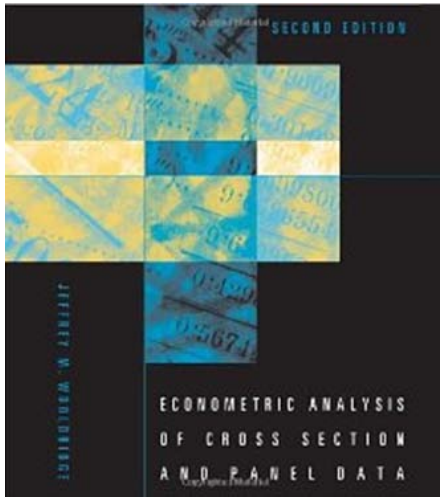


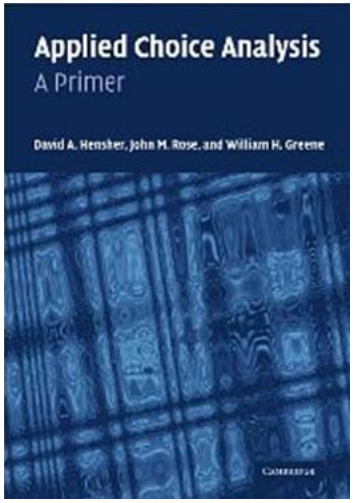




Part 7: Sample Selection in Nonlinear and Panel Models







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. These may not be the same as the full population
- Can one generalize from the subpopulation to the full population?

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."

Heckman's Canonical Model

A behavioral model:

Offered wage = o* = $\beta_1'\mathbf{x}$ + v (x = age,experience,educ...)

Reservation wage = r* = $\beta_2'\mathbf{z}$ + u (z = age, kids, family stuff)

Labor force participation:

LFP = 1 if o* $\geq$ r*, 0 otherwise

$$\text{Prob(LFP=1)} = \Phi\left[(\beta_1'\mathbf{x} - \quad)/\sqrt{\sigma_v^2 + \sigma_u^2} \right]$$

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

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* | **w**,LFP = 1]? Not $\gamma'\mathbf{w}$.

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."**

Sample Selection Observations

- The selection ‘problem’ is caused by the correlation of the unobservables
  - Selection on observables is often manageable within the conventional model.
  - Selection on unobservables often requires a more detailed specification of the model – where does the effect come from?
- The ‘bias’ relates to the inconsistency of familiar estimators such as least squares
- The data are not biased; the (an) estimator is biased.

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$[(0,0),(1,\rho\sigma,\sigma^2)]$

$$E[y_i \mid y_i \text{ is observed}] = E[y_i \mid d_i = 1]$$

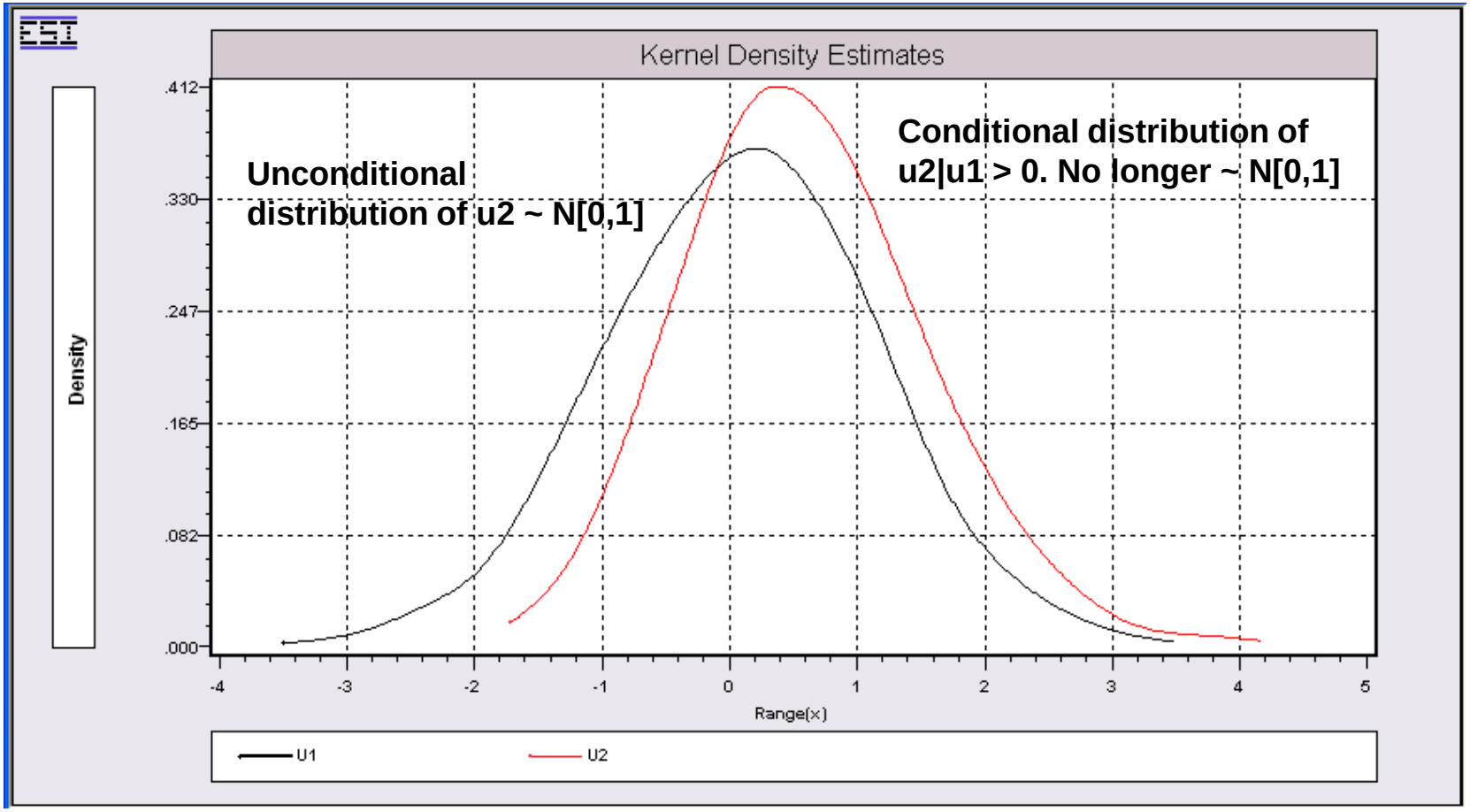
$$= \beta' \mathbf{x}_i + E[\varepsilon_i \mid d_i = 1]$$

$$= \beta' \mathbf{x}_i + E[\varepsilon_i \mid u_i > -\alpha' z_i]$$

$$= \beta' \mathbf{x}_i + (\rho\sigma) \frac{\phi(\alpha' \mathbf{z}_i)}{\Phi(\alpha' \mathbf{z}_i)}$$

$$= \beta' \mathbf{x}_i + \theta\lambda_i$$

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 $\theta = 0 \leftrightarrow \rho = 0$
- The "selection bias" is plim $(\mathbf{b} - \boldsymbol{\beta})$

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. Not more or less robust
  - Simple to do – 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 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).

