Day 5
Limited Dependent Variable Models (Brief)
Binary, multinomial, censored, treatment effects

© A. Colin Cameron
Univ. of Calif. - Davis

Advanced Econometrics
Bavarian Graduate Program in Economics

.

Based on A. Colin Cameron and Pravin K. Trivedi (2005), Microeconometrics: Methods and Applications (MMA), C.U.P.
A. Colin Cameron and Pravin K. Trivedi (2009, 2010), Microeconometrics using Stata (MUS), Stata Press.

July 22-26, 2013

1. Introduction

- Abbreviated handout: assumes previous exposure to nonlinear models.
- Binary outcomes
  - y takes only one of two values, say 0 or 1.
  - model $\Pr[y = 1|\mathbf{x}]$
  - logit and probit are standard
- Multinomial outcomes
  - y takes only m possible outcomes.
  - model $\Pr[y = j|\mathbf{x}]$ for $j = 1, ..., m$
  - many models including multinomial logit.
- Censored and truncated models (e.g. Tobit) and selection models
  - Considerably more difficult conceptually.
  - Sample is not reflective of the population (selection on y)
  - Standard methods rely on strong distributional assumptions.
- Treatment evaluation

Outline

2. Logit model: Definition

- Data y takes only one of two values, say 0 or 1.
  - OLS has problem that $\mathrm{E}[y_i|\mathbf{x}_i] = \mathbf{x}_i'\boldsymbol{\beta} > 1$ or < 0 is possible
  - And OLS is inefficient (based on homoskedasticity, normality).
  - So what do we do?
- Starting point from statistics is Bernoulli (binomial with 1 trial):

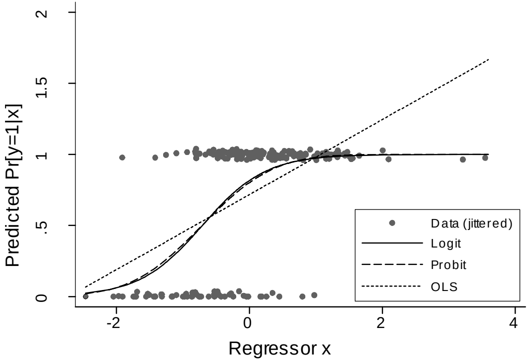
$$\begin{aligned} \Pr[y=1] &= p \\ \Pr[y=0] &= 1-p. \end{aligned}$$

  - with $\mathrm{E}[y] = p$ and $\mathrm{V}[y] = p(1-p)$.
- For regression the probability $0 < p_i < 1$ varies with regressors $\mathbf{x}_i$

Logit	$p_i = \Lambda(\mathbf{x}_i'\boldsymbol{\beta}) = \frac{\exp(\mathbf{x}_i'\boldsymbol{\beta})}{1+\exp(\mathbf{x}_i'\boldsymbol{\beta})}$	$\Lambda(\cdot)$ is logistic c.d.f.
Probit	$p_i = \Phi(\mathbf{x}_i'\boldsymbol{\beta})$	$\Phi(\cdot)$ is standard normal c.d.f.

Example

- A single regressor example allows a nice plot.
- Compare predictions of $\Pr[y = 1|x]$ from logit, probit and OLS.
    - Scatterplot of $y = 0$ or 1 (jittered) on scalar x (data are generated).



- Logit similar to probit with predictions between 0 and 1.
  OLS predicts outside the $(0, 1)$ interval.

Logit and Probit MLE

- Useful notation: The Bernoulli density can be written in compact notation as

$$f(y_i|\mathbf{x}_i) = p_i^{y_i}(1-p_i)^{1-y_i}.$$

- Log-likelihood function:

$$\begin{aligned} \ln L(\boldsymbol{\beta}) &= \ln\left(\prod_{i=1}^N f(y_i|\mathbf{x}_i)\right) \\ &= \sum_{i=1}^N \ln f(y_i|\mathbf{x}_i) \\ &= \sum_{i=1}^N \ln\left(p_i^{y_i}(1-p_i)^{1-y_i}\right) \\ &= \sum_{i=1}^N \{y_i \ln p_i + (1-y_i)\ln(1-p_i)\} \end{aligned}$$

- MLE solves $\partial \ln L(\boldsymbol{\beta})/\partial\boldsymbol{\beta} = \mathbf{0}$. After considerable algebra

$$\begin{array}{lll} \text{Logit} & p_i = \Lambda(\mathbf{x}_i'\boldsymbol{\beta}) & \sum_{i=1}^N (y_i - \Lambda(\mathbf{x}_i'\boldsymbol{\beta}))\mathbf{x}_i = \mathbf{0} \\ \text{Probit} & p_i = \Phi(\mathbf{x}_i'\boldsymbol{\beta}) & \sum_{i=1}^N (y_i - \Phi(\mathbf{x}_i'\boldsymbol{\beta}))\frac{\Phi'(\mathbf{x}_i'\boldsymbol{\beta})}{\Phi(\mathbf{x}_i'\boldsymbol{\beta})(1-\Phi(\mathbf{x}_i'\boldsymbol{\beta}))}\mathbf{x}_i = \mathbf{0}. \end{array}$$

Properties of MLE

- The distribution is necessarily Bernoulli
  - If $\Pr[y_i = 1|\mathbf{x}_i] = p_i$ then necessarily $\Pr[y_i = 0|\mathbf{x}_i] = 1 - p_i$ since the two probabilities must some to one.
  - Only possible error is in p_i.
- So the MLE is consistent if p_i is correctly specified
  - $p_i = \Lambda(\mathbf{x}_i'\boldsymbol{\beta})$ for logit and $p_i = \Phi(\mathbf{x}_i'\boldsymbol{\beta})$ for probit.
- The information matrix equality necessarily holds if data are independent over i and

$$\begin{array}{lll} \text{Logit} & \widehat{\boldsymbol{\beta}}_{\text{ML}} \overset{a}{\sim} \mathcal{N}\left[\boldsymbol{\beta},\ \left(\sum_{i=1}^N \Lambda(\mathbf{x}_i'\boldsymbol{\beta})(1-\Lambda(\mathbf{x}_i'\boldsymbol{\beta}))\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\right] \\ \text{Probit} & \widehat{\boldsymbol{\beta}}_{\text{ML}} \overset{a}{\sim} \mathcal{N}\left[\boldsymbol{\beta},\ \left(\sum_{i=1}^N \frac{(\Phi'(\mathbf{x}_i'\boldsymbol{\beta})^2}{\Phi(\mathbf{x}_i'\boldsymbol{\beta})(1-\Phi(\mathbf{x}_i'\boldsymbol{\beta}))}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\right]. \end{array}$$

- Default ML standard errors implement by using $\widehat{\boldsymbol{\beta}}$ in place of $\boldsymbol{\beta}$.
  - For independent data there is no need for robust se's in this case.

