1A: Binary outcomes: Basics

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

.

OeNB Summer School 2010
Microeconometrics
Oesterreichische Nationalbank (OeNB), Vienna, Austria

Based on
A. Colin Cameron and Pravin K. Trivedi,
Microeconometrics: Methods and Applications (MMA), ch.14
Microeconometrics using Stata (MUS), ch.14.
Data examples are from MUS.

Aug 30 - Sept 3, 2010

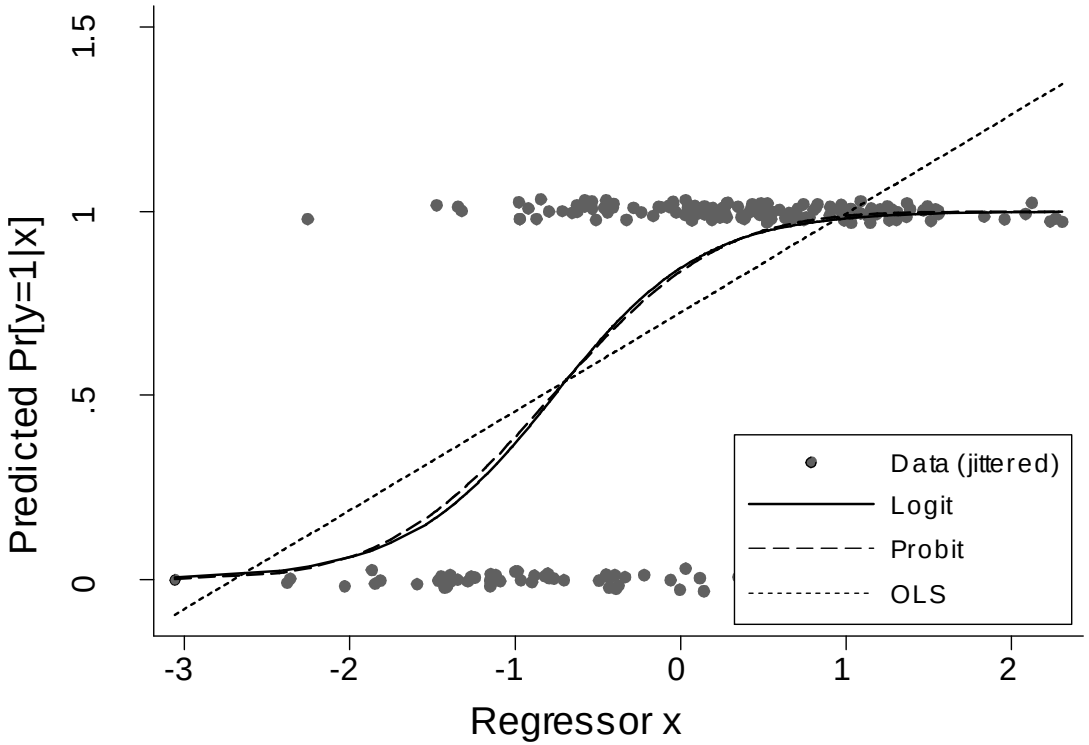
1. Introduction

- Discrete outcome or qualitative response models:
  y takes only a finite number of discrete values (categorical data).
  - Binary outcome models: only two possible outcomes.
    - ★ Without loss of generality we let these values be 1 and 0.
    - ★ We model $\Pr[y = 1|\mathbf{x}]$ using logit and probit models.
- Binary logit and probit models are nonlinear models
  - We illustrate the complications that arise with a nonlinear model.
- Other limited dependent variable models are
  - Multinomial outcome models: m possible outcomes.
    - ★ We model $\Pr[y = j|\mathbf{x}]$ for $j = 1, ..., m$.
  - Censored and truncated models (Tobit)
    - ★ Considerably more difficult conceptually.
    - ★ Sample is not reflective of the population (selection on y)
    - ★ Standard methods rely on strong distributional assumptions.

Outline

2. Binary Data: Examples

- First: a single regressor example allows a nice plot.
- Compare predictions of $\Pr[y = 1|x]$ from logit, probit and OLS.
  - Generated data followed by Stata command `logit y x`
  - Scatterplot of $y = 0$ or 1 on scalar regressor x (y is jittered).



