Nonlinear Panel Data Models

Prf. José Fajardo
Fundação Getulio Vargas

What is a *Nonlinear* Model?

- Model: $E[g(y)|\mathbf{x}] = m(\mathbf{x},\boldsymbol{\theta})$
- Objective:
  - Learn about $\boldsymbol{\theta}$ from $\mathbf{y}$, $\mathbf{X}$
  - Usually "estimate" $\boldsymbol{\theta}$
- Linear Model: Closed form; $\hat{\boldsymbol{\theta}} = h(\mathbf{y}, \mathbf{X})$
- Nonlinear Model
  - Not wrt $m(\mathbf{x},\boldsymbol{\theta})$. E.g., $y=\exp(\boldsymbol{\theta}'\mathbf{x} + \boldsymbol{\varepsilon})$
  - Wrt estimator: Implicitly defined. $h(\mathbf{y}, \mathbf{X}, \hat{\boldsymbol{\theta}})=0$, E.g., $E[y|\mathbf{x}]= \exp(\boldsymbol{\theta}'\mathbf{x})$

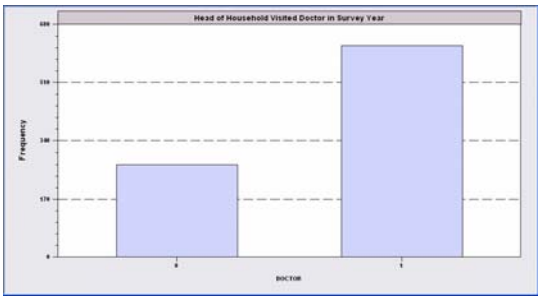
Binary Choice Models

- Binary choice modeling – the leading example of formal nonlinear modeling
- Binary choice modeling with panel data
  - Models for heterogeneity
  - Estimation strategies
    - Unconditional and conditional
    - Fixed and random effects
- The incidental parameter problem
- JW chapter 15, Baltagi, ch. 11, Hsiao ch. 7, Greene ch. 23.

A Random Utility Approach

- Underlying Preference Scale, U*(choices)
- Revelation of Preferences:
  - U*(choices) $\leq$ 0 → Choice "0"
  - U*(choices) > 0 → Choice "1"

Binary Outcome: Visit Doctor



A Model for Binary Choice

- Yes or No decision (Buy/NotBuy, Do/NotDo)
- Example, choose to visit physician or not
- Model: Net utility of visit at least once

  $U_{visit} = \alpha+\beta_1 Age + \beta_2 Income + \gamma Sex + \varepsilon$

  Choose to visit if net utility is positive

  Net utility = $U_{visit} - U_{not\ visit}$

  Random Utility

- Data: **X** = [1,age,income,sex]
  y = 1 if choose visit, $\Leftrightarrow U_{visit} > 0$, 0 if not.

Choosing Between the Two Alternatives

Modeling the Binary Choice

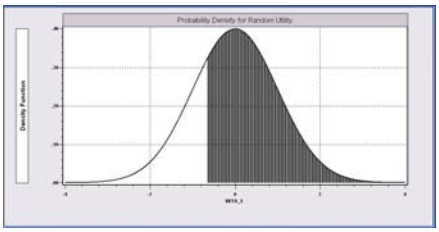
$$U_{visit} = \alpha + \beta_1 \text{Age} + \beta_2 \text{Income} + \beta_3 \text{Sex} + \varepsilon$$

Chooses to visit: $U_{visit} > 0$

$$\alpha + \beta_1 \text{Age} + \beta_2 \text{Income} + \beta_3 \text{Sex} + \varepsilon > 0$$

$$\varepsilon > -[\alpha + \beta_1 \text{Age} + \beta_2 \text{Income} + \beta_3 \text{Sex}]$$

Probability Model for Choice Between Two Alternatives



$$\varepsilon > -[\alpha + \beta_1 \text{Age} + \beta_2 \text{Income} + \beta_3 \text{Sex}]$$

Application

27,326 Observations

- 1 to 7 years, panel
- 7,293 households observed
- We use the 1994 year, 3,337 household observations

Descriptive Statistics for 4 variables

Variable	Mean	Std.Dev.	Minimum	Maximum	Cases	Missing
DOCTOR	.657980	.474456	0.0	1.0	3377	0
AGE	42.62659	11.58599	25.0	64.0	3377	0
INCOME	.444764	.216586	.034000	3.0	3377	0
FEMALE	.463429	.498735	0.0	1.0	3377	0