Data Example: Private health insurance

- `ins=1` if have private health insurance.
- Summary statistics (sample is 50-86 years from 2000 HRS)

```
. describe ins retire age hstatusg hhincome educyear married hisp

              storage  display     value
variable name   type   format      label      variable label
------------------------------------------------------------------------------------------
ins             float  %9.0g                  1 if have private health insurance
retire          double %12.0g                 1 if retired
age             double %12.0g                 age in years
hstatusg        float  %9.0g                  1 if health status good of better
hhincome        float  %9.0g                  household annual income in $000's
educyear        double %12.0g                 years of education
married         double %12.0g                 1 if married
hisp            double %12.0g                 1 if hispanic

. summarize ins retire age hstatusg hhincome educyear married hisp

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         ins |      3206    .3870867    .4871597          0          1
      retire |      3206    .6247661    .4842588          0          1
         age |      3206    66.91391    3.675794         52         86
    hstatusg |      3206    .7046163    .4562862          0          1
    hhincome |      3206    45.26391    64.33936          0   1312.124
-------------+--------------------------------------------------------
    educyear |      3206    11.89863    3.304611          0         17
     married |      3206    .7330006     .442461          0          1
        hisp |      3206    .0726762    .2596448          0          1
```

- Summary statistics: by whether or not have private health insurance.

```
. bysort ins: summarize retire age hstatusg hhincome educyear married hisp, sep(0)
```

-> ins = 0

Variable	Obs	Mean	Std. Dev.	Min	Max
retire	1965	.5938931	.49123	0	1
age	1965	66.8229	3.851651	52	86
hstatusg	1965	.653944	.4758324	0	1
hhincome	1965	37.65601	58.98152	0	1197.704
educyear	1965	11.29313	3.475632	0	17
married	1965	.6814249	.4660424	0	1
hisp	1965	.1007634	.3010917	0	1

-> ins = 1

Variable	Obs	Mean	Std. Dev.	Min	Max
retire	1241	.6736503	.469066	0	1
age	1241	67.05802	3.375173	53	82
hstatusg	1241	.7848509	.4110914	0	1
hhincome	1241	57.31028	70.3737	.124	1312.124
educyear	1241	12.85737	2.755311	2	17
married	1241	.8146656	.3887253	0	1
hisp	1241	.0282031	.1656193	0	1

- `ins=1` more likely if retired, older, good health status, richer, more educated, married and nonhispanic.

Logit data example

- Stata command `logit` gives the logit MLE $(p = \Lambda(\mathbf{x}'\boldsymbol{\beta}))$.
  - $\text{ME}_j = \frac{\partial \Pr[y=1|\mathbf{x}]}{\partial x_j} = \Lambda'(\mathbf{x}'\boldsymbol{\beta})\beta_j = \Lambda(\mathbf{x}'\boldsymbol{\beta})(1-\Lambda(\mathbf{x}'\boldsymbol{\beta}))\beta_j$

```
. * Logit regression
. logit ins retire age hstatusg hhincome educyear married hisp

Iteration 0:   log likelihood = -2139.7712
Iteration 1:   log likelihood = -1998.8563
Iteration 2:   log likelihood = -1994.9129
Iteration 3:   log likelihood = -1994.8784
Iteration 4:   log likelihood = -1994.8784

Logistic regression                               Number of obs   =       3206
                                                  LR chi2(7)      =     289.79
                                                  Prob > chi2     =     0.0000
Log likelihood = -1994.8784                       Pseudo R2       =     0.0677
```

ins	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
retire	.1969297	.0842067	2.34	0.019	.0318875	.3619718
age	-.0145955	.0112871	-1.29	0.196	-.0367178	.0075267
hstatusg	.3122654	.0916739	3.41	0.001	.1325878	.491943
hhincome	.0023036	.000762	3.02	0.003	.00081	.0037972
educyear	.1142626	.0142012	8.05	0.000	.0864288	.1420963
married	.578636	.0933198	6.20	0.000	.3957327	.7615394
hisp	-.8103059	.1957522	-4.14	0.000	-1.193973	-.4266387
_cons	-1.715578	.7486219	-2.29	0.022	-3.18285	-.2483064

- Average marginal effect
  $\text{AME}_j = \frac{1}{N}\sum_{i=1}^{N} \frac{\partial \Pr[y_i=1|\mathbf{x}_i]}{\partial x_j} = \frac{1}{N}\sum_{i=1}^{N} \Lambda(\mathbf{x}'\boldsymbol{\beta})(1-\Lambda(\mathbf{x}'\boldsymbol{\beta}))\beta_j$
- Compute AME after `logit` using Stata 11 `margins, dydx(*)` or Stata 10 add-on command `margeff`.

```
. margins, dydx(*)
Warning: cannot perform check for estimable functions.

Average marginal effects                          Number of obs   =       3206
Model VCE    : OIM

Expression   : Pr(ins), predict()
dy/dx w.r.t. : retire age hstatusg hhincome educyear married hisp
```

	dy/dx	Delta-method Std. Err.	z	P>|z|	[95% Conf.	Interval]
retire	.0427616	.018228	2.35	0.019	.0070354	.0784878
age	-.0031693	.0024486	-1.29	0.196	-.0079686	.00163
hstatusg	.0678058	.0197778	3.43	0.001	.0290419	.1065696
hhincome	.0005002	.0001646	3.04	0.002	.0001777	.0008228
educyear	.0248111	.0029705	8.35	0.000	.0189891	.0306332
married	.1256459	.0198205	6.34	0.000	.0867985	.1644933
hisp	-.175951	.0421962	-4.17	0.000	-.258654	-.0932481

- Marginal effects: 0.043, -0.003, 0.067, 0.0005, 0.025, 0.126, -0.176vs. Coefficients: 0.197, -0.015, 0.312, 0.0023, 0.114, 0.579, -0.810.
  - Marginal effect here is about one-fifth the size of the coefficient.

Probit data example

- Stata command probit gives the probit MLE.

```
. probit ins retire age hstatusg hhincome educyear married hisp

Iteration 0:   log likelihood = -2139.7712
Iteration 1:   log likelihood = -1996.0367
Iteration 2:   log likelihood = -1993.6288
Iteration 3:   log likelihood = -1993.6237

Probit regression                                 Number of obs   =       3206
                                                  LR chi2(7)      =     292.30
                                                  Prob > chi2     =     0.0000
Log likelihood = -1993.6237                       Pseudo R2       =     0.0683

------------------------------------------------------------------------------
         ins |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      retire |   .1183567   .0512678     2.31   0.021     .0178737    .2188396
         age |  -.0088696    .006899    -1.29   0.199    -.0223914    .0046521
    hstatusg |   .1977357   .0554868     3.56   0.000     .0889836    .3064877
    hhincome |    .001233   .0003866     3.19   0.001     .0004754    .0019907
    educyear |   .0707477   .0084782     8.34   0.000     .0541308    .0873646
     married |    .362329   .0560031     6.47   0.000     .2525651     .472093
        hisp |  -.4731099   .1104385    -4.28   0.000    -.6895655   -.2566544
       _cons |  -1.069319   .4580791    -2.33   0.020    -1.967138   -.1715009
------------------------------------------------------------------------------
```

- Scaled differently to logit but similar t-statistics (see below).