Mroz 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  ; Lhs = WHRS ; Rhs = X,Lambda $
REJECT     ; LFP = 0 $
REGRESS  ; Lhs = WHRS ; Rhs = X $
```

Participation Equation

```
+-----------------------------------------+
| Binomial Probit Model                   |
| Dependent variable               LFP    |
| Weighting variable              None    |
| Number of observations           753    |
+-----------------------------------------+
```

Variable	Coefficient	Standard Error	b/St.Er.	P[|Z|>z]	Mean of X
	Index function for probability				
Constant	1.00264501	.49994379	2.006	.0449	
KL6	-.90399802	.11434394	-7.906	.0000	.23771580
K618	-.05452607	.04021041	-1.356	.1751	1.35325365
WA	-.02602427	.01332588	-1.953	.0508	42.5378486
WE	.16038929	.02773622	5.783	.0000	12.2868526
HA	-.01642514	.01329110	-1.236	.2165	45.1208499
HE	-.05191039	.02040378	-2.544	.0110	12.4913679

Hours Equation

```
+----------------------------------------------------+
| Sample Selection Model                             |
| Two stage   least squares regression               |
| LHS=WHRS     Mean                 =   1302.930     |
|              Standard deviation   =   776.2744     |
| WTS=none     Number of observs.   =        428     |
| Model size   Parameters           =          8     |
|              Degrees of freedom   =        420     |
| Residuals    Sum of squares       =   .2267214E+09 |
|              Standard error of e  =   734.7195     |
| Correlation of disturbance in regression           |
| and Selection Criterion (Rho).....   -.84541       |
+----------------------------------------------------+
```

Variable	Coefficient	Standard Error	b/St.Er.	P[|Z|>z]	Mean of X
Constant	2442.26665	1202.11143	2.032	.0422	
KL6	115.109657	282.008565	.408	.6831	.14018692
K618	-101.720762	38.2833942	-2.657	.0079	1.35046729
AGE	14.6359451	53.1916591	.275	.7832	41.9719626
AGESQ	-.10078602	.61856252	-.163	.8706	1821.12150
WE	-102.203059	39.4096323	-2.593	.0095	12.6588785
FAMINC	.01379467	.00345041	3.998	.0001	24130.4229
LAMBDA	-793.857053	494.541008	-1.605	.1084	.61466207

Selection "Bias" of OLS

Variable	Coefficient	Standard Error	b/St.Er.	P[|Z|>z]	Mean of X
Constant	2442.26665	1202.11143	2.032	.0422	
KL6	115.109657	282.008565	.408	.6831	.14018692
K618	-101.720762	38.2833942	-2.657	.0079	1.35046729
AGE	14.6359451	53.1916591	.275	.7832	41.9719626
AGESQ	-.10078602	.61856252	-.163	.8706	1821.12150
WE	-102.203059	39.4096323	-2.593	.0095	12.6588785
FAMINC	.01379467	.00345041	3.998	.0001	24130.4229
LAMBDA	-793.857053	494.541008	-1.605	.1084	.61466207

Variable	Coefficient	Standard Error	t-ratio	P[|T|>t]	Mean of X
Constant	1812.12538	1144.33342	1.584	.1140	
KL6	-299.128041	100.033124	-2.990	.0030	.14018692
K618	-126.399697	30.8728451	-4.094	.0001	1.35046729
AGE	11.2795338	53.8442084	.209	.8342	41.9719626
AGESQ	-.26103541	.62632815	-.417	.6771	1821.12150
WE	-47.3271780	17.2968137	-2.736	.0065	12.6588785
FAMINC	.01261889	.00338906	3.723	.0002	24130.4229

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

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

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

MLE

```
+--------------------------------------------+
| ML Estimates of Selection Model            |
| Maximum Likelihood Estimates               |
| Number of observations          753        |
| Iterations completed             47        |
| Log likelihood function      -3894.471     |
| Number of parameters             16        |
| FIRST  7 estimates are probit equation.    |
+--------------------------------------------+
```

Variable	Coefficient	Standard Error	b/St.Er.	P[|Z|>z]
	Selection (probit) equation for LFP			
Constant	1.01350651	.54823177	1.849	.0645
KL6	-.90129694	.11081111	-8.134	.0000
K618	-.05292375	.04137216	-1.279	.2008
WA	-.02491779	.01428642	-1.744	.0811
WE	.16396194	.02911763	5.631	.0000
HA	-.01763340	.01431873	-1.231	.2181
HE	-.05596671	.02133647	-2.623	.0087
	Corrected regression, Regime 1			
Constant	1946.84517	1167.56008	1.667	.0954
KL6	-209.024866	222.027462	-.941	.3465
K618	-120.969192	35.4425577	-3.413	.0006
AGE	12.0375636	51.9850307	.232	.8169
AGESQ	-.22652298	.59912775	-.378	.7054
WE	-59.2166488	33.3802882	-1.774	.0761
FAMINC	.01289491	.00332219	3.881	.0001
SIGMA(1)	748.131644	59.7508375	12.521	.0000
RHO(1,2)	-.22965163	.50082203	-.459	.6466

MLE vs. Two Step