Example10.do

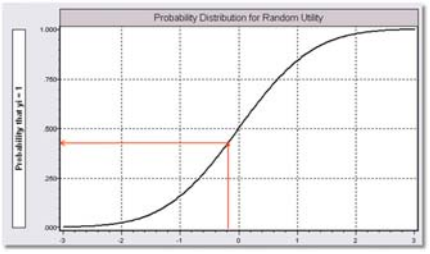
Binary Choice Data

Listing of current sample

Line	Observation	ID	DOCTOR	AGE	INCOME	FEMALE
1	1	1	1	54	.30500	0
2	2	1	0	55	.45101	0
3	3	1	0	56	.35000	0
4	4	2	0	44	.30500	1
5	5	2	1	45	.31828	1
6	6	2	1	46	.35000	1
7	7	2	1	48	.35305	1
8	8	3	0	58	.14340	1
9	9	3	0	60	.30000	1
10	10	3	1	61	.11000	1
11	11	3	1	62	.10000	1
12	12	4	1	29	.13000	1
13	13	5	1	27	.06500	0
14	14	5	1	28	.06000	0
15	15	5	0	31	.15500	0
16	16	6	1	25	.16000	0
17	17	6	1	26	.30000	0
18	18	6	0	27	.30000	0
19	19	6	1	28	.20000	0
20	20	6	1	31	.18000	0
21	21	7	0	26	.30000	1
22	22	7	0	27	.20000	1
23	23	7	1	30	.18000	1
24	24	8	1	64	.15000	0
25	25	9	1	30	.24000	0

An Econometric Model

- Choose to visit iff $U_{visit} > 0$
  - $U_{visit} = \alpha + \beta_1 \text{Age} + \beta_2 \text{Income} + \beta_3 \text{Sex} + \varepsilon$
  - $U_{visit} > 0 \Leftrightarrow \varepsilon > -(\alpha + \beta_1 \text{Age} + \beta_2 \text{Income} + \beta_3 \text{Sex})$
    $\varepsilon < \alpha + \beta_1 \text{Age} + \beta_2 \text{Income} + \beta_3 \text{Sex}$
- Probability model: For any person observed by the analyst,
  $\text{Prob(visit)} = \text{Prob}[\varepsilon \leq \alpha + \beta_1 \text{Age} + \beta_2 \text{Income} + \beta_3 \text{Sex}]$
- Note the relationship between the unobserved ε and the outcome

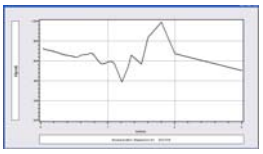


$\alpha + \beta_1 \text{Age} + \beta_2 \text{ Income} + \beta 3 \text{ Sex}$

Modeling Approaches

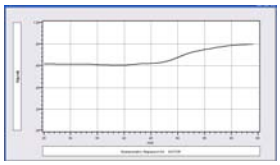
- Nonparametric – "relationship"
  - Minimal Assumptions
  - Minimal Conclusions
- Semiparametric – "index function"
  - Stronger assumptions
  - Robust to model misspecification (heteroscedasticity)
  - Still weak conclusions
- Parametric – "Probability function and index"
  - Strongest assumptions – complete specification
  - Strongest conclusions
  - Possibly less robust. (Not necessarily)

Nonparametric Regressions



P(Visit)=f(Income)

P(Visit)=f(Age)



Linear Probability Model

- Prob(y=1|x)=$\boldsymbol{\beta}'\mathbf{x}$
- Upside
  - Easy to compute using LS. (Not really)
  - Can use 2SLS
- Downside
  - Probabilities not between 0 and 1
  - "Disturbance" is binary – makes no statistical sense
  - Heteroscedastic
  - Statistical underpinning is inconsistent with the data

Fully Parametric

- Index Function: $U^* = \boldsymbol{\beta}'\mathbf{x} + \varepsilon$
- Observation Mechanism: $y = 1[U^* > 0]$
- Distribution: $\varepsilon \sim f(\varepsilon)$; Normal, Logistic, …
- Maximum Likelihood Estimation:

$$\text{Max}(\boldsymbol{\beta})\ \log L = \Sigma_i \log \text{Prob}(Y_i = y_i | x_i)$$