OLS data example

- OLS estimates for private health insurance
  - If do OLS need to use heteroskedastic-robust standard errors

```
. regress ins retire age hstatusg hhincome educyear married hisp, vce(robust)

Linear regression                                      Number of obs =    3206
                                                       F(  7,  3198) =   58.98
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0826
                                                       Root MSE      =  .46711

------------------------------------------------------------------------------
             |               Robust
         ins |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      retire |   .0408508   .0182217     2.24   0.025     .0051234    .0765782
         age |  -.0028955   .0023254    -1.25   0.213    -.0074549    .0016638
    hstatusg |   .0655583   .0190126     3.45   0.001     .0282801    .1028365
    hhincome |   .0004921   .0001874     2.63   0.009     .0001247    .0008595
    educyear |   .0233686   .0027081     8.63   0.000     .0180589    .0286784
     married |   .1234699   .0186521     6.62   0.000     .0868987    .1600411
        hisp |  -.1210059   .0269459    -4.49   0.000    -.1738389    -.068173
       _cons |   .1270857   .1538816     0.83   0.409    -.1746309    .4288023
------------------------------------------------------------------------------
```

Compare logit, probit and OLS estimates

- Coefficients in different models are not directly comparable!
  - Though the t-statistics are similar.

```
. * Compare coefficient estimates across models with default and robust standard e
. estimates table blogit bprobit bols blogitr bprobitr bolsr, ///
>   stats(N ll) b(%7.3f) t(%7.2f) stfmt(%8.2f)
```

Variable	blogit	bprobit	bols	blogitr	bprobitr	bolsr
retire	0.197	0.118	0.041	0.197	0.118	0.041
	2.34	2.31	2.24	2.32	2.30	2.24
age	-0.015	-0.009	-0.003	-0.015	-0.009	-0.003
	-1.29	-1.29	-1.20	-1.32	-1.32	-1.25
hstatusg	0.312	0.198	0.066	0.312	0.198	0.066
	3.41	3.56	3.37	3.40	3.57	3.45
hhincome	0.002	0.001	0.000	0.002	0.001	0.000
	3.02	3.19	3.58	2.01	2.21	2.63
educyear	0.114	0.071	0.023	0.114	0.071	0.023
	8.05	8.34	8.15	7.96	8.33	8.63
married	0.579	0.362	0.123	0.579	0.362	0.123
	6.20	6.47	6.38	6.15	6.46	6.62
hisp	-0.810	-0.473	-0.121	-0.810	-0.473	-0.121
	-4.14	-4.28	-3.59	-4.18	-4.36	-4.49
_cons	-1.716	-1.069	0.127	-1.716	-1.069	0.127
	-2.29	-2.33	0.79	-2.36	-2.40	0.83
N	3206	3206	3206	3206	3206	3206
ll	-1994.88	-1993.62	-2104.75	-1994.88	-1993.62	-2104.75

legend: b/t

Compare predicted probabilities from models

- Predicted probabilities $\frac{1}{N}\sum_{i=1}^{N} F(\mathbf{x}_i'\widehat{\beta})$ for different models.

```
. * Comparison of predicted probabilities from logit, probit and OLS
. quietly logit ins retire age hstatusg hhincome educyear married hisp

. predict plogit, p

. quietly probit ins retire age hstatusg hhincome educyear married hisp

. predict pprobit, p

. quietly regress ins retire age hstatusg hhincome educyear married hisp

. quietly predict pOLS

. summarize ins plogit pprobit pOLS
```

Variable	Obs	Mean	Std. Dev.	Min	Max
ins	3206	.3870867	.4871597	0	1
plogit	3206	.3870867	.1418287	.0340215	.9649615
pprobit	3206	.3861139	.1421416	.0206445	.9647618
pOLS	3206	.3870867	.1400249	-.1557328	1.197223

- Average probabilities are very close (and for logit and OLS $= \bar{y}$).
- Range similar for logit and probit but OLS gives $\widehat{p}_i < 0$ and $\widehat{p}_i > 1$.

Marginal effects: Approximations for logit and probit

- In general for $p = F(\mathbf{x}'\boldsymbol{\beta})$, $\text{ME}_j = \frac{\partial p}{\partial x_j} = F'(\mathbf{x}'\boldsymbol{\beta}) \times \beta_j$.
  - For OLS: $\text{ME}_j = \widehat{\beta}_j$.
  - For logit: $\text{ME}_j \leq 0.25\widehat{\beta}_j$ as $F'(\mathbf{x}'\boldsymbol{\beta}) = \Lambda(\mathbf{x}'\boldsymbol{\beta})(1-\Lambda(\mathbf{x}'\boldsymbol{\beta})) \leq 0.25$.
  - For probit: $\text{ME}_j \leq 0.40\widehat{\beta}_j$ as $F'(\mathbf{x}'\boldsymbol{\beta}) = \phi(\mathbf{x}'\boldsymbol{\beta}) \leq (1/\sqrt{2\pi}) \simeq 0.40$.
- This leads to the following rule of thumb for slope parameters

$$\begin{aligned} \widehat{\beta}_{\text{Logit}} &\simeq 4\widehat{\beta}_{\text{OLS}} \\ \widehat{\beta}_{\text{Probit}} &\simeq 2.5\widehat{\beta}_{\text{OLS}} \\ \widehat{\beta}_{\text{Logit}} &\simeq 1.6\widehat{\beta}_{\text{Probit}}. \end{aligned}$$

- Also for logit a useful approximation is $\text{ME}_j \simeq \bar{y}(1-\bar{y})\widehat{\beta}_j$.

Which model?

- Logit: binary model most often used by statisticians.
  - generalizes simply to multinomial data ($>$ two outcomes)
  - $\widehat{\beta}_j$ measures change in log-odds ratio $p/(1-p)$ due to x_j change.
- Probit: binary model most often used by economists.
  - motivated by a latent normal random variable.
  - generalizes to Tobit models and multinomial probit.
- Empirically: either logit or probit can be used
  - give similar predictions and marginal effects
  - greatest difference is in prediction of probabilities close to 0 or 1.
- Complementary log-odds model
  - sometimes used when outcomes are mostly 0 or mostly 1.
- OLS: can be useful for preliminary data analysis
  - but final results should use probit or logit.

3. Multinomial models: Definition

- There are m mutually-exclusive alternatives.
  - y takes value j if the outcome is alternative j, $j = 1, ..., m$.
  - Probability that the outcome is alternative j is

$$p_j = \Pr[y = j], \qquad j = 1, ..., m.$$

- Introduce m binary variables for each observed y

$$y_j = \begin{cases} 1 & \text{if } y = j \\ 0 & \text{if } y \neq j. \end{cases}.$$

  - $y_j = 1$ if alternative j is chosen and $y_j = 0$ for all non-chosen alternatives.
  - For an individual exactly one of $y_1, y_2, ..., y_m$ will be non-zero.
- Density for one observation is conveniently written as

$$f(y) = p_1^{y_1} \times p_2^{y_2} \times ... \times p_m^{y_m} = \prod_{j=1}^{m} p_j^{y_j}.$$

Regression Model

- Introduce individual characteristics
  - parameterize p_{ij} in terms of observed data $\mathbf{x}_i$ and parameters $\boldsymbol{\beta}$:

$$p_{ij} = \Pr[y_i = j] = F_j(\mathbf{x}_i, \boldsymbol{\beta}), \quad j = 1, ..., m.$$

  - these probabilities should lie between 0 and 1 and sum over j to one.
- MLE maximizes the log-likelihood function

$$\begin{aligned} \ln L(\cdot) &= \ln \left(\prod_{i=1}^{N} f(y_i) \right) = \ln \left(\prod_{i=1}^{N} \prod_{j=1}^{m} p_j^{y_j} \right) \\ &= \sum_{i=1}^{N} \sum_{j=1}^{m} y_{ij} \ln p_{ij} \end{aligned}$$

- Different models have different models for p_{ij}.
  - e.g. multinomial logit

$$p_{ij} = \Pr[y_i = j] = \frac{\exp(\mathbf{x}_i' \boldsymbol{\beta}_j)}{\sum_{k=1}^{m} \exp(\mathbf{x}_i' \boldsymbol{\beta}_k)}, \ j = 1, ..., m, \quad \boldsymbol{\beta}_1 = \mathbf{0}.$$

  - nested logit, multinomial probit, ordered logit, ... use different p_{ij}.