```
Two Step
 Constant      2442.26665     1202.11143     2.032    .0422
 KL6            115.109657     282.008565      .408    .6831     .14018692
 K618          -101.720762     38.2833942    -2.657    .0079    1.35046729
 AGE            14.6359451     53.1916591      .275    .7832    41.9719626
 AGESQ          -.10078602      .61856252     -.163    .8706    1821.12150
 WE            -102.203059     39.4096323    -2.593    .0095    12.6588785
 FAMINC          .01379467      .00345041     3.998    .0001    24130.4229
 LAMBDA        -793.857053     494.541008    -1.605    .1084     .61466207
|              Standard error of e  =   734.7195              |
| Correlation of disturbance in regression                    |
| and Selection Criterion (Rho).....   -.84541 |
MLE
 Constant      1946.84517     1167.56008     1.667    .0954
 KL6           -209.024866     222.027462     -.941    .3465
 K618          -120.969192     35.4425577    -3.413    .0006
 AGE            12.0375636     51.9850307      .232    .8169
 AGESQ          -.22652298      .59912775     -.378    .7054
 WE            -59.2166488     33.3802882    -1.774    .0761
 FAMINC          .01289491      .00332219     3.881    .0001
 SIGMA(1)       748.131644     59.7508375    12.521    .0000
 RHO(1,2)       -.22965163      .50082203     -.459    .6466
```

Extension – Treatment Effect

What is the value of an elite college education?

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

$y_i^* = \mathbf{x}\boldsymbol{\beta} + \delta d_i + \varepsilon_i$ observed for everyone

$[\varepsilon_i, u_i] \sim$ Bivariate Normal$[0,0,\sigma^2,\rho,1]$

$$E[y_i^* | x_i, d_i=1] = \mathbf{x}\boldsymbol{\beta} + \delta + E[\varepsilon_i \mid x_i, d_i = 1]$$

$$= \mathbf{x}\boldsymbol{\beta} + \delta + E[\varepsilon_i \mid x_i, u_i > -\mathbf{z}_i'\boldsymbol{\gamma}]$$

$$= \mathbf{x}\boldsymbol{\beta} + \delta + \rho\sigma\left(\frac{\phi(\mathbf{z}_i'\boldsymbol{\gamma})}{\Phi(\mathbf{z}_i'\boldsymbol{\gamma})}\right)$$

$$= \mathbf{x}\boldsymbol{\beta} + \delta + \rho\sigma\lambda_i$$

$$E[y_i^* | x_i, d_i=0] = \mathbf{x}\boldsymbol{\beta} + E[\varepsilon_i \mid x_i, d_i = 0]$$

$$= \mathbf{x}\boldsymbol{\beta} + \rho\sigma\left(\frac{-\phi(-\mathbf{z}_i'\boldsymbol{\gamma})}{\Phi(-\mathbf{z}_i'\boldsymbol{\gamma})}\right)$$

Least squares is still biased and inconsistent. Left out variable

Treatment Effect

$$E[y_i^* | x_i, d_i = 1] - E[y_i^* | x_i, d_i = 0]$$

$$= \quad \mathbf{x}\boldsymbol{\beta} + \delta + \rho\sigma\left(\frac{\phi(\mathbf{z}_i'\boldsymbol{\gamma})}{\Phi(\mathbf{z}_i'\boldsymbol{\gamma})}\right)$$

$$- \mathbf{x}\boldsymbol{\beta} - \rho\sigma\left(\frac{-\phi(-\mathbf{z}_i'\boldsymbol{\gamma})}{\Phi(-\mathbf{z}_i'\boldsymbol{\gamma})}\right)$$

$$= \delta + \rho\sigma\left[\left(\frac{\phi(\mathbf{z}_i'\boldsymbol{\gamma})}{\Phi(\mathbf{z}_i'\boldsymbol{\gamma})}\right) - \left(\frac{-\phi(-\mathbf{z}_i'\boldsymbol{\gamma})}{\Phi(-\mathbf{z}_i'\boldsymbol{\gamma})}\right)\right]$$

$$= \text{Treatment} + \text{Selection Effect}$$

Sample Selection in Exponential Regression

An approach modeled on Heckman's model

Regression Equation: $\text{Prob}[y=j|x,u]=P(\lambda)$;

$$\lambda=\exp(\mathbf{x}\boldsymbol{\beta}+\theta u)$$

Selection Equation: $d=1[\mathbf{z}\boldsymbol{\delta}+\theta u>0]$ (The usual probit)

$[u,\varepsilon]\sim n[0,0,1,1,\rho]$ (Var[u] is absorbed in θ)

Estimation: Nonlinear Least Squares: [Terza (1998).]

$$E[y|x,d=1]=\exp(\mathbf{x}\boldsymbol{\beta}+\rho)\,\frac{\Phi(\mathbf{z}\boldsymbol{\delta}+\rho)}{\Phi(\mathbf{z}\boldsymbol{\delta})}$$

Panel Data and Selection

Selection equation with time invariant individual effect

$d_{it} = 1[\mathbf{z}_{it}'\boldsymbol{\gamma} + \theta_i + \eta_{it} > 0]$

Observation mechanism: $(y_{it}, \mathbf{x}_{it})$ observed when $d_{it} = 1$

Primary equation of interest

Common effects linear regression model

$y_{it} \mid (d_{it} = 1) = \mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i + \varepsilon_{it}$

"Selectivity" as usual arises as a problem when the unobservables are correlated; $\text{Corr}(\varepsilon_{it}, \eta_{it}) \neq 0$.

The common effects, θ_i and α_i make matters worse.

Panel Data and Sample Selection Models: A Nonlinear Time Series

I. 1990-1992: Fixed and Random Effects Extensions

II. 1995 and 2005: Model Identification through Conditional Mean Assumptions

III. 1997-2005: Semiparametric Approaches based on Differences and Kernel Weights

IV. 2007: Return to Conventional Estimators, with Bias Corrections

Panel Data Sample Selection Models

Verbeek, Economics Letters, 1990.

$d_{it} = 1[\mathbf{z}_{it}'\boldsymbol{\gamma} + w_i + \eta_{it} > 0]$ **(Random effects probit)**

$y_{it} \mid (d_{it} = 1) = \mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i + \varepsilon_{it}$; **(Fixed effects regression)**

Proposed "marginal likelihood" based on joint normality

$$\log L_i = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\prod_{t=1}^{T_i}\Phi\left[(2d_{it}-1)\frac{\mathbf{z}_{it}'\boldsymbol{\gamma} + \Delta_{it} + u_{i,1} + d_{it}u_{i,2}}{\sqrt{\sigma_\eta^2(1-d_{it}\rho^2)}}\right] f(u_{i,1},u_{i,2})\,du_{i,1}du_{i,2}$$

$$\Delta_{it} = (\rho/\sigma_\varepsilon)d_{it}\left[(y_{it}-\bar{y}_i) - (\mathbf{x}_{it}-\bar{\mathbf{x}}_i)'\boldsymbol{\beta}\right]$$

(Integrate out the random effects; difference out the fixed effects.)

$u_{i,1}, u_{i,2}$ are time invariant uncorrelated standard normal variables

How to do the integration? Natural candidate for simulation.

(Not mentioned in the paper. Too early.)

[Verbeek and Nijman: Selectivity "test" based on this model, International Economic Review, 1992.]

Zabel – Economics Letters

- Inappropriate to have a mix of FE and RE models
- Two part solution
  - Treat both effects as “fixed”
  - Project both effects onto the group means of the variables in the equations (Mundlak approach)
  - Resulting model is two random effects equations
- Use both random effects

Selection with Fixed Effects

$$y_{it}^* = \eta_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it},\ \eta_i = \bar{\mathbf{x}}_i'\boldsymbol{\pi} + \tau w_i, w_i \sim N[0,1]$$

$$d_{it}^* = \theta_i + \mathbf{z}_{it}'\boldsymbol{\alpha} + u_{it},\ \theta_i = \bar{\mathbf{z}}_i'\boldsymbol{\delta} + \omega v_i, v_i \sim N[0,1]$$

$$(\varepsilon_{it}, u_{it}) \sim N_2[(0,0),(\sigma^2, 1, \rho\sigma)].$$

$$L_i = \int_{-\infty}^{\infty} \prod_{d_{it}=0} \Phi\left[-\mathbf{z}_{it}'\boldsymbol{\alpha} - \bar{\mathbf{z}}_i'\boldsymbol{\delta} - \omega v_i\right] \phi(v_i) dv_i$$

$$\times \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \prod_{d_{it}=1} \left(\Phi\left[\frac{\mathbf{z}_{it}'\boldsymbol{\alpha} + \bar{\mathbf{z}}_i'\boldsymbol{\delta} + \omega v_i + (\rho/\sigma)\varepsilon_{it}}{\sqrt{1-\rho^2}} \right] \right) \frac{1}{\sigma}\phi\left(\frac{\varepsilon_{it}}{\sigma}\right) \phi_2(v_i, w_i)\, dv_i dw_i$$

$$\varepsilon_{it} = y_{it} - \mathbf{x}_{it}'\boldsymbol{\beta} - \bar{\mathbf{x}}_i'\boldsymbol{\pi} - \tau w_i$$

Practical Complications

The bivariate normal integration is actually the product of two univariate normals, because in the specification above, v_i and w_i are assumed to be uncorrelated. Vella notes, however, "… given the computational demands of estimating by maximum likelihood induced by the requirement to evaluate multiple integrals, we consider the applicability of available simple, or two step procedures."

Simulation

The first line in the log likelihood is of the form $E_v[\Pi_{d=0}\Phi(\ldots)]$ and the second line is of the form $E_w[E_v[\Phi(\ldots)\phi(\ldots)/\sigma]]$. Using simulation instead, the simulated likelihood is

$$L_i^S = \frac{1}{R}\sum_{r=1}^{R}\prod_{d_{it}=0}\Phi\left[-\mathbf{z}_{it}'\boldsymbol{\alpha}-\bar{\mathbf{z}}_i'\boldsymbol{\delta}-\omega v_{i,r}\right]$$

$$\times\ \frac{1}{R}\sum_{r=1}^{R}\prod_{d_{it}=1}\Phi\left[\frac{\mathbf{z}_{it}'\boldsymbol{\alpha}+\bar{\mathbf{z}}_i'\boldsymbol{\delta}+\omega v_{i,r}+(\rho/\sigma)\varepsilon_{it,r}}{\sqrt{1-\rho^2}}\right]\frac{1}{\sigma}\phi\left(\frac{\varepsilon_{it,r}}{\sigma}\right)$$

$$\varepsilon_{it,r} = y_{it}-\mathbf{x}_{it}'\boldsymbol{\beta}-\bar{\mathbf{x}}_i'\boldsymbol{\pi}-\tau w_{i,r}$$

Correlated Effects

Suppose that w_i and v_i are bivariate standard normal with correlation ρ_{vw}. We can project w_i on v_i and write

$$w_i = \rho_{vw} v_i + (1-\rho_{vw}^2)^{1/2} h_i$$

where h_i has a standard normal distribution. To allow the correlation, we now simply substitute this expression for wi in the simulated (or original) log likelihood, and add ρ_{vw} to the list of parameters to be estimated. The simulation is then over still independent normal variates, v_i and h_i.

Conditional Means

Wooldridge (1995) proposes an estimator that can be based on straightforward applications of conventional, everyday methods.

$$y_{it}^* = \eta_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it},$$
$$d_{it}^* = \theta_i + \mathbf{z}_{it}'\boldsymbol{\alpha} + u_{it},$$
$$(\varepsilon_{it}, u_{it}) \sim N_2[(0,0),(\sigma^2,1,\rho\sigma)].$$

Under the mean independence assumption

$$E[\varepsilon_{it} \mid \eta_i, \theta_i, \mathbf{z}_{i1},\ldots,\mathbf{z}_{iT}, v_{i1},\ldots,v_{iT},d_{i1},\ldots,d_{iT}] = \rho u_{it}, \quad v_{it} = \theta_i + u_{it}$$

$$E[y_{it}|\mathbf{x}_{i1}.,,,.\mathbf{x}_{iT}, \eta_i, \theta_i, \mathbf{z}_{i1},\ldots,\mathbf{z}_{iT}, v_{i1},\ldots,v_{iT},d_{i1},\ldots,d_{iT}] = \eta_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \rho u_{it}.$$

This suggests an approach to estimating the model parameters, however it requires computation of u_{it}. That would require estimation of θ_i which cannot be done, at least not consistently – and that precludes simple estimation of u_{it}.

A Feasible Estimator

To escape the dilemma, Wooldridge suggests Chamberlain's approach to the fixed effects model,

$$\theta_i = f_0 + \mathbf{z}_{i1}'\mathbf{f}_1 + \mathbf{z}_{i2}'\mathbf{f}_2 + \ldots + \mathbf{z}_{iT}'\mathbf{f}_T + h_i.$$

With this substitution,

$$\begin{aligned} d_{it}^* &= \mathbf{z}_{it}'\boldsymbol{\alpha} + f_0 + \mathbf{z}_{i1}'\mathbf{f}_1 + \mathbf{z}_{i2}'\mathbf{f}_2 + \ldots + \mathbf{z}_{iT}'\mathbf{f}_T + h_i + u_{it} \\ &= \mathbf{z}_{it}'\boldsymbol{\alpha} + f_0 + \mathbf{z}_{i1}'\mathbf{f}_1 + \mathbf{z}_{i2}'\mathbf{f}_2 + \ldots + \mathbf{z}_{iT}'\mathbf{f}_T + w_{it} \end{aligned}$$

where w_{it} is independent of $\mathbf{z}_{it}$, $t = 1,\ldots,T$. This now implies that

$$\begin{aligned} E[y_{it}|\mathbf{x}_{i1},,,,,\mathbf{x}_{iT}, \eta_i, \theta_i, \mathbf{z}_{i1},\ldots,\mathbf{z}_{iT}, v_{i1},\ldots,v_{iT},d_{i1},\ldots,d_{iT}] &= \eta_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \rho(w_{it} - h_i) \\ &= (\eta_i - \rho h_i) + \mathbf{x}_{it}'\boldsymbol{\beta} + \rho w_{it}. \end{aligned}$$

Estimation

To complete the estimation procedure, we now compute T cross sectional probit models (reestimating f_0, f_1,… each time), and compute $\hat{\lambda}_{it}$ from each one. The resulting equation,

$$y_{it} = a_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \rho\,\hat{\lambda}_{it} + v_{it}$$

now forms the basis for estimation of $\boldsymbol{\beta}$ and ρ by using a conventional fixed effects linear regression with the observed data.

Kyriazidou - Semiparametrics