Parametric Model Estimation

- How to estimate $\alpha, \beta_1, \beta_2, \beta_3$?
  - The technique of maximum likelihood

$$L = \prod_{y=0} \text{Prob}[y = 0 \mid \mathbf{x}] \times \prod_{y=1} \text{Prob}[y = 1 \mid \mathbf{x}]$$

  - $\text{Prob}[y=1] = \text{Prob}[\varepsilon > -(\alpha + \beta_1 \text{Age} + \beta_2 \text{Income} + \beta_3 \text{Sex})]$

    $\text{Prob}[y=0] = 1 - \text{Prob}[y=1]$
- Requires a model for the probability

Completing the Model: F(ε)

- The distribution
  - Normal: **PROBIT**, natural for behavior
  - Logistic: **LOGIT**, allows "thicker tails"
  - Gompertz: **EXTREME VALUE**, asymmetric
  - Others…
- Does it matter?
  - Yes, large difference in estimates
  - Not much, quantities of interest are more stable.

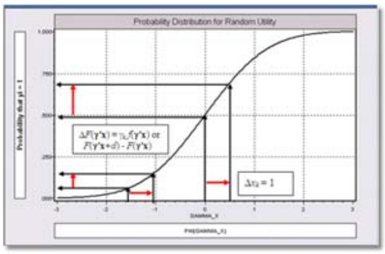
Estimated Binary Choice Models

	LOGIT		PROBIT		EXTREME VALUE	
Variable	Estimate	t-ratio	Estimate	t-ratio	Estimate	t-ratio
Constant	-0.42085	-2.662	-0.25179	-2.600	0.00960	0.078
Age	0.02365	7.205	0.01445	7.257	0.01878	7.129
Income	-0.44198	-2.610	-0.27128	-2.635	-0.32343	-2.536
Sex	0.63825	8.453	0.38685	8.472	0.52280	8.407
Log-L	-2097.48		-2097.35		-2098.17	
Log-L(0)	-2169.27		-2169.27		-2169.27	

Ignore the t ratios for now.

Example10.do

Effect on Predicted Probability of an Increase in Age



$\alpha + \beta_1 (Age+1) + \beta_2$ (Income) $+ \beta_3$ Sex

(β_1 is positive)

Partial Effects in Probability Models

- Prob[Outcome] = some $F(\alpha+\beta_1 Income\ldots)$
- "Partial effect" = $\partial F(\alpha+\beta_1 Income\ldots) / \partial$"x" (derivative)
  - Partial effects are derivatives
  - Result varies with model
    - Logit: $\partial F(\alpha+\beta_1 Income\ldots)/\partial x$ = Prob * (1-Prob) × β
    - Probit: $\partial \Phi(\alpha+\beta_1 Income\ldots)/\partial x$ = Normal density × β
    - Extreme Value: $\partial F(\alpha+\beta_1 Income\ldots)/\partial x$ = Prob * (-log Prob) × β
  - Scaling usually erases model differences

Estimated Partial Effects

	LOGIT		PROBIT		EXTREME VALUE	
	Estimate	t ratio	Estimate	t ratio	Estimate	t ratio
Age	.00527	7.235	.00527	7.269	.00506	6.291
Income	-.09844	-2.611	-.09897	-2.636	-.09711	-2.527
Female	.14026	8.663	.13958	8.264	.13539	8.747

Example10.do

Partial Effect for a Dummy Variable

- Prob[$y_i = 1|\mathbf{x}_i, d_i$] = $F(\boldsymbol{\beta}'\mathbf{x}_i+\gamma d_i)$
  = conditional mean
- Partial effect of d
  Prob[$y_i = 1|\mathbf{x}_i, d_i=1$] - Prob[$y_i = 1|\mathbf{x}_i, d_i=0$]
- Probit: $\delta(d_i) = \Phi\left(\hat{\boldsymbol{\beta}}'\bar{\mathbf{x}} + \hat{\gamma}\right) - \Phi\left(\hat{\boldsymbol{\beta}}'\bar{\mathbf{x}}\right)$