Data example: Fishing site

- Multinomial variable y has outcome one of
  - $y = 1$ if fish from beach
  - $y = 2$ if fish from pier
  - $y = 3$ if fish from private boat
  - $y = 4$ if fish from charter boat
- Regressors are
  - price: varies by alternative and individual
  - catch rate: varies by alternative and individual
  - income: varies by individual but not alternative

- Variable definitions

```
. describe

Contains data from  mus15data.dta
  obs:         1,182
 vars:            16                          12 May 2008 20:46
 size:        85,104 (99.2% of memory free)
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
mode            float  %9.0g       modetype     Fishing mode
price           float  %9.0g                    price for chosen alternative
crate           float  %9.0g                    catch rate for chosen
                                                  alternative
dbeach          float  %9.0g                    1 if beach mode chosen
dpier           float  %9.0g                    1 if pier mode chosen
dprivate        float  %9.0g                    1 if private boat mode chosen
dcharter        float  %9.0g                    1 if charter boat mode chosen
pbeach          float  %9.0g                    price for beach mode
ppier           float  %9.0g                    price for pier mode
pprivate        float  %9.0g                    price for private boat mode
pcharter        float  %9.0g                    price for charter boat mode
qbeach          float  %9.0g                    catch rate for beach mode
qpier           float  %9.0g                    catch rate for pier mode
qprivate        float  %9.0g                    catch rate for private boat mode
qcharter        float  %9.0g                    catch rate for charter boat mode
income          float  %9.0g                    monthly income in thousands $
-------------------------------------------------------------------------------
```

- Data organization
  - here wide form with one observation per individual
  - each observation has data for all the possible alternatives.

```
. list mode d* p* income in 1/2, clean

         mode    dbeach    dpier   dprivate    dcharter     price    pbeach     ppier    pprivat
> e  pcharter   pmlogit1   pmlogit2   pmlogit3   pmlogit4    income
  1.  charter         0        0          0           1    182.93    157.93    157.93      157.9
> 3    182.93   .1125092   .0919656   .4516733   .3438518   7.083332
  2.  charter         0        0          0           1    34.534    15.114    15.114      10.53
> 4    34.534   .1122198   .2117394   .2635553   .4124855       1.25
```

- Here person 2 chose charter fishing (`mode=charter` or `dcharter=1`) when beach, pier, private and charter fishing cost, respectively, 15.11, 15.11, 10.53 and 34.53.

- Summary statistics
  - Columns $y = 1, ..., 4$ give sample means for those with $y = 1, ..., 4$.

	Sub-sample averages				
Explanatory Variable	y=1	y=2	y=3	y=4	All y
	Beach	Pier	Private	Charter	Overall
Income ($1,000's per month)	4.052	3.387	4.654	3.881	4.099
Price beach ($)	36	31	138	121	103
Price pier ($)	36	31	138	121	103
Price private ($)	98	82	42	45	55
Price charter ($)	125	110	71	75	84
Catch rate beach	0.28	0.26	0.21	0.25	0.24
Catch rate pier	0.22	0.20	0.13	0.16	0.16
Catch rate private	0.16	0.15	0.18	0.18	0.17
Catch rate charter	0.52	0.50	0.65	0.69	0.63
Sample probability	0.113	0.151	0.354	0.382	1.000
Observations	134	178	418	452	1182

- On average a person chooses to fish where it is cheapest to fish.

- Multinomial logit of fishing mode regressed on intercept and income
  - $\Pr[y_{ij}=1] = \frac{e^{\mathbf{x}_i'(\alpha_j+\beta_j \texttt{income})}}{\sum_{k=1}^{4} e^{\mathbf{x}_i'(\alpha_k+\beta_k \texttt{income})}}$, $j = 1, 2, 3, 4$, $\alpha_1 = 0$, $\beta_1 = 0$.
  - normalization that base outcome is beach fishing $(y = 1)$

```
. * Multinomial logit with base outcome alternative 1
. mlogit mode income, baseoutcome(1)

Iteration 0:   log likelihood =   -1497.7229
Iteration 1:   log likelihood =   -1477.5265
Iteration 2:   log likelihood =   -1477.1514
Iteration 3:   log likelihood =   -1477.1506

Multinomial logistic regression                   Number of obs   =       1182
                                                  LR chi2( 3)     =      41.14
                                                  Prob > chi2     =     0.0000
Log likelihood =  -1477.1506                      Pseudo R2       =     0.0137
```

mode	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
pier						
income	-.1434029	.0532882	-2.69	0.007	-.2478459	-.03896
_cons	.8141503	.2286316	3.56	0.000	.3660405	1.26226
private						
income	.0919064	.0406638	2.26	0.024	.0122069	.1716059
_cons	.7389208	.1967309	3.76	0.000	.3533352	1.124506
charter						
income	-.0316399	.0418463	-0.76	0.450	-.1136571	.0503774
_cons	1.341291	.1945167	6.90	0.000	.9600457	1.722537

(mode==beach is the base outcome)

- Predicted probabilities of each outcome:

$$\widehat{\Pr}[y_{ij}=1]=\frac{e^{\mathbf{x}_i'(\widehat{\alpha}_j+\widehat{\beta}_j \texttt{income})}}{\sum_{k=1}^{4} e^{\mathbf{x}_i'(\widehat{\alpha}_k+\widehat{\beta}_k \texttt{income})}}$$

```
. * Compare average predicted probabilities to sample average frequencies
. predict pmlogit1 pmlogit2 pmlogit3 pmlogit4, pr

. summarize pmlogit* dbeach dpier dprivate dcharter, separator(4)
```

Variable	Obs	Mean	Std. Dev.	Min	Max
pmlogit1	1182	.1133672	.0036716	.0947395	.1153659
pmlogit2	1182	.1505922	.0444575	.0356142	.2342903
pmlogit3	1182	.3536379	.0797714	.2396973	.625706
pmlogit4	1182	.3824027	.0346281	.2439403	.4158273
dbeach	1182	.1133672	.3171753	0	1
dpier	1182	.1505922	.3578023	0	1
dprivate	1182	.3536379	.4783008	0	1
dcharter	1182	.3824027	.4861799	0	1

- As expected average predicted probabilities sum to one.
- Furthermore average predicted probabilities of each outcome equals frequency of that outcome
  - Property of multinomial logit and conditional logit
  - Analog of OLS residuals sum to zero so $\overline{\widehat{y}}=\overline{y}$.