Assume 2 periods

Estimate selection equation by FE logit

Use first differences and weighted least squares:

$$\hat{\boldsymbol{\beta}} = \left[\Sigma_{i=1}^{N} d_{i1} d_{i2} \Psi_i \Delta \mathbf{x}_i \Delta \mathbf{x}_i'\right]^{-1} \left[\Sigma_{i=1}^{N} d_{i1} d_{i2} \hat{\Psi}_i \Delta \mathbf{x}_i \Delta y_i\right]$$

$$\hat{\Psi}_i = \frac{1}{h} K\left[\frac{\Delta w_i' \hat{\boldsymbol{\alpha}}}{h}\right]$$ kernel function.

Use with longer panels - any pairwise differences

Extensions based on pairwise differences by Rochina-Barrachina and Dustman/Rochina-Barrachina (1999)

Bias Corrections

- Val and Vella, 2007 (Working paper)
- Assume fixed effects
  - Bias corrected probit estimator at the first step
  - Use fixed probit model to set up second step Heckman style regression treatment.

Postscript

- What selection process is at work?
  - All of the work examined here (and in the literature) assumes the selection operates anew in each period
  - An alternative scenario: Selection into the panel, once, at baseline.
  - Alternative: Sequential selection = endogenous attrition (Wooldridge 2002, inverse probability weighting)
- Why aren't the time invariant components correlated?
- Other models
  - All of the work on panel data selection assumes the main equation is a linear model.
  - Any others? Discrete choice? Counts?

Attrition

- In a panel, t=1,...,T individual I leaves the sample at time K_i and does not return.
- If the determinants of attrition (especially the unobservables) are correlated with the variables in the equation of interest, then the now familiar problem of sample selection arises.

Dealing with Attrition in a QOL Study

- The attrition issue: Appearance for the second interview was low for people with initial low QOL (death or depression) or with initial high QOL (don't need the treatment). Thus, missing data at exit were clearly related to values of the dependent variable.
- Solutions to the attrition problem
  - Heckman selection model (used in the study)
    - Prob[Present at exit|covariates] = $\Phi(z'\theta)$ (Probit model)
    - Additional variable added to difference model $\lambda_i = \Phi(\mathbf{z}_i'\theta)/\Phi(\mathbf{z}_i'\theta)$
  - The FDA solution: fill with zeros. (!)

An Early Attrition Model

Hausman, J. and Wise, D., "Attrition Bias in Experimental and Panel Data: The Gary Income Maintenance Experiment," Econometrica, 1979.

A two period model:

Structural response model (Random Effects Regression)

$$y_{i1} = \mathbf{x}_{i1}\boldsymbol{\beta} + \varepsilon_{i1} + u_i$$

$$y_{i2} = \mathbf{x}_{i2}\boldsymbol{\beta} + \varepsilon_{i2} + u_i$$

Attrition model for observation in the second period (Probit)

$$z_{i2}^* = \delta y_{i2} + \mathbf{x}_{i2}\boldsymbol{\theta} + \mathbf{w}_{i2}\boldsymbol{\alpha} + v_{i2}$$

$$z_{i2} = 1(z_{i2}^* > 0)$$

Endogeneity "problem"

$$\rho_{12} = \text{Corr}[\varepsilon_{i1} + u_i, \varepsilon_{i2} + u_i] = \sigma_u^2 / (\sigma_\varepsilon^2 + \sigma_u^2)$$

$$\tau = \text{Corr}[v_{i2}, \varepsilon_{i2} + u_i] = \text{Corr}[v_{i2} + \delta(\varepsilon_{i2} + u_i), \varepsilon_{i2} + u_i)$$

Methods of Estimating the Attrition Model

- Heckman style “selection” model
- Two step maximum likelihood
- Full information maximum likelihood
- Two step method of moments estimators
- Weighting schemes that account for the “survivor bias”

Selection Model

Reduced form probit model for second period observation equation

$$z_{i2}^* = x_{i2}'(\theta + \delta\beta) + w_i'\alpha + \delta(\varepsilon_{i2} + u_i + v_i)$$
$$= r_{i2}'\gamma + h_{i2}$$
$$z_{i2} = 1(z_{i2}^* > 0)$$

Conditional means for observations observed in the second period

$$E[y_{i2} \mid x_{i2}, z_{i2} = 1] = x_{i2}'\beta + (\rho_{12}\sigma_\varepsilon)\frac{\phi(r_{i2}'\gamma)}{\Phi(r_{i2}'\gamma)}$$

First period conditional means for observations observed in the second period

$$E[y_{i1} \mid x_{i1}, z_{i2} = 1] = x_{i1}'\beta + (\rho_{12}\tau\sigma_\varepsilon)\frac{\phi(r_{i2}'\gamma)}{\Phi(r_{i2}'\gamma)}$$

(1) Estimate probit equation

(2) Combine these two equations with a period dummy variable, use OLS with a constructed regressor in the second period

THE TWO DISTURBANCES ARE CORRELATED.

TREAT THIS IS A SUR MODEL. (EQUIVALENT TO MDE)

Maximum Likelihood

$$\text{LogL}_i = \frac{-\log 2\pi}{2} - \log\sigma_\varepsilon - \frac{(y_{i1} - x_{i1}'\beta)^2}{2\sigma_\varepsilon^2}$$

$$+ \quad z_{i2}\left[\log\sigma_\varepsilon + \log\sqrt{1-\rho_{12}^2} - \frac{[(y_{i2} - \rho_{12}y_{i1}) - (x_{12} - \rho_{12}x_{i1})'\beta)]^2}{2\sigma_\varepsilon^2(1-\rho_{12}^2)} + \log\Phi\left(\frac{r_{i2}'\gamma + (\tau/\sigma_\varepsilon)(y_{i2} - x_{i2}'\beta)}{\sqrt{1-\tau^2}}\right)\right]$$

$$+ (1 - z_{i2})\left[\log\Phi\left(-\frac{r_{i2}'\gamma + (\rho_{12}\tau/\sigma_\varepsilon)(y_{i1} - x_{i1}'\beta)}{\sqrt{1-\rho_{12}\tau^2}}\right)\right]$$

(1) See H&W for FIML estimation

(2) Use the invariance principle to reparameterize

(3) Estimate γ separately and use a two step ML with Murphy and Topel correction of asymptotic covariance matrix.

TABLE IV[a]

PARAMETER ESTIMATES OF THE EARNINGS FUNCTION STRUCTURAL MODEL WITH AND WITHOUT A CORRECTION FOR ATTRITION

	With attrition correction: maximum likelihood estimates (standard errors)		Without attrition correction: generalized least squares estimates (standard errors)
Variables	Earnings function parameters	Attrition parameters	Earnings function parameters
Constant	5.8539	−.6347	5.8911
	(0.0903)	(.3351)	(.0829)
Experimental effect	−.0822	.2414	−.0793
	(.0402)	(.1211)	(.0390)
Time effect	.0940	—	.0841
	(.0520)	—	(.0358)
Education	.0209	−.0204	.0136
	(.0052)	(.0244)	(.0050)
Experience	.0037	−.0038	.0020
	(.0013)	(.0061)	(.0013)
Income	−.0131	.1752	−.0115
	(.0050)	(.0470)	(.0044)
Union	.2159	1.4290	.2853
	(.0362)	(0.1252)	(.0330)
Poor health	−.0601	.2480	−.0578
	(.0330)	(.1237)	(.0326)
	$\hat{\sigma}_\eta^2 = .1832$	$l^* = 64.35$	$\hat{\sigma}_\eta^2 = .1236$
	(.0057)		
	$\hat{\rho}_{12} = .2596$	$\rho_{23} = -.1089$	$\hat{\rho}_{12} = .2003$
	(.0391)	(1.0429)	

A Model of Attrition

- Nijman and Verbeek, Journal of Applied Econometrics, 1992
- Consumption survey (Holland, 1984 – 1986)
  - Exogenous selection for participation (rotating panel)
  - Voluntary participation (missing not at random – attrition)

Attrition Model

The main equation

$y_{i,t} = \beta_0 + \mathbf{x}'_{i,t}\boldsymbol{\beta} + \tilde{\alpha}_i + \varepsilon_{i,t}$, Random effects consumption function

$\tilde{\alpha}_i = \bar{\mathbf{x}}'_i\boldsymbol{\theta} + u_i$, Mundlak device; u_i uncorrelated with $\mathbf{X}_i$

$y_{i,t} = \beta_0 + \mathbf{x}'_{i,t}\boldsymbol{\beta} + \bar{\mathbf{x}}'_i\boldsymbol{\theta} + u_i + \varepsilon_{i,t}$, Reduced form random effects model

The selection mechanism

$a_{it} = 1$[individual i asked to participate in period t] Purely exogenous

a_{it} may depend on observables, but does not depend on unobservables

$r_{it} = 1$[individual i chooses to participate if asked] Endogenous.

r_{it} is the endogenous participation dummy variable

$a_{it} = 0 \Rightarrow r_{it} = 0$

$a_{it} = 1 \Rightarrow$ the selection mechanism operates

Selection Equation

The main equation

$y_{i,t} = \beta_0 + \mathbf{x}'_{i,t}\boldsymbol{\beta} + \bar{\mathbf{x}}'_i\boldsymbol{\theta} + u_i + \varepsilon_{i,t}$, Reduced form random effects model

The selection mechanism

$r_{it} = 1[\text{individual i chooses to participate if asked}]$ Endogenous.

r_{it} is the endogenous participation dummy variable

$a_{it} = 0 \Rightarrow r_{it} = 0$

$a_{it} = 1 \Rightarrow$ the selection mechanism operates

$r_{it} = 1[\gamma_0 + \mathbf{x}'_{i,t}\boldsymbol{\gamma} + \bar{\mathbf{x}}'_i\boldsymbol{\mu} + \mathbf{z}'_{i,t}\boldsymbol{\delta} + v_i + w_{i,t} > 0]$ all observed if $a_{it} = 1$

State dependence: $\mathbf{z}$ may include $r_{i,t-1}$

Latent persistent unobserved heterogeneity: $\sigma_v^2 > 0$.

"Selection" arises if $\text{Cov}[\varepsilon_{i,t}, w_{i,t}] \neq 0$ or $\text{Cov}[u_i, v_i] \neq 0$

Estimation Using One Wave

- Use any single wave as a cross section with observed lagged values.
- Advantage: Familiar sample selection model
- Disadvantages
  - Loss of efficiency
  - “One can no longer distinguish between state dependence and unobserved heterogeneity.”

One Wave Model

A standard sample selection model.

$$y_{it} = \beta_0 + \mathbf{x}_{it}'\boldsymbol{\beta} + \bar{\mathbf{x}}_i'\boldsymbol{\theta} + (u_i + \varepsilon_{it})$$

$$r_{it} = 1[\gamma_0 + \mathbf{x}_{it}'\boldsymbol{\gamma} + \bar{\mathbf{x}}_i'\boldsymbol{\mu} + \delta_1 r_{i,t-1} + \delta_2 a_{i,t-1} + (v_i + w_{it}) > 0]$$

With only one period of data and $r_{i,t-1}$ exogenous,

this is the Heckman sample selection model.

If $>$ 0, then $r_{i,t-1}$ is correlated with v_i and the Heckman approach fails.

An assumption is required:

(1) Include $r_{i,t-1}$ and assume no unobserved heterogeneity

(2) Exclude $r_{i,t-1}$ and assume there is no state dependence.

In either case, now if $\text{Cov}[(u_i + \varepsilon_{it}),(v_i + w_{it})]$ we can use OLS.

Otherwise, use the maximum likelihood estimator.

Maximum Likelihood Estimation

- Because numerical integration is required in one or two dimensions for every individual in the sample at each iteration of a high dimensional numerical optimization problem, this is, though feasible, not computationally attractive.
  - The dimensionality of the optimization is irrelevant
  - This is much easier in 2008 than it was in 1992 (especially with simulation) The authors did the computations with Hermite quadrature.

Testing for Selection?

- Selectivity is parameterized in these models – coefficients are correlations or covariances.
- Maximum Likelihood Results
  - Covariances were highly insignificant.
  - LR statistic=0.46.
- Two step results produced the same conclusion based on a Hausman test
- ML Estimation results looked like the two step results.

Selectivity in Nonlinear Models

- 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

A Bivariate Probit Model

Labor Force Participation Equation

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

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

Full Time or Part Time?

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

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

Probability Model:

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

Participant and Full Time

$$\text{Prob}[f=1,d=1] = \text{Prob}[f=1|d=1]\text{Prob}[d=1]$$
$$= \text{Bivariate Normal}(\beta'\mathbf{x}, \alpha'\mathbf{z}, \rho)$$

Participant and Part Time

$$\text{Prob}[f=0,d=1] = \text{Prob}[f=0|d=1]\text{Prob}[d=1]$$
$$= \text{Bivariate Normal}(\beta'\mathbf{x}, -\alpha'\mathbf{z}, -\rho)$$

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

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

Covariates and Unobserved Heterogeneity

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

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

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

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

$y_i, \mathbf{x}_i$ 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

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

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

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