Partial Effect for Nonlinear Terms

$$\text{Prob} = \Phi[\alpha + \beta_1 \text{Age} + \beta_2 \text{Age}^2 + \beta_3 \text{Income} + \beta_4 \text{Female}]$$

$$\frac{\partial \text{Prob}}{\partial \text{Age}} = \phi[\alpha + \beta_1 \text{Age} + \beta_2 \text{Age}^2 + \beta_3 \text{Income} + \beta_4 \text{Female}] \times (\beta_1 + 2\beta_2 \text{Age})$$

$$= \phi(1.30811 - .06487\,Age + .0091\,Age^2 - .17362\,Income + .39666\,Female) \times [(-.06487 + 2(.0091)\,Age]$$

Must be computed at specific values of Age, Income and Female

Example10.do

Odds Ratios

This calculation is not meaningful if the model is not a binary logit model

$$OR(\mathbf{x},z) = \frac{Prob(y=1\,|\,\mathbf{x},z)}{Prob(y=0\,|\,\mathbf{x},z)} = \frac{\exp(\boldsymbol{\beta}'\mathbf{x}+\gamma z)}{1}$$
$$= \exp(\boldsymbol{\beta}'\mathbf{x}+\gamma z)$$
$$= \exp(\boldsymbol{\beta}'\mathbf{x})\exp(\gamma z)$$

Example10.do

Odds Ratio

- Exp(γ) = **multiplicative** change in the odds ratio when z changes by 1 unit.
- $dOR(\mathbf{x},z)/d\mathbf{x} = OR(\mathbf{x},z)*\boldsymbol{\beta}$, not $\exp(\boldsymbol{\beta})$
- The "odds ratio" is not a partial effect – it is not a derivative.
- It is only meaningful when the odds ratio is itself of interest and the change of the variable by a whole unit is meaningful.
- "Odds ratios" might be interesting for dummy variables

Cautions About reported Odds Ratios

```
. logit  grade gpa tuce psi, nolog

Logit estimates                                   Number of obs   =         32
                                                  LR chi2(3)      =      15.40
                                                  Prob > chi2     =     0.0015
Log likelihood = -12.889633                       Pseudo R2       =     0.3740

------------------------------------------------------------------------------
       grade |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         gpa |   2.826113   1.262941     2.24   0.025                         
        tuce |   .0951577   .1415542     0.67   0.501    -.1822835    .3725988
         psi |   2.378688   1.064564     2.23   0.025      .29218    4.465195
       _cons |  -13.02135   4.931325    -2.64   0.008    -22.68657   -3.35613
------------------------------------------------------------------------------

. logit  grade gpa tuce psi, or nolog

Logit estimates                                   Number of obs   =         32
                                                  LR chi2(3)      =      15.40
                                                  Prob > chi2     =     0.0015
Log likelihood = -12.889633                       Pseudo R2       =     0.3740

------------------------------------------------------------------------------
       grade | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         gpa |                                                               
        tuce |   1.099852   .1556857     0.67   0.501     .8333681    1.451904
         psi |   10.79073   11.48743     2.23   0.025     1.339344    86.93802
------------------------------------------------------------------------------
```

Measuring Fit

How Well Does the Model Fit?

- **There is no R squared.**
  - Least squares for linear models is computed to maximize R^2
  - There are no residuals or sums of squares in a binary choice model
  - The model is not computed to optimize the fit of the model to the data
- **How can we measure the "fit" of the model to the data?**
  - "Fit measures" computed from the log likelihood
    - **"Pseudo R squared" = 1 – logL/logL0**
    - Also called the "likelihood ratio index"
    - Others… - these do not measure fit.
  - Direct assessment of the effectiveness of the model at predicting the outcome

Log Likelihoods

- logL = $\sum_i$ log density ($y_i|\mathbf{x}_i,\boldsymbol{\beta}$)
- For probabilities
  - Density is a probability
  - Log density is < 0
  - LogL is < 0
- For other models, log density can be positive or negative.
  - For linear regression, logL=-N/2(1+log2π+log(**e'e**/N)]
  - Positive if $s^2 < .058497$

Likelihood Ratio Index

$$\log L = \sum_{i=1}^{N} \{(1-y_i)\log[1-F(\boldsymbol{\beta}'\mathbf{x}_i)] + y_i \log F(\boldsymbol{\beta}'\mathbf{x}_i)\}$$

1. Suppose the model predicted $F(\boldsymbol{\beta}'\mathbf{x}_i) = 1$ whenever y=1 and $F(\boldsymbol{\beta}'\mathbf{x}_i) = 0$ whenever y=0. Then, logL = 0. [$F(\boldsymbol{\beta}'\mathbf{x}_i)$ cannot equal 0 or 1 at any finite β.]
2. Suppose the model always predicted the same value, $F(\beta_0)$

$$\begin{aligned}\text{LogL}_0 &= \sum_{i=1}^{N} \{(1-y_i)\log[1-F(\beta_0)] + y_i \log F(\beta_0)\} \\ &= N_0 \log[1-F(\beta_0)] + N_1 \log F(\beta_0) \\ &< 0\end{aligned}$$

$\text{LRI} = 1 - \dfrac{\log L}{\log L_0}$. Since $\text{logL} > \text{logL}_0$ $0 \le \text{LRI} < 1$.