- Parameter interpretation is complex.
- There are many marginal effects: one for each outcome value.
  - Here $\text{ME}_{ij} = \partial p_{ij} / \partial \mathbf{x}_i = p_{ij}(\boldsymbol{\beta}_j - \overline{\boldsymbol{\beta}}_i)$ where $\overline{\boldsymbol{\beta}}_i = \sum_l p_{il}\boldsymbol{\beta}_l$.
  - e.g. average marginal effect (AME) of \$1,000 increase in annual income on probability fish from private boat (the third outcome) if a \$1,000 increase in monthly income increases Pr[charter fish] by 0.032.

```
. * AME of income change for outcome 3
. margins, dydx(*) predict(outcome(3))
Warning: cannot perform check for estimable functions.

Average marginal effects                          Number of obs   =       1182
Model VCE    : OIM

Expression   : Pr(mode==3), predict(outcome(3))
dy/dx w.r.t. : income
```

	dy/dx	Delta-method Std. Err.	z	P>|z|	[95% Conf.	Interval]
income	.0317562	.0052589	6.04	0.000	.021449	.0420633

Further details

- $\widehat{\boldsymbol{\beta}}$ is consistently asymptotically normal by the usual asymptotic theory if the d.g.p. is correctly specified.
  - The distribution is necessarily multinomial.
  - So key is correct specification of $p_{ij} = F_j(\mathbf{x}_i, \boldsymbol{\beta})$.
  - And no need to use `vce(robust)` option if independent data.
- Distinguish between two different types of regressors.
  - Alternative-specific or case-specific or alternative-invariant regressors do not vary across alternatives.
    - ★ e.g. income (in our example), gender.
  - Alternative-varying regressors may vary across alternatives.
    - ★ e.g. price.
  - Multinomial logit: all regressors are individual-specific.
  - Conditional logit: same as multinomial logit regressors are alternative varying.

Unordered models

- Unordered model: no obvious ordering of alternatives.
- Additive random utility model (ARUM) specifies utility of each alternative (of m) as

$$\begin{array}{lll} U_1 & = V_1 + \varepsilon_1 \\ U_2 & = V_2 + \varepsilon_2 \\ \vdots & \quad \vdots \quad & \vdots \\ U_m & = V_m + \varepsilon_m \end{array}$$

  - Here V_j is deterministic part of utility, e.g. $V_j = \mathbf{x}'\boldsymbol{\beta}_j$ or $\mathbf{x}_j'\boldsymbol{\beta}$, and ε_j are errors.
- Then j is chosen if it has the highest utility

$$\begin{aligned} \Pr[y = j] & = \Pr[U_j \geq U_k, \text{ all } k \neq j] \\ & = \Pr[\varepsilon_k - \varepsilon_j \leq -(V_k - V_j), \text{ all } k \neq j] \end{aligned}$$

- Different error distributions lead to different multinomial models.

Examples of unordered Models

- **1.** Multinomial logit and conditional logit:
  - errors ε_j are i.i.d. type I extreme value.
- **2.** Nested logit
  - ε_j are correlated type I extreme value.
- **3.** Random parameters logit:
  - ε_j are i.i.d. type I extreme value
  - but additionally parameters $\boldsymbol{\beta}_i$ are multivariate normal
  - no analytical solution for p_{ij}.
- **4.** Multinomial probit:
  - ε_j are correlated multivariate normal
  - no analytical solution for p_{ij}.

- Model 1: multinomial logit, conditional logit
  - attraction is that tractable (easy to estimate) but too limited
  - independence of irrelevant alternatives
    - $\Pr[y_{ik} = 1 | y_{ik} = 1 \text{ or } y_{ij} = 1]$ depends only on alternatives j and k
    - assumes ε_{ij} independent of ε_{ik}
    - red bus - blue bus problem.
- Model 2: nested logit
  - richer and still easy but requires specifying error correlation structure
  - two versions - only one consistent with ARUM
- Model 3: random parameters logit
  - currently very popular (use simulated ML or Bayesian)
- Model 4: multinomial probit
  - potentially rich but hard to estimate and fits poorly.

Ordered multinomial models

- For outcomes for which there is a natural ordering
  - e.g. y^* is a person's health status.
    We observe poor or fair ($y = 1$), good ($y = 2$) or excellent ($y_i = 3$).
- Model is based on a single latent variable $y^* = \mathbf{x}'\boldsymbol{\beta} + u$.
- Multinomial outcomes depend on magnitude of y^*. For 3 outcomes:

$$y_i = \begin{cases} 1 & \text{if } y^* \leq \alpha_1 \\ 2 & \text{if } \alpha_1 < y^* \leq \alpha_2 \\ 3 & \text{if } y^* > \alpha_2. \end{cases}$$

- Ordered probit model specifies $u \sim \mathcal{N}[0, 1]$. Then

$$\begin{aligned} p_1 &= \Pr[y^* \leq \alpha_1] = \Pr[\mathbf{x}'\boldsymbol{\beta} + u \leq \alpha_1] = \Phi(\alpha_1 - \mathbf{x}_i'\boldsymbol{\beta}) \\ p_2 &= \Pr[\alpha_1 < \mathbf{x}'\boldsymbol{\beta} + u \leq \alpha_2] = \Phi(\alpha_2 - \mathbf{x}'\boldsymbol{\beta}) - \Phi(\alpha_1 - \mathbf{x}_i'\boldsymbol{\beta}) \\ p_3 &= 1 - p_1 - p_2. \end{aligned}$$

  - ML estimation is straightforward.
  - Ordered logit model specifies $u \sim$ logistic: replace $\Phi(\cdot)$ above by $\Lambda(\cdot)$.

Stata commands

- Stata commands

Command	Model
mlogit	multinomial logit
asclogit	conditional logit
clogit	older command for conditional logit
nlogit	nested logit (ARUM version)
mprobit	multinomial probit
asmprobit	multinomial probit
mixlogit	random parameters logit (Stata add-on)

- Commands `mlogit` and `mprobit` for individual-specific regressors only
  - data in wide form (one obs is all alternatives for individual)
- Other commands allow individual-varying regressors (e.g. price)
  - data in long form (one obs is one alternative for individual)
  - commands `reshape` to move from wide to long form.

4. Censored data: Tobit

- Problem: with censored or truncated data:
  - The incomplete sample is not representative of the population. Instead, sample is selected on basis of y (vs. selection on $\mathbf{x}$ is okay).
  - Simple estimators are inconsistent and get wrong marginal effects. So need alternative estimators. These require strong assumptions.
- Censored Data: For part of the range of y we observe only that y is in that range, rather than observing the exact value of y.
  - e.g. Annual income top-coded at \$75,000 (censored from above).
  - e.g. Expenditures or hours worked bunched at 0 (censored from below).
- Truncated data: For part of range of y we do not observe y at all.
  - e.g. Sample excludes those with annual income $>$ \$75,000 per year.
  - e.g. Those with expenditures of \$0 are not observed.

Tobit Model Definition

- Latent dependent variable y^* follows regular linear regression

$$\begin{aligned} y^* &= \mathbf{x}'\boldsymbol{\beta} + \varepsilon \\ \varepsilon &\sim \mathcal{N}[0, \sigma^2] \end{aligned}$$

  - But this latent variable is only partially observed.
- Censored regression (from below at 0): we observe

$$y = \begin{cases} y^* & \text{if } y^* > 0 \\ 0 & \text{if } y^* \leq 0. \end{cases}$$

- Truncated regression (from below at 0): we observe only

$$y = y^* \quad \text{if } y^* > 0.$$

- In either case can estimate by MLE (skip this)
  - very fragile: e.g. inconsistent if ε is nonnormal or is heteroskedastic.
- We focus on conditional means, for intuition and later work.