```
Poisson  Model with Sample Selection.
Dependent variable                DOCVIS
Log likelihood function     -3592.42064
Restricted log likelihood   -6076.83457
Chi squared [   2 d.f.]      4968.82786
Significance level               .00000
McFadden Pseudo R-squared       .4088336
Estimation based on N =  27326, K =  12
Inf.Cr.AIC  =   7208.8 AIC/N =      .264
Mean of LHS Variable =           3.12451
Restr. Log-L is Poisson+Probit (indep).
LogL for initial probit =    -2442.4091
LogL for initial Poisson=    -3634.4254
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	1.22286	1.03286	1.18	.2364	-.80151	3.24722
AGE	.01528**	.00654	2.34	.0194	.00247	.02809
EDUC	-.06984**	.02745	-2.54	.0110	-.12364	-.01603
HHNINC	-.38472	.36241	-1.06	.2884	-1.09504	.32559
MARRIED	-.22330	.15519	-1.44	.1502	-.52748	.08087
HHKIDS	-.18695	.14693	-1.27	.2032	-.47493	.10103
	Parameters of Probit Selection Model					
Constant	-3.40612***	.12282	-27.73	.0000	-3.64684	-3.16540
AGE	.00634***	.00178	3.57	.0004	.00286	.00982
EDUC	.06403***	.00757	8.45	.0000	.04919	.07887
HHNINC	.75095***	.09452	7.95	.0000	.56570	.93620
	Standard Deviation of Heterogeneity					
Sigma	1.07226***	.05985	17.92	.0000	.95495	1.18957
	Correlation of Heterogeneity & Selection					
Rho	-.02597	.25188	-.10	.9179	-.51964	.46770