Fit Measures Based on Predictions

- Computation
  - Use the model to compute predicted probabilities
  - Use the model and a rule to compute predicted y = 0 or 1
- Fit measure compares predictions to actuals

Predicting the Outcome

- Predicted probabilities

  $\mathbf{P = F(a + b_1 Age + b_2 Income + b_3 Female + \ldots)}$
- Predicting outcomes
  - Predict y=1 if P is "large"
  - Use 0.5 for "large" (more likely than not)
  - Generally, use $\hat{y} = 1$ if $\hat{P} > P^*$
- Count successes and failures

Cramer Fit Measure

$\hat{F}$ = Predicted Probability

$$\hat{\lambda} = \frac{\Sigma_{i=1}^{N} y_i \hat{F}}{N_1} - \frac{\Sigma_{i=1}^{N} (1-y_i)\hat{F}}{N_0}$$

$$\hat{\lambda} = \left(\text{Mean } \hat{F} \,|\, \text{when } y = 1\right) - \left(\text{Mean } \hat{F} \,|\, \text{when } y = 0\right)$$

= reward for correct predictions minus penalty for incorrect predictions

```
+----------------------------------------+
| Fit Measures Based on Model Predictions|
| Efron                          =   .04825|
| Ben Akiva and Lerman           =   .57139|
| Veall and Zimmerman            =   .08365|
| Cramer                         =   .04771|
+----------------------------------------+
```

Hypothesis Testing in Binary Choice Models

Base Model for Hypothesis Tests

```
---------------------------------------------------------------------
Binary Logit Model for Binary Choice
Dependent variable                DOCTOR
Log likelihood function      -2085.92452
Restricted log likelihood    -2169.26982
Chi squared [   5 d.f.]        166.69058
Significance level               .00000
McFadden Pseudo R-squared      .0384209
Estimation based on N =   3377, K =   6
Information Criteria: Normalization=1/N
               Normalized   Unnormalized
AIC              1.23892     4183.84905
--------+------------------------------------------------------------
Variable| Coefficient    Standard Error  b/St.Er. P[|Z|>z]   Mean of X
--------+------------------------------------------------------------
        |Characteristics in numerator of Prob[Y = 1]
Constant|    1.86428***       .67793        2.750   .0060
     AGE|    -.10209***       .03056       -3.341   .0008      42.6266
   AGESQ|     .00154***       .00034        4.556   .0000      1951.22
  INCOME|     .51206          .74600         .686   .4925       .44476
 AGE_INC|    -.01843          .01691       -1.090   .2756      19.0288
  FEMALE|     .65366***       .07588        8.615   .0000       .46343
--------+------------------------------------------------------------
```

H_0: Age is not a significant determinant of Prob(Doctor = 1)

H_0: $\beta_2 = \beta_3 = \beta_5 = 0$

Example10.do

Endogeneity

Endogenous RHS Variable

- $U^* = \boldsymbol{\beta}'\mathbf{x} + \theta h + \varepsilon$
  $y = 1[U^* > 0]$
  $E[\varepsilon|h] \neq 0$ (h is endogenous)
  - Case 1: h is continuous
  - Case 2: h is binary = a treatment effect
- Approaches
  - Parametric: Maximum Likelihood
  - Semiparametric (not developed here):
    - GMM
    - Various approaches for case 2

Endogenous Continuous Variable

$U^* = \boldsymbol{\beta}'\mathbf{x} + \theta h +$ 
$y = 1[U^* > 0]$
$h = \boldsymbol{\alpha}'\mathbf{z} +$

⇐ Correlation = ρ. This is the source of the endogeneity

$E[\varepsilon|h] \neq 0 \Leftrightarrow \text{Cov}[u, \varepsilon] \neq 0$

Additional Assumptions:

$(u,\varepsilon) \sim N[(0,0),(\sigma_u^2, \rho\sigma_u, 1)]$

$\mathbf{z}$ = a valid set of exogenous variables, uncorrelated with (u,ε)

This is not IV estimation. Z may be uncorrelated with X without problems.