Tobit example with Simulated Data

- Specify a linear relationship between
  - y : annual hours worked, and
  - x : log hourly wage.
- Desired hours of work, y^*, generated by model

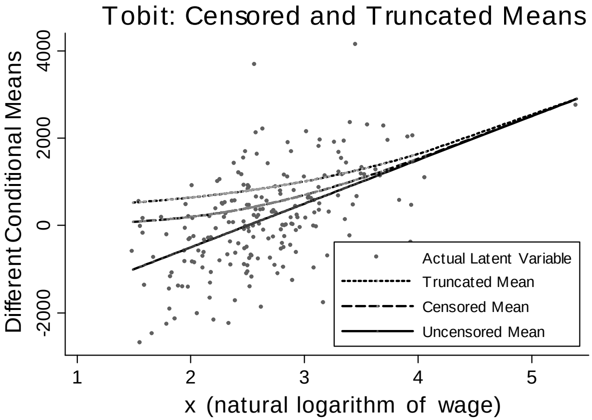
$$\begin{aligned} y_i^* &= -2500 + 1000x_i + \varepsilon_i, \quad i = 1, ..., 250, \\ \varepsilon_i &\sim \mathcal{N}[0, 1000^2], \\ x_i &\sim \mathcal{N}[2.75, 0.6^2] \; (\Rightarrow w_i \sim [18.73, 12.32^2]). \end{aligned}$$

- Tobit model: Instead of observing y^* we observe y where

$$y_i = \begin{cases} y_i^* & \text{if } y_i^* > 0 \\ 0 & \text{if } y_i^* \leq 0. \end{cases}$$

  - Here if desired hours are negative people do not work and $y = 0$.

- Scatterplot & true regression curves (derived later) for three samples:
  - truncated (top), censored (middle) and completely observed (bottom).



- Censored and truncated data the model is now nonlinear
  - and linear model will be flatter line than true line $(\widehat{\beta} \simeq 0.5\beta)$.

Truncated Mean in Tobit model

- Truncated mean: We observe y only when $y > 0$.
- The truncated conditional mean (suppressing conditioning on $\mathbf{x}$) is

$$
\begin{array}{ll}
\mathsf{E}[y|y>0] & \\
= \mathsf{E}\left[\mathbf{x}'\boldsymbol{\beta}+\varepsilon|\mathbf{x}'\boldsymbol{\beta}+\varepsilon>0\right] & \text{as } y=\mathbf{x}'\boldsymbol{\beta}+\varepsilon \\
= \mathbf{x}'\boldsymbol{\beta}+\mathsf{E}\left[\varepsilon|\varepsilon>-\mathbf{x}'\boldsymbol{\beta}\right] & \text{as } \mathbf{x} \text{ and } \varepsilon \text{ independent} \\
= \mathbf{x}'\boldsymbol{\beta}+\sigma\mathsf{E}\left[\frac{\varepsilon}{\sigma}|\frac{\varepsilon}{\sigma}>\frac{-\mathbf{x}'\boldsymbol{\beta}}{\sigma}\right] & \text{transform to } \varepsilon/\sigma\sim\mathcal{N}[0,1] \\
= \mathbf{x}'\boldsymbol{\beta}+\sigma\lambda\left(\frac{\mathbf{x}'\boldsymbol{\beta}}{\sigma}\right) & \text{using next slide: key result for } \mathcal{N}[0,1].
\end{array}
$$

  - where $\lambda(z)=\phi(z)/\Phi(z)$ is called the inverse Mills ratio.
- The regression function is not just $\mathbf{x}'\boldsymbol{\beta}$ (and is nonlinear).
  - OLS of y on $\mathbf{x}$ is inconsistent for $\boldsymbol{\beta}$
  - Need NLS or MLE for consistent estimates.

- Derivation: Truncated mean $\mathrm{E}[z|z>c]$ for the standard normal
  - key result used in the previous slide
  - consider $z \sim \mathcal{N}[0,1]$, with density $\phi(z)$ and c.d.f. $\Phi(z)$.
  - conditional density of $z|z>c$ is $\phi(z)/(1-\Phi(c))$.
  - truncated conditional mean is

$$
\begin{aligned}
\mathrm{E}[z|z>c] &= \int_c^\infty z\,(\phi(z)/(1-\Phi(c)))\;dz \\
&= \int_c^\infty z\frac{1}{\sqrt{2\pi}}\exp(-\tfrac{1}{2}z^2)\;dz\Big/\,(1-\Phi(c)) \\
&= \left[-\frac{1}{\sqrt{2\pi}}\exp(-\tfrac{1}{2}z^2)\right]_c^\infty\Big/\,(1-\Phi(c)) \\
&= \frac{\phi(c)}{1-\Phi(c)} \\
&= \frac{\phi(-c)}{\Phi(-c)} \\
&= \lambda(-c),\;\text{ where } \lambda(c)=\phi(c)/\Phi(c).
\end{aligned}
$$

Tobit Model: Censored Mean

- Censored mean: We observe $y = 0$ if $y^* < 0$ and $y = y^*$otherwise.
- The censored conditional mean (suppressing conditioning on $\mathbf{x}$) is

$$
\begin{aligned}
\mathrm{E}[y] &= \mathrm{E}_{y^*}[\mathrm{E}[y|y^*]] \\
&= \Pr[y^* \leq 0] \times 0 + \Pr[y^* > 0] \times \mathrm{E}[y^*|y^* > 0] \\
&= \Phi(\mathbf{x}'\boldsymbol{\beta}/\sigma) \left\{ \mathbf{x}'\boldsymbol{\beta} + \sigma \frac{\phi(\mathbf{x}'\boldsymbol{\beta}/\sigma)}{\Phi(\mathbf{x}'\boldsymbol{\beta}/\sigma)} \right\} \\
\mathrm{E}[y|\mathbf{x}] &= \Phi(\mathbf{x}'\boldsymbol{\beta}/\sigma)\mathbf{x}'\boldsymbol{\beta} + \sigma\phi(\mathbf{x}'\boldsymbol{\beta}/\sigma),
\end{aligned}
$$

using earlier result for the truncated mean $\mathrm{E}[y^*|y^* > 0]$.

- This conditional mean is again nonlinear.
  - OLS of y on $\mathbf{x}$ is inconsistent for $\boldsymbol{\beta}$
  - Need NLS or MLE for consistent estimates.

Tobit MLE: Data Example

- Data from 2001 Medical Expenditure Survey (MUS chapter 16).
  - `ambexp` (ambulatory expenditure = physician and hospital outpatient).
  - `dambexp` (`=1 if ambexp`>`0` and `=0 if ambexp=0`).
  - Regressors: age (in tens of years), `female`, `educ` (years of completed schooling), `blhisp` (=1 if black or hispanic) , `totchr` (number of chronic conditions), and `ins` (=1 if PPO or HMO health insurance).

Variable	Obs	Mean	Std. Dev.	Min	Max
ambexp	3328	1386.519	2530.406	0	49960
dambexp	3328	.8419471	.3648454	0	1
age	3328	4.056881	1.121212	2.1	6.4
female	3328	.5084135	.5000043	0	1
educ	3328	13.40565	2.574199	0	17
blhisp	3328	.3085938	.4619824	0	1
totchr	3328	.4831731	.7720426	0	5
ins	3328	.3650841	.4815261	0	1

- 16% of sample are censored (since `dambexp` has mean 0.84).