Note: ***, **, * ==> Significance at 1%, 5%, 10% level.

```
Poisson Regression
Dependent variable                DOCVIS
Log likelihood function     -1659.65094
Restricted log likelihood   -1751.09275
Chi squared [   6 d.f.]       182.88363
Significance level              .00000
McFadden Pseudo R-squared      .0522199
Estimation based on N =  27326, K =   7
Inf.Cr.AIC  =   3333.3 AIC/N =     .122
Chi-squared =  2887.14468  RsqP= .1766
G  -squared =  2219.26013  RsqD= .0761
Overdispersion tests: g=mu(i)  :  5.121
Overdispersion tests: g=mu(i)^2:  5.728
Cov. matrix corrected for 2 step estimation
```

DOCVIS	Coefficient	Standard Error	z	Prob. |z|>Z*	95% Confidence Interval	
Constant	1.13394	4082.032	.00	.9998	-7999.50166	8001.76954
AGE	.01440	6.37321	.00	.9982	-12.47686	12.50566
EDUC	-.05828	63.44852	.00	.9993	-124.41510	124.29854
HHNINC	-.19541	742.8324	.00	.9998	-1456.12012	1455.72930
MARRIED	-.37640	.62847	-.60	.5492	-1.60818	.85539
HHKIDS	-.19118	.27342	-.70	.4844	-.72707	.34472
MlsRatio	.21226	1130.177	.00	.9999	-2214.89378	2215.31831

Note: ***, **, * ==> Significance at 1%, 5%, 10% level.

Stochastic Frontier Model: ML

$$y_i = \beta' \mathbf{x}_i + v_i - u_i$$

where $u_i = |\sigma_u U_i| = \sigma_u |U_i|$, $U_i \sim N[0,1^2]$,

$v_i = \sigma_v V_i$, $V_i \sim N[0,1^2]$.

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

where $\varepsilon_i = y_i - \beta' \mathbf{x}_i = v_i - u_i$,

$\lambda = \sigma_u/\sigma_v$,

$\sigma = \sqrt{\sigma_v^2 + \sigma_u^2}$

Sample Selected SF Model

$d_i = 1[\alpha'\mathbf{z}_i + w_i > 0]$, $w_i \sim N[0,1^2]$

$y_i = \beta'\mathbf{x}_i + \varepsilon_i$, $\varepsilon_i \sim N[0,\sigma_\varepsilon^2]$

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

$\varepsilon_i = v_i - u_i$

$u_i = |\sigma_u U_i| = \sigma_u |U_i|$ where $U_i \sim N[0,1^2]$

$v_i = \sigma_v V_i$ where $V_i \sim N[0,1^2]$.

$(w_i,v_i) \sim N_2[(0,1), (1, \rho\sigma_v, \sigma_v^2)]$

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

Sample Selection in a Nonlinear Model

$$d_i = 1(\boldsymbol{\alpha}'\mathbf{z}_i + w_i > 0) \quad w_i \sim N[0,1],$$

$$g_i|\varepsilon_i = g(\boldsymbol{\beta}'\mathbf{x}_i, \sigma_\varepsilon \varepsilon_i) \quad \varepsilon_i \sim N[0,1]$$

$$y_i\,|\mathbf{x}_i, \varepsilon_i \sim f\,[y_i \mid g(\boldsymbol{\beta}'\mathbf{x}_i, \sigma_\varepsilon \varepsilon_i)]$$

$$[w_i, \varepsilon_i] \sim \mathrm{N}[(0,1),(1,\rho,1)]$$

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

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

Simulated Log Likelihood for a Simpler Model

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

A 2 Step MSL Approach

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

where $a_i = \hat{\alpha}'\mathbf{z}_i$

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

Simulated ML for the SF Model

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

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

$$p(\mid U_i \mid) = \frac{2\exp[-\frac{1}{2} \mid U_i \mid^2]}{\sqrt{2\pi}}, \mid U_i \mid \geq 0.$$

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

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

This is simply a linear regression with a random constant term, $\alpha_i = \alpha - \sigma_u \mid U_i \mid$

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.

A Study of Health Status in the Presence of Attrition

Journal of
APPLIED ECONOMETRICS

Research Article

The dynamics of health in the British Household Panel Survey

Paul Contoyannis[1], Andrew M. Jones[2,*], Nigel Rice[3]

Article first published online: 9 AUG 2004

DOI: 10.1002/jae.755

Issue

Journal of Applied Econometrics

Volume 19, Issue 4, pages 473–503, July/August 2004

Model for Self Assessed Health

- British Household Panel Survey (BHPS)
  - Waves 1-8, 1991-1998
  - Self assessed health on 0,1,2,3,4 scale
  - Sociological and demographic covariates
  - Dynamics – inertia in reporting of top scale
- Dynamic ordered probit model
  - Balanced panel – analyze dynamics
  - Unbalanced panel – examine attrition

Dynamic Ordered Probit Model

Latent Regression - Random Utility

$$h_{it}^{*} = \boldsymbol{\beta}'\mathbf{x}_{it} + \boldsymbol{\gamma}'\mathbf{H}_{i,t-1} + \alpha_i + \varepsilon_{it}$$

$\mathbf{x}_{it}$ = relevant covariates and control variables

It would not be appropriate to include $h_{i,t-1}$ itself in the model as this is a label, not a measure

$\mathbf{H}_{i,t-1}$ = 0/1 indicators of reported health status in previous period

$H_{i,t-1}(j) = 1[\text{Individual i reported } h_{it} = j \text{ in previous period}]$, j=0,...,4

Ordered Choice Observation Mechanism

$$h_{it} = j \text{ if } \mu_{j-1} < h_{it}^{*} \leq \mu_j, \; j = 0,1,2,3,4$$

Ordered Probit Model - $\varepsilon_{it} \sim N[0,1]$

Random Effects with Mundlak Correction and Initial Conditions

$$\alpha_i = \alpha_0 + \boldsymbol{\alpha}_1'\mathbf{H}_{i,1} + \boldsymbol{\alpha}_2'\bar{\mathbf{x}}_i + u_i, \quad u_i \sim N[0,\sigma^2]$$

Random Effects Dynamic Ordered Probit Model

Random Effects Dynamic Ordered Probit Model

$$h_{it}^{*} = \mathbf{x}_{it}'\boldsymbol{\beta} + \Sigma_{j=1}^{J}\gamma_j h_{i,t-1}(j) + \alpha_i + \varepsilon_{i,t}$$

$$h_{i,t} = j \text{ if } \mu_{j-1} < h_{it}^{*} < \mu_j$$

$$h_{i,t}(j) = 1 \text{ if } h_{i,t} = j$$

$$P_{it,j} = P[h_{it} = j] = \Phi(\mu_j - \mathbf{x}_{it}'\boldsymbol{\beta} - \Sigma_{j=1}^{J}\gamma_j h_{i,t-1}(j) - \alpha_i) - \Phi(\mu_{j-1} - \mathbf{x}_{it}'\boldsymbol{\beta} - \Sigma_{j=1}^{J}\gamma_j h_{i,t-1}(j) - \alpha_i)$$