- Logit $\simeq$ probit, while OLS predicts outside the $(0, 1)$ interval!

Data Example: Private health insurance [MUS ch.14.4]

- `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.

Example: Logit model

- Probability that $y_i = 1$ given regressors is specified to be

$$\Pr[y_i = 1|\mathbf{x}_i] = \Lambda(\mathbf{x}_i'\boldsymbol{\beta}) = \frac{\exp(\mathbf{x}_i'\boldsymbol{\beta})}{1 + \exp(\mathbf{x}'\boldsymbol{\beta})}.$$

  - Clearly $0 < \Pr[y_i = 1|\mathbf{x}_i] < 1$.
- MLE $\widehat{\boldsymbol{\beta}}$ is shown below to solve

$$\sum_{i=1}^{n}(y_i - \Lambda(\mathbf{x}_i'\boldsymbol{\beta}))\mathbf{x}_i = \mathbf{0}.$$

  - This is nonlinear in $\boldsymbol{\beta}$, so need to use iterative estimation procedure.
- Marginal effect of a change in the j^{th} regressor is

$$\text{ME}_j = \frac{\partial \Pr[y = 1|\mathbf{x}]}{\partial x_j} = \Lambda'(\mathbf{x}'\boldsymbol{\beta})\beta_j = \Lambda(\mathbf{x}'\boldsymbol{\beta})(1 - \Lambda(\mathbf{x}'\boldsymbol{\beta}))\beta_j$$

  - This varies with the evaluation point $\mathbf{x}$
  - This does not equal β_j, though $\text{sign}[\text{ME}_j] = \text{sign}[\beta_j]$.

- Stata command `logit` gives the logit MLE.

```
. * 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
------------------------------------------------------------------------------
```

- All except perhaps `hstatusg` have the expected sign.

- The average marginal effect $\text{AME}_j = \frac{1}{N}\sum_{i=1}^{N} \frac{\partial \Pr[y_i=1|\mathbf{x}_i]}{\partial x_j}$
  - In Stata 11 use command `margins, dydx(*)` after `logit`
  - In Stata 10 use add-on command `margeff` after `logit`.

```
. margeff

Average marginal effects on Prob(ins==1) after logit
```

ins	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interv
retire	.0426943	.0181787	2.35	0.019	.0070647	.0783
age	-.0031693	.0024486	-1.29	0.196	-.0079685	.0016
hstatusg	.0675283	.0196091	3.44	0.001	.0290951	.1059
hhincome	.0005002	.0001646	3.04	0.002	.0001777	.0008
educyear	.0248111	.0029706	8.35	0.000	.0189889	.0306
married	.1235562	.0191419	6.45	0.000	.0860388	.1610
hisp	-.1608825	.0339246	-4.74	0.000	-.2273735	-.0943

- Marginal effects: 0.043, -0.003, 0.067, 0.0005, 0.025, 0.124, -0.161 vs. 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.

3. Binary data: Estimation Theory

- For cross-section data
  - distribution for binary y is clearly Bernoulli (binomial with one trial)
  - maximum likelihood estimator (MLE) is clearly best estimator
  - it is fine to use default standard errors (robust is not needed).
- The main complications are
  - different models arise due to different specifications for $\Pr[y_i = 1|\mathbf{x}_i]$
  - interpretation of model estimates is complicated as nonlinear model
    - ★ emphasize marginal effects and parameter interpretation.