- Stata command `tobit, ll(0)` yields

```
. * Tobit on censored data
. tobit ambexp age female educ blhisp totchr ins, ll(0)

Tobit regression                                  Number of obs   =       3328
                                                  LR chi2(6)      =     694.07
                                                  Prob > chi2     =     0.0000
Log likelihood = -26359.424                       Pseudo R2       =     0.0130

------------------------------------------------------------------------------
      ambexp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   314.1479   42.63358     7.37   0.000     230.5572    397.7387
      female |   684.9918   92.85445     7.38   0.000     502.9341    867.0495
        educ |    70.8656   18.57361     3.82   0.000     34.44873    107.2825
      blhisp |   -530.311   104.2667    -5.09   0.000    -734.7443   -325.8776
      totchr |   1244.578   60.51364    20.57   0.000      1125.93    1363.226
         ins |  -167.4714   96.46068    -1.74   0.083    -356.5998    21.65696
       _cons |  -1882.591   317.4299    -5.93   0.000    -2504.969   -1260.214
-------------+----------------------------------------------------------------
      /sigma |   2575.907   34.79296                      2507.689    2644.125
------------------------------------------------------------------------------
  Obs. summary:        526  left-censored observations at ambexp<=0
                      2802     uncensored observations
                         0 right-censored observations
```

- Question: How do we interpret the coefficients?
  - Uncensored mean: $\partial \mathrm{E}[y^*|\mathbf{x}]/\partial x_j = \beta_j$
  - Censored mean: $\partial \mathrm{E}[y|\mathbf{x}]/\partial x_j = \Phi(\mathbf{x}'\boldsymbol{\alpha})\beta_j$ after some algebra

- The Tobit model is vary fragile
  - MLE is inconsistent if errors are nonnormal and even if they are normal but heteroskedastic.
  - This has led to semiparametric estimators.
- In particular censored least absolute deviations (CLAD) estimator
  - Basic idea is that censoring and truncation effect the mean, but not the median (if less than 50% censored)
  - LAD is the regression analog of the median estimate
  - Censored LAD can work well particularly for top coded data.
- Also when there is censoring from below at zero, the process for zeroes can differ from that for nonzeroes.
  - We consider this next.

5. Sample Selection Model: Overview

- There are many generalizations of standard Tobit, often involving sample selection or self-selection.
- We consider the most common, Heckman's sample selection model
    - Also called type 2 Tobit, Tobit with stochastic threshold, Tobit with probit selection.
    - For censoring below this is often more realistic than standard Tobit, as it allows different equations for participation and the outcome.

Sample Selection Model: Definition

- Define two latent variables as follows:

$$\begin{array}{ll} \text{Participation:} & y_1^* = \mathbf{x}_1'\boldsymbol{\beta}_1 + \varepsilon_1 \\ \text{Outcome:} & y_2^* = \mathbf{x}_2'\boldsymbol{\beta}_2 + \varepsilon_2 \end{array}$$

- Neither y_1^* nor y_2^* are completely observed.
  - Participation: We observe whether y_1^* is positive or negative

$$y_1 = \begin{cases} 1 & \text{if } y_1^* > 0 \\ 0 & \text{if } y_1^* \leq 0. \end{cases}$$

  - Outcome: Only positive values of y_2^* are observed

$$y_2 = \begin{cases} y_2^* & \text{if } y_1^* > 0 \\ 0 & \text{if } y_1^* \leq 0. \end{cases}$$

- MLE is used if error terms are specified to be joint normal
  - $(\varepsilon_1, \varepsilon_2) \sim \mathcal{N}\left[(0,0),\ (\sigma_1^2 = 1, \sigma_{12}, \sigma_2^2)\right]$
  - Fragile: e.g. inconsistent if ε is nonnormal or is heteroskedastic.

Sample Selection Model: Heckman 2-step estimator

- Assume instead that errors $(\varepsilon_1, \varepsilon_2)$ satisfy

$$\varepsilon_2 = \delta \times \varepsilon_1 + v,$$

where $\varepsilon_1 \sim \mathcal{N}[0,1]$ and v is independent of ε_1.

  - This is implied by $(\varepsilon_1, \varepsilon_2)$ joint normal.
  - But it is a weaker assumption.

- Then $y_2 = \mathbf{x}_2'\boldsymbol{\beta}_2 + \varepsilon_2$ if $y_1^* > 0$ implies

$$\begin{aligned} \mathrm{E}[y_2|y_1^* > 0] &= \mathbf{x}_2'\boldsymbol{\beta}_2 + \mathrm{E}[\varepsilon_2|\mathbf{x}_1'\boldsymbol{\beta}_1 + \varepsilon_1 > 0] \\ &= \mathbf{x}_2'\boldsymbol{\beta}_2 + \mathrm{E}\left[(\delta \times \varepsilon_1 + v)|\varepsilon_1 > -\mathbf{x}_1'\boldsymbol{\beta}_1\right] \\ &= \mathbf{x}_2'\boldsymbol{\beta}_2 + \delta \times \mathrm{E}[\varepsilon_1|\varepsilon_1 > -\mathbf{x}_1'\boldsymbol{\beta}_1] \\ &= \mathbf{x}_2'\boldsymbol{\beta}_2 + \delta \times \lambda(\mathbf{x}_1'\boldsymbol{\beta}_1) \end{aligned}$$

where third equality uses v independent of ε_1 and $\lambda(c) = \phi(c)/\Phi(c)$ is the inverse Mills ratio.

- For the observed outcomes:

$$E[y_2|y_1^* > 0] = \mathbf{x}_2'\boldsymbol{\beta}_2 + \delta\lambda(\mathbf{x}_1'\boldsymbol{\beta}_1).$$

  - OLS of y_2 on $\mathbf{x}_2$ only is inconsistent as regressor $\lambda(\mathbf{x}_1'\boldsymbol{\beta}_1)$ is omitted.
  - Heckman included an estimate of $\lambda(\mathbf{x}_1'\boldsymbol{\beta}_1)$ as an additional regressor.

- Heckman's two-step procedure:
  - **1.** Estimate $\boldsymbol{\beta}_1$ by probit for $y_1^* > 0$ or $y_1^* < 0$ with regressors $\mathbf{x}_{1i}$.
  - Calculate $\widehat{\lambda}_i = \lambda(\mathbf{x}_{1i}'\widehat{\boldsymbol{\beta}}_1) = \phi(\mathbf{x}_{1i}'\widehat{\boldsymbol{\beta}}_1)/\Phi(\mathbf{x}_{1i}'\widehat{\boldsymbol{\beta}}_1)$.
  - **2.** For observed y_2 estimate $\boldsymbol{\beta}_2$ and σ in the OLS regression

$$y_{2i} = \mathbf{x}_{2i}'\boldsymbol{\beta}_2 + \delta\widehat{\lambda}_i + w_i.$$

  - Need standard errors that correct for w_i heteroskedastic and $\widehat{\lambda}_i$ estimated. Stata command `heckman` does this.

- Exclusion restriction:
  - desirable to include some regressors in participation equation $(\mathbf{x}_1)$ that can be excluded from the outcome equation $(\mathbf{x}_2)$
  - otherwise identification solely from nonlinearity.
- Selection on observables only
  - If $\text{Cov}[\varepsilon_1, \varepsilon_2] = 0$ model then there is no longer selection on unobservables
  - Model reduces to a two-part model
    - ★ Probit for whether $y > 0$
    - ★ Regular OLS for the positives.
    - ★ Can be reasonable for individual's hospital expenditure data.
- Logs for the outcome
  - Often the outcome is expenditure
  - Then better to use a log model for the outcome
  - But will then need to transform to levels for prediction.