Parameterize Random Effects

$$\alpha_i = \alpha_0 + \Sigma_{j=1}^{J}\alpha_{1,j} h_{i,1}(j) + \boldsymbol{\alpha}'\overline{\mathbf{x}}_i + u_i$$

Simulation or Quadrature Based Estimation

$$\ln L = \sum_{i=1}^{N} \ln \int_{\alpha_i} \prod_{t=1}^{T_i} P_{it,j} f(\alpha_j) d\alpha_j$$

Data

Table I. Variable definitions

SAH	Self-Assessed Health: 5 if excellent, 4 if good, 3 if fair, 2 if poor, 1 if very poor
WIDOW	1 if widowed, 0 otherwise
SINGLE	1 if never married, 0 otherwise
DIV/SEP	1 if divorced or separated, 0 otherwise
NON-WHITE	1 if a member of ethnic group other than white, 0 otherwise
DEGREE	1 if highest academic qualification is a degree or higher degree, 0 otherwise
HND/A	1 if highest academic qualification is HND or A level, 0 otherwise
O/CSE	1 if highest academic qualification is O level or CSE, 0 otherwise
HHSIZE	Number of people in household including respondent
NCH04	Number of children in household aged 0–4
NCH511	Number of children in household aged 5–11
NCH1218	Number of children in household aged 12–18
INCOME	Equivalized annual real household income in pounds
AGE	Age in years at 1st December of current wave

Variable of Interest

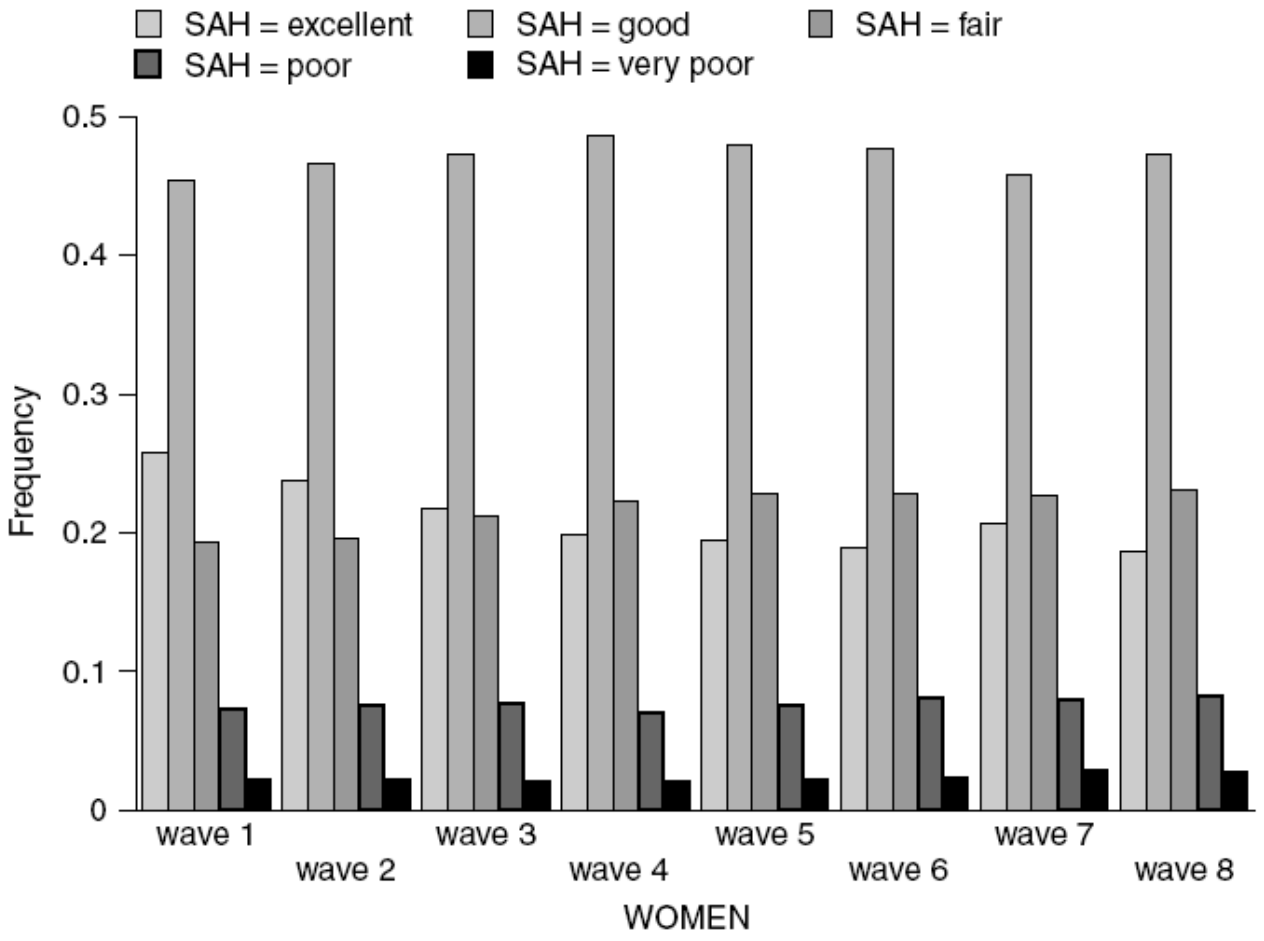


Figure 1. Self-assessed health status by wave

Dynamics

Table II. Transition matrices, balanced panel

(a) Men

SAH	EX	GOOD	FAIR	POOR	VERY POOR	*N*
EX	0.600	0.342	0.046	0.010	0.002	5485
GOOD	0.184	0.651	0.142	0.019	0.004	9263
FAIR	0.055	0.361	0.471	0.100	0.012	3433
POOR	0.029	0.120	0.340	0.418	0.093	1031
VERY POOR	0.032	0.073	0.133	0.423	0.339	248
N	5231	9287	3565	1111	266	19 460

(b) Women

SAH	EX	GOOD	FAIR	POOR	VERY POOR	*N*
EX	0.572	0.353	0.059	0.013	0.004	5164
GOOD	0.150	0.657	0.162	0.026	0.005	11 306
FAIR	0.040	0.362	0.465	0.116	0.017	4928
POOR	0.021	0.156	0.360	0.365	0.098	1587
VERY POOR	0.014	0.106	0.192	0.326	0.362	423
N	4884	11 329	5082	1649	464	23 408

Attrition

Table V. Sample size, drop-outs and attrition rates by wave

(a) All data

	FULL SAMPLE				EX at $t-1$	GOOD at $t-1$	FAIR at $t-1$	POOR at $t-1$	VPOOR at $t-1$
Wave	No. individuals	Survival rate	Drop-outs	Attrition rate	Attrition rate	Attrition rate	Attrition rate	Attrition rate	Attrition rate
1	10 256								
2	8957	87.33%	1299	12.67%	11.54%	12.57%	13.01%	13.73%	23.74%
3	8162	79.58%	795	8.88%	8.08%	8.13%	9.65%	12.62%	19.46%
4	7825	76.30%	337	4.13%	6.67%	6.54%	6.73%	10.35%	14.74%
5	7430	72.45%	395	5.05%	6.21%	6.18%	7.87%	9.11%	16.34%
6	7238	70.57%	192	2.58%	3.11%	3.24%	5.06%	10.47%	18.83%
7	7102	69.25%	136	1.88%	3.15%	3.85%	4.79%	8.83%	8.75%
8	6839	66.68%	263	3.70%	3.43%	3.82%	5.30%	5.88%	17.01%

Testing for Attrition Bias

Table 9: Verbeek and Nijman tests for attrition: based on dynamic ordered probit models with Wooldridge specification of correlated effects and initial conditions