Estimation by ML (Control Function)

Probit fit of y to $\mathbf{x}$ and h will not consistently estimate $(\boldsymbol{\beta},\theta)$ because of the correlation between h and ε induced by the correlation of u and ε. Using the bivariate normality,

$$\text{Prob}(y=1\,|\,\mathbf{x},h)=\Phi\left[\frac{\boldsymbol{\beta}'\mathbf{x}+\theta h+(\rho/\sigma_u)u}{\sqrt{1-\rho^2}}\right]$$

Insert $u_i = (h_i - \boldsymbol{\alpha}'\mathbf{z})/\sigma_u$ and include f(h|$\mathbf{z}$) to form logL

$$\text{logL}=\sum_{i=1}^{N}\left\{\log\Phi\left[(2y_i-1)\left(\frac{\boldsymbol{\beta}'\mathbf{x}_i+\theta h_i+\rho\left(\frac{h_i-\boldsymbol{\alpha}'\mathbf{z}_i}{\sigma_u}\right)}{\sqrt{1-\rho^2}}\right)\right]+\log\frac{1}{\sigma_u}\phi\left[\left(\frac{h_i-\boldsymbol{\alpha}'\mathbf{z}_i}{\sigma_u}\right)\right]\right\}$$

Two Approaches to ML

(1) **Full information ML.** Maximize the full log likelihood with respect to $(\boldsymbol{\beta},\theta,\sigma_u,\boldsymbol{\alpha},\rho)$
(The built in Stata routine IVPROBIT does this. It is not an instrumental variable estimator; it is a FIML estimator.)
Note also, this does not imply replacing h with a prediction from the regression then using probit with $\hat{h}$ instead of h.

(2) **Two step limited information ML. (Control Function)**
(a) Use OLS to estimate $\boldsymbol{\alpha}$ and σ_u with $\mathbf{a}$ and s.
(b) Compute $\hat{v}_i = \hat{u}_i/s = (h_i - \mathbf{a}'\mathbf{z}_i)/s$

(c) $$\log\hat{\Phi}\left[\frac{\boldsymbol{\beta}'\mathbf{x}_i+\theta h_i+\rho\hat{v}_i}{\sqrt{1-\rho^2}}\right]=\log\hat{\Phi}\left[\boldsymbol{\delta}'\mathbf{x}_i+\lambda h_i+\tau\hat{v}_i\right]$$

The second step is to fit a probit model for y to $(\mathbf{x},h,\hat{v})$ then solve back for $(\boldsymbol{\beta},\theta,\rho)$ from $(\boldsymbol{\delta},\lambda,\tau)$ and from the previously estimated $\mathbf{a}$ and s. Use the delta method to compute standard errors.

FIML Estimates

```
------------------------------------------------------------------------------
Probit with Endogenous RHS Variable
Dependent variable                HEALTHY
Log likelihood function     -6464.60772
--------+---------------------------------------------------------------------
Variable| Coefficient    Standard Error  b/St.Er. P[|Z|>z]   Mean of X
--------+---------------------------------------------------------------------
        |Coefficients in Probit Equation for HEALTHY
Constant|    1.21760***        .06359      19.149   .0000
     AGE|     -.02426***        .00081     -29.864   .0000      43.5257
 MARRIED|     -.02599           .02329      -1.116   .2644       .75862
  HHKIDS|      .06932***        .01890       3.668   .0002       .40273
  FEMALE|    -.14180***        .01583      -8.959   .0000       .47877
  INCOME|      .53778***        .14473       3.716   .0002       .35208
        |Coefficients in Linear Regression for INCOME
Constant|    -.36099***        .01704     -21.180   .0000
     AGE|      .02159***        .00083      26.062   .0000      43.5257
   AGESQ|     -.00025***     .944134D-05   -26.569   .0000      2022.86
    EDUC|      .02064***        .00039      52.729   .0000      11.3206
 MARRIED|      .07783***        .00259      30.080   .0000       .75862
  HHKIDS|    -.03564***        .00232     -15.332   .0000       .40273
  FEMALE|      .00413**         .00203       2.033   .0420       .47877
        |Standard Deviation of Regression Disturbances
Sigma(w)|     .16445***        .00026     644.874   .0000
```