Heckman 2-step: Data Example

- 2-step where outcome is for $\ln y$.

```
.
. * Heckman 2-step without exclusion restrictions
. heckman lny $xlist, select(dy = $xlist) twostep

Heckman selection model -- two-step estimates   Number of obs      =      3328
(regression model with sample selection)        Censored obs       =       526
                                                Uncensored obs     =      2802

                                                Wald chi2(6)       =    189.46
                                                Prob > chi2        =    0.0000

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
lny          |
         age |    .202124   .0242974     8.32   0.000     .1545019    .2497462
      female |   .2891575    .073694     3.92   0.000     .1447199    .4335951
        educ |   .0119928   .0116839     1.03   0.305    -.0109072    .0348928
      blhisp |  -.1810582   .0658522    -2.75   0.006    -.3101261   -.0519904
      totchr |   .4983315   .0494699    10.07   0.000     .4013724    .5952907
         ins |  -.0474019   .0531541    -0.89   0.373     -.151582    .0567782
       _cons |   5.302572   .2941363    18.03   0.000     4.726076    5.879069
-------------+----------------------------------------------------------------
dy           |
         age |    .097315   .0270155     3.60   0.000     .0443656    .1502645
      female |   .6442089   .0601499    10.71   0.000     .5263172    .7621006
        educ |   .0701674   .0113435     6.19   0.000     .0479345    .0924003
      blhisp |  -.3744867   .0617541    -6.06   0.000    -.4955224   -.2534509
      totchr |   .7935208   .0711156    11.16   0.000     .6541367    .9329048
         ins |   .1812415   .0625916     2.90   0.004     .0585642    .3039187
       _cons |  -.7177087   .1924667    -3.73   0.000    -1.094937   -.3404809
-------------+----------------------------------------------------------------
mills        |
      lambda |  -.4801696   .2906565    -1.65   0.099    -1.049846    .0895067
-------------+----------------------------------------------------------------
         rho |   -0.37130
       sigma |  1.2932083
      lambda |  -.4801696   .2906565
------------------------------------------------------------------------------
```

Stata commands

- Stata commands

Command	Model
`tobit`	Tobit MLE (censored)
`truncreg`	Tobit MLE (truncated)
`cnreg`	Tobit (varying known threshold)
`intreg`	Interval normal data (e.g. \$1-\$100, \$101-\$200,..)
`heckman, mle`	Sample selection MLE
`heckman, 2step`	Sample selection two step

6. Treatment effects models

- What is the effect of a binary treatment?
- Outcome y (e.g. earnings) depends on whether or not get treatment d (e.g. training).
- Model

$$\begin{array}{ll} \text{Treatment} & d_i = 0 \text{ or } d_i = 1 \\ \text{Outcome} & y_i = \begin{cases} y_{1i} & \text{if } y_i = 1 \\ y_{0i} & \text{if } y_i = 1 \end{cases} \end{array}$$

- Problem: We want treatment effect $y_{1i} - y_{0i}$.
  - But we observe only one of y_{1i} and y_{0i}.
  - And people self-select into training
    - not randomized like an experiment.
- Solutions: many. Key distinction between
  - selection on observables only (just $x's$)
  - selection on observables and unobservables ($x's$ and $\varepsilon's$)

Selection on observables only

- A. Naive: Compare means
  - use $\bar{y}_1 - \bar{y}_0$
  - same as $\widehat{\alpha}$ in OLS of $y_i = \alpha d_i + u_i$
  - consistent if $\text{Cov}(u_i, d_i) = 0$
  - method for a randomized experiment, otherwise likely invalid.
- B. Control function
  - add $x_i's$ to control for d_i being chosen
  - use $\widehat{\alpha}$ in OLS of $y_i = \alpha d_i + \mathbf{x}_i'\boldsymbol{\beta} + u_i$
  - consistent if $\text{Cov}(u_i, d_i|\mathbf{x}_i) = 0$
- C. Propensity score matching
  - propensity score $p = \Pr[\text{treated}|\mathbf{x}] = \Pr[d = 1|\mathbf{x}]$
  - calculate using a very flexible logit model (interactions ...)
  - compare $y_1's$ (treated) with $y_0's$ (untreated) for those with similar p.
  - practical variation of matching those with similar $\mathbf{x}'s$.
- D. Sharp regression discontinuity design
  - suppose $y_i = f(s_i) + \alpha d_i + \mathbf{x}_i'\boldsymbol{\beta} + u_i$ and $d_i = \mathbf{1}(s_i > s_i^*)$.
  - compare y_i for those with s_i either side of threshold s_i^*

Selection on observables and unobservables

- A. Panel data
  - $y_{it} = \alpha d_{it} + \mathbf{x}_{it}'\boldsymbol{\beta} + v_i + \varepsilon_{it}$
  - first difference (or mean difference) gets rid of v_i
    - ★ OLS on $\Delta y_{it} = \alpha \Delta d_{it} + \Delta \mathbf{x}_{it}'\boldsymbol{\beta} + \Delta \varepsilon_{it}$
  - consistent if $\text{Cov}(\varepsilon_{it}, d_{it}|\mathbf{x}_{it}) = 0$ but allows $\text{Cov}(v_i, d_{it}|\mathbf{x}_{it}) \neq 0$
    - ★ okay if treatment correlated only with time invariant part of the error
- B. Difference in differences
  - variation of preceding that does not require panel data.
  - suppose treatment occurs only in second time period (not in first)
    - ★ use $\widehat{\alpha} = \Delta \overline{y}_{\text{treated}} - \Delta \overline{y}_{\text{untreated}} = (y_{1,\text{tr}} - y_{0,\text{tr}}) - (y_{1,\text{untr}} - y_{0,\text{untr}})$.
    - ★ more generally OLS on $\Delta y_i = \alpha d_i + \Delta \mathbf{x}_i'\boldsymbol{\beta} + u_i$
    - ★ requires common time trend for treated and untreated groups
  - Extends to more time periods (model in level with d_{it})
  - Extend to contrasts other than in time e.g. male/female
  - Extension is event history analysis.

- C. Instrumental variables
  - IV estimation with instrument $\mathbf{z}_i$ in $y_i = \alpha d_i + \mathbf{x}_i'\boldsymbol{\beta} + u_i$
  - consistent if $\text{Cov}(u_i, d_i|\mathbf{x}_i) = 0$
- D. Fuzzy regression discontinuity design
  - in fuzzy design not everyone with $s_i > s_i^*$ gets the treatment.
  - this introduces a role for unobservables.
- E. Parametric model e,g, Roy model:
  - introduce latent variables $d_i^*, y_{1i}^*, y_{0i}^*$ for d_i, y_{1i}, y_{0i}.
  - then $\text{E}[y_{1i}] = \text{E}[y_{1i}^*|d_i = 1] = \text{E}[y_{1i}^*|d_i^* > 0]$
    $= \text{E}[\mathbf{x}_{1i}'\boldsymbol{\beta} + \varepsilon_{1i}|\mathbf{z}_i'\gamma + v_i > 0] = \mathbf{x}_{1i}'\boldsymbol{\beta} + \text{E}[\varepsilon_{1i}|v_i > -\mathbf{z}_i'\gamma]$
  - so $\text{E}[y_{1i}] = \mathbf{x}_{1i}'\boldsymbol{\beta} + \delta_1\lambda(\mathbf{z}_i'\gamma)$ where $\lambda(\cdot)$ is inverse Mills ratio
    if $\varepsilon_{1i} = \delta_1 v_i + \xi_i > 0$, $v_i \sim \mathcal{N}[0,1]$, ξ_i independent.
- F. LATE (local average treatment effects)
  - allows α to vary with i and applies to many estimators.
  - for example consider IV interpreted as local effect
    - ★ e.g. in earnings-education regression with instrument law change that increased school leaving age, the earnings effect is for those with low levels of education.