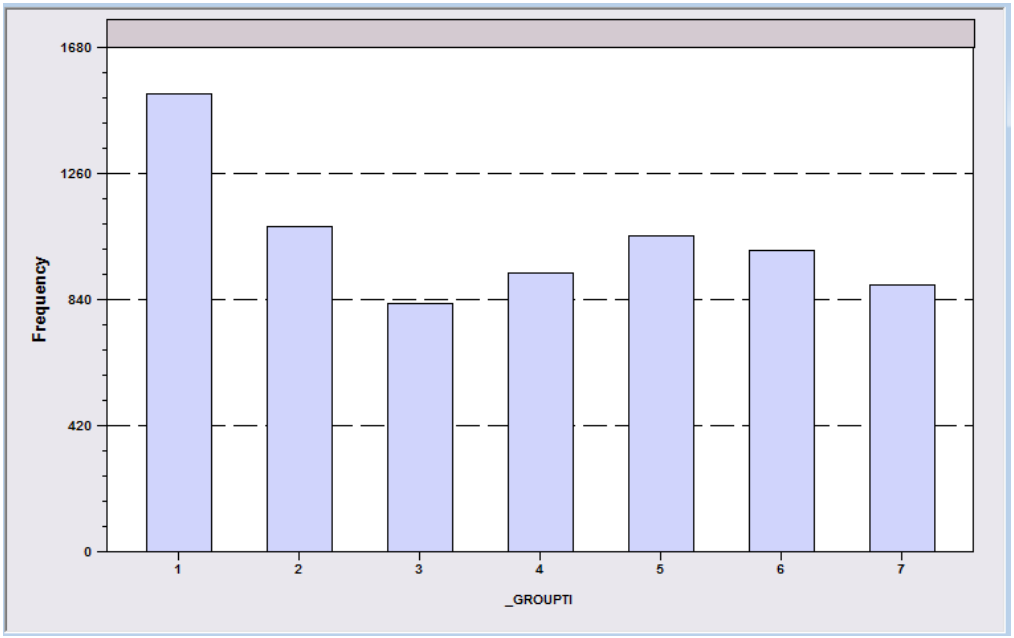
	MEN				WOMEN			
	β	Std.err.	t-test	p-value	β	Std.err.	t-test	p-value
NEXT WAVE	.199	.035	5.67	.000	.060	.034	1.77	.077
ALL WAVES	.139	.031	4.46	.000	.071	.029	2.45	.014
NUMBER OF WAVES	.031	.009	3.54	.000	.016	.008	1.88	.060

Three dummy variables added to full model with unbalanced panel suggest presence of attrition effects.

Probability Weighting Estimators

- A Patch for Attrition
- (1) Fit a participation probit equation for each wave.
- (2) Compute p(i,t) = predictions of participation for each individual in each period.
  - Special assumptions needed to make this work
- Ignore common effects and fit a weighted pooled log likelihood: $\Sigma_i \Sigma_t [d_{it}/p(i,t)]\log LP_{it}$.

Attrition Model with IP Weights



Assumes (1) Prob(attrition|all data) = Prob(attrition|selected variables) (ignorability)
(2) Attrition is an 'absorbing state.' No reentry.
Obviously not true for the GSOEP data above.
Can deal with point (2) by isolating a subsample of those present at wave 1 and the monotonically shrinking subsample as the waves progress.

Inverse Probability Weighting

Panel is based on those present at WAVE 1, N1 individuals

Attrition is an absorbing state. No reentry, so N1 $\geq$ N2 $\geq$ ... $\geq$ N8.

Sample is restricted at each wave to individuals who were present at the previous wave.

d_{it} = 1[Individual is present at wave t].

$d_{i1} = 1 \ \forall$ i, $d_{it} = 0 \Rightarrow d_{i,t+1} = 0$.

$\tilde{\mathbf{x}}_{i1}$ = covariates observed for all i at entry that relate to likelihood of being present at subsequent waves.

(health problems, disability, psychological well being, self employment, unemployment, maternity leave, student, caring for family member, ...)

Probit model for $d_{it} = 1[\delta'\tilde{\mathbf{x}}_{i1} + w_{it}]$, t = 2,...,8. $\hat{\pi}_{it}$ = fitted probability.

Assuming attrition decisions are independent, $\hat{P}_{it} = \prod_{s=1}^{t} \hat{\pi}_{is}$

Inverse probability weight $\hat{W}_{it} = \dfrac{d_{it}}{\hat{P}_{it}}$

Weighted log likelihood $\log L_W = \sum_{i=1}^{N} \sum_{t=1}^{8} \log L_{it}$ (No common effects.)

Estimated Partial Effects by Model

Table 12: Average partial effects on probability of reporting excellent health for selected variables

a) Men

	(1) Pooled model, balanced sample	(2) Pooled model, unbalanced sample	(3) Pooled model, IPW-1	(4) Pooled model, IPW-2	(5) Random effects, balanced sample	(6) Random effects, unbalanced sample
Ln(INCOME)	.009 (.004)	.009 (.004)	.009 (.004)	.011 (.005)	.013 (.006)	.012 (.005)
Mean Ln(INCOME)	.049 (.024)	.043 (.022)	.042 (.021)	.045 (.022)	.066 (.028)	.056 (.025)
DEGREE	.010 (.005)	.017 (.009)	.018 (.009)	.018 (.009)	.015 (.006)	.027 (.012)
HND/A	.019 (.009)	.021 (.011)	.021 (.010)	.022 (.011)	.028 (.011)	.030 (.013)
O/CSE	.016 (.008)	.020 (.010)	.020 (.010)	.020 (.010)	.024 (.010)	.028 (.012)
SAHEX(t-1)	.234 (.087)	.231 (.090)	.231 (.090)	.230 (.089)	.082 (.031)	.085 (.035)
SAHFAIR(t-1)	-.170 (.085)	-.163 (.084)	-.162 (.084)	-.162 (.083)	-.080 (.034)	-.077 (.036)
SAHPOOR(t-1)	-.242 (.167)	-.233 (.163)	-.232 (.162)	-.232 (.162)	-.151 (.077)	-.145 (.078)
SAHVPOOR(t-1)	-.260 (.198)	-.253 (.197)	-.255 (.199)	-.255 (.200)	-.184 (.104)	-.179 (.106)

Partial Effect for a Category

Table 12: Average partial effects on probability of reporting excellent health for selected variables

a) Men

	(1) Pooled model, balanced sample	(2) Pooled model, unbalanced sample	(3) Pooled model, IPW-1	(4) Pooled model, IPW-2	(5) Random effects, balanced sample	(6) Random effects, unbalanced sample
Ln(INCOME)	.009 (.004)	.009 (.004)	.009 (.004)	.011 (.005)	.013 (.006)	.012 (.005)
Mean Ln(INCOME)	.049 (.024)	.043 (.022)	.042 (.021)	.045 (.022)	.066 (.028)	.054 (.025)
DEGREE	.010 (.005)	.017 (.009)	.018 (.009)	.018 (.009)	.015 (.006)	.027 (.012)
HND/A	.019 (.009)	.021 (.011)	.021 (.010)	.023 (.011)	.028 (.011)	.030 (.013)
O/CSE	.016 (.008)	.020 (.010)	.020 (.010)	.020 (.010)	.024 (.010)	.028 (.012)
SAHEX(t-1)	.234 (.087)	.231 (.090)	.231 (.090)	.230 (.089)	.082 (.031)	.085 (.035)
SAHFAIR(t-1)	-.170 (.085)	-.163 (.084)	-.162 (.084)	-.162 (.083)	-.080 (.034)	-.077 (.036)
SAHPOOR(t-1)	-.242 (.167)	-.233 (.163)	-.232 (.162)	-.232 (.162)	-.151 (.077)	-.145 (.078)
SAHVPOOR(t-1)	-.260 (.198)	-.253 (.197)	-.255 (.199)	-.255 (.200)	-.184 (.104)	-.179 (.106)

SAHEX(t-1)	.234 (.087)
SAHFAIR(t-1)	-.170 (.085)
SAHPOOR(t-1)	-.242 (.167)
SAHVPOOR(t-1)	-.260 (.198)

These are 4 dummy variables for state in the previous period. Using first differences, the 0.234 estimated for SAHEX means transition from EXCELLENT in the previous period to GOOD in the previous period, where GOOD is the omitted category. Likewise for the other 3 previous state variables. The margin from 'POOR' to 'GOOD' was not interesting in the paper. The better margin would have been from EXCELLENT to POOR, which would have (EX,POOR) change from (1,0) to (0,1).