Example10.do

Endogenous Binary Variable

$U^* = \boldsymbol{\beta}'\mathbf{x} + \theta h + \varepsilon$
$y = 1[U^* > 0]$
$h^* = \boldsymbol{\alpha}'\mathbf{z} + u$
$h = 1[h^* > 0]$

⇐ Correlation = ρ. This is the source of the endogeneity

$E[\varepsilon|h^*] \neq 0 \Leftrightarrow Cov[u, \varepsilon] \neq 0$

Additional Assumptions:

$(u,\varepsilon) \sim N[(0,0),(\sigma_u^2, \rho\sigma_u, 1)]$

$\mathbf{z}$ = a valid set of exogenous variables, uncorrelated with (u,ε)

This is not IV estimation. Z may be uncorrelated with X without problems.

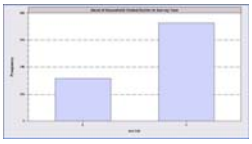
Endogenous Binary Variable

$P(Y = y, H = h) = P(Y = y|H = h) \times P(H=h)$

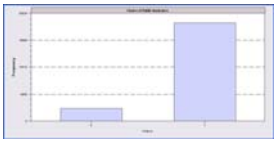
This is a simple bivariate probit model.

Not a simultaneous equations model - the estimator is FIML, not any kind of least squares.

Doctor = F(age,age²,income,female,Public)



Public = F(age,educ,income,married,kids,female)



FIML Estimates

```
------------------------------------------------------------------------------
FIML Estimates of Bivariate Probit Model
Dependent variable                DOCPUB
Log likelihood function    -25671.43905
Estimation based on N =  27326, K =  14
--------+---------------------------------------------------------------------
Variable| Coefficient     Standard Error  b/St.Er. P[|Z|>z]    Mean of X
--------+---------------------------------------------------------------------
        |Index    equation for DOCTOR
Constant|     .59049***       .14473       4.080   .0000
     AGE|    -.05740***       .00601      -9.559   .0000      43.5257
   AGESQ|     .00082***     .681660D-04    12.100   .0000      2022.86
  INCOME|     .08883*         .05094       1.744   .0812       .35208
  FEMALE|     .34583***       .01629      21.225   .0000       .47877
  PUBLIC|     .43533***       .07357       5.917   .0000       .88571
        |Index    equation for PUBLIC
Constant|    3.55054***       .07446      47.681   .0000
     AGE|     .00067         .00115        .581   .5612      43.5257
    EDUC|    -.16839***       .00416     -40.499   .0000      11.3206
  INCOME|    -.98656***       .05171     -19.077   .0000       .35208
 MARRIED|    -.00985         .02922       -.337   .7361       .75862
   HHKIDS|   -.08095***       .02510      -3.225   .0013       .40273
  FEMALE|     .12139***       .02231       5.442   .0000       .47877
```

Example10.do

Partial Effects

Conditional Mean

$$E[y \mid \mathbf{x}, h] = \Phi(\boldsymbol{\beta}'\mathbf{x} + \theta h)$$

$$\begin{aligned} E[y \mid \mathbf{x}, \mathbf{z}] &= E_h E[y \mid \mathbf{x}, h] \\ &= \text{Prob}(h = 0 \mid \mathbf{z}) E[y \mid \mathbf{x}, h = 0] + \text{Prob}(h = 1 \mid \mathbf{z}) E[y \mid \mathbf{x}, h = 1] \\ &= \Phi(-\boldsymbol{\alpha}'\mathbf{z})\Phi(\boldsymbol{\beta}'\mathbf{x}) + \Phi(\boldsymbol{\alpha}'\mathbf{z})\Phi(\boldsymbol{\beta}'\mathbf{x} + \theta) \end{aligned}$$