Estimation: iid case

- Begin with coin toss example of introductory statistics.
  - $y = 1$ denotes heads and $y = 0$ denotes tails.
  - p denotes the probability of a head $(y = 1)$ on one coin toss.
  - Then
$$\begin{aligned} \Pr[y = 1] &= p \\ \Pr[y = 0] &= 1 - p. \end{aligned}$$
  - The mean and variance are
$$\begin{aligned} \mathrm{E}[y] &= p \\ \mathrm{V}[y] &= p(1-p). \end{aligned}$$
- For N tosses y_i is the i^{th} of N independent realizations of head or tail.
  - The MLE for p is the sample mean $\overline{y}$,
    i.e. the proportion of tosses that are heads

Estimation: Binary regression models

- For economics examples p_i varies across individuals via regressors $\mathbf{x}_i$
  - e.g. work / no work
  - e.g. commute by car / bus.
- Specify model for the probability

$$\Pr[y_i = 1|\mathbf{x}_i] = p_i = F(\mathbf{x}_i'\boldsymbol{\beta}),$$

  where $0 \leq F(\cdot) \leq 1$ so that $0 \leq p \leq 1$.
- Single-index model
  - parameters $\boldsymbol{\beta}$ appear only via single index $\mathbf{x}'\boldsymbol{\beta}$ that is then transformed to be between 0 and 1.
- Choose $F(\cdot)$ to be a cumulative distribution function (c.d.f.).
  - Logit model uses logistic c.d.f.: $F(\cdot) = \Lambda(\cdot)$ with $\Lambda(z) = e^z/(1+e^z)$
  - Probit model uses standard normal c.d.f.: $F(\cdot) = \Phi(\cdot)$.

Estimation: Maximum Likelihood Estimation

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

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

- **Likelihood** is product of densities given independence over i:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{N} f(y_i|\mathbf{x}_i) = \prod_{i=1}^{N} p_i^{y_i}(1-p_i)^{1-y_i}$$

- MLE maximizes $L(\boldsymbol{\beta})$ which is equivalent to maximize $\ln L(\boldsymbol{\beta})$.
- **Log-likelihood function:**

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

- **MLE** $\widehat{\beta}$ maximizes $\ln L(\beta)$ if $\partial \ln L(\beta)/\partial \beta = \mathbf{0}$. Some algebra:

$$\begin{aligned}
\frac{\partial \ln L(\beta)}{\partial \beta} &= \sum_{i=1}^{N} \left\{ \frac{y_i}{F(\mathbf{x}_i'\beta)} F'(\mathbf{x}_i'\beta)\mathbf{x}_i - \frac{1-y_i}{1-F(\mathbf{x}_i'\beta)} F'(\mathbf{x}_i'\beta)\mathbf{x}_i \right\} \\
&= \sum_{i=1}^{n} \left\{ \left(\frac{y_i}{F(\mathbf{x}_i'\beta)} - \frac{1-y_i}{1-F(\mathbf{x}_i'\beta)} \right) F'(\mathbf{x}_i'\beta)\mathbf{x}_i \right\} \\
&= \sum_{i=1}^{N} \left\{ \left(\frac{y_i(1-F(\mathbf{x}_i'\beta)) + (1-y_i)F(\mathbf{x}_i'\beta)}{F(\mathbf{x}_i'\beta)1-F(\mathbf{x}_i'\beta)} \right) F'(\mathbf{x}_i'\beta)\mathbf{x}_i \right\} \\
&= \sum_{i=1}^{n} \frac{y_i - F(\mathbf{x}_i'\beta)}{F(\mathbf{x}_i'\beta)(1-F(\mathbf{x}_i'\beta))} F'(\mathbf{x}_i'\beta)\mathbf{x}_i
\end{aligned}$$

- Resulting **first-order conditions** (where $F'(z) = \partial F(z)/\partial z$).

$$\sum_{i=1}^{n} \frac{y_i - F(\mathbf{x}_i'\beta)}{F(\mathbf{x}_i'\beta)(1 - F(\mathbf{x}_i'\beta))} F'(\mathbf{x}_i'\beta)\mathbf{x}_i = \mathbf{0}.$$

- No explicit solution so use iterative **gradient methods** to compute $\widehat{\beta}$.

Consistency of MLE

- What are weakest conditions for consistency?
  - Analogy principle: $\widehat{\boldsymbol{\beta}}$ solving $\sum_{i=1}^{N} \mathbf{h}_i(\widehat{\boldsymbol{\beta}}) = \mathbf{0}$ is consistent for $\boldsymbol{\beta}$ if $\boldsymbol{\beta}$