Partial Effects

$$= \left[\Phi(-\boldsymbol{\alpha}'\mathbf{z})\phi(\boldsymbol{\beta}'\mathbf{x}) + \Phi(\boldsymbol{\alpha}'\mathbf{z})\phi(\boldsymbol{\beta}'\mathbf{x} + \theta)\right] \boldsymbol{\beta}$$

$$= \left[-\phi(-\boldsymbol{\alpha}'\mathbf{z})\Phi(\boldsymbol{\beta}'\mathbf{x}) + \phi(\boldsymbol{\alpha}'\mathbf{z})\Phi(\boldsymbol{\beta}'\mathbf{x} + \theta)\right] \boldsymbol{\alpha}$$

$$= \phi(\boldsymbol{\alpha}'\mathbf{z})\left[\Phi(\boldsymbol{\beta}'\mathbf{x} + \theta) - \Phi(\boldsymbol{\beta}'\mathbf{x})\right] \boldsymbol{\alpha}$$

Sample Selection Problem

Canonical Sample Selection Model

Regression Equation

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

Sample Selection Mechanism

$d^* = z'\gamma + u$; $d = 1[d^* > 0]$ (probit)

$y = y^*$ if $d = 1$; not observed otherwise

Is the sample 'nonrandomly selected?'

$$\begin{aligned} E[y^* \mid x, d=1] &= x'\beta + E[\varepsilon \mid x, d = 1] \\ &= x'\beta + E[\varepsilon \mid x, u > -z'\gamma] \\ &= x'\beta + \text{something if } \text{Cor}[\varepsilon, u \mid x] \neq 0 \end{aligned}$$

A left out variable problem (again)

Incidental truncation

Heckman's Model

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

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

$y_i = y_i^*$ if $d_i = 1$; not observed otherwise

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

$$E[y_i^* | x_i, d_i = 1] = \mathbf{x}_i'\boldsymbol{\beta} + E[\varepsilon_i | x_i, d_i = 1]$$
$$= \mathbf{x}_i'\boldsymbol{\beta} + E[\varepsilon_i | x_i, u_i > -\mathbf{z}_i'\boldsymbol{\gamma}]$$
$$= \mathbf{x}_i'\boldsymbol{\beta} + \rho\sigma\left(\frac{\phi(\mathbf{z}_i'\boldsymbol{\gamma})}{\Phi(\mathbf{z}_i'\boldsymbol{\gamma})}\right)$$
$$= \mathbf{x}_i'\boldsymbol{\beta} + \rho\sigma\lambda_i$$

Least squares is biased and inconsistent again. Left out variable

Two Step Estimation

Step 1: Estimate the probit model

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

Estimation of $\boldsymbol{\gamma}$ by $\hat{\boldsymbol{\gamma}}$. Now compute $\hat{\lambda}_i = \left(\frac{\phi(\mathbf{z}_i'\hat{\boldsymbol{\gamma}})}{\Phi(\mathbf{z}_i'\hat{\boldsymbol{\gamma}})}\right)$

Step 2: Estimate the regression model with estimated regressor

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

$y_i = y_i^*$ if $d_i = 1$; not observed otherwise

$$E[y_i^* | x_i, d_i = 1] = \mathbf{x}_i'\boldsymbol{\beta} + E[\varepsilon_i | x_i, d_i = 1]$$
$$= \mathbf{x}_i'\boldsymbol{\beta} + \theta\lambda_i$$

Linearly regress y_i on $x_i, \hat{\lambda}_i$.

The "LAMBDA"

Step2a. Fix standard errors (Murphy and Topel). Estimate ρ and σ using $\hat{\theta}$ and **e'e**/n

Classic Application

- Mroz, T., Married women's labor supply, Econometrica, 1987.
  - N =753
  - N_1 = 428
- A specification
  - LFP=f(age,age^2,family income, education, kids)
  - Wage=g(experience, exp^2, education, city)

```
use E:\paneldata\paneldata\EBAPE\aula7\mroz.dta
gen age2=age*age
heckman wage exper expersq city, select(inlf =age age2 faminc
   educ kidslt6 kidsge6)
heckman lwage educ exper expersq, select(nwifeinc educ exper
   expersq age kidslt6 kidsge6) twostep
```

Sample Selection in Probit

We use the data from Pindyck and Rubinfeld (1998). In this dataset, the variables are whether children attend private school (private), number of years the family has been at the present residence (years), log of property tax (logptax), log of income (loginc), and whether one voted for an increase in property taxes (vote). In this example, we alter the meaning of the data. Here we assume that we observe whether children attend private school only if the family votes for increasing the property taxes. This assumption is not true in the dataset, and we make it only to illustrate the use of this command.

```
webuse school
heckprob private years logptax, sel(vote=years
loginc logptax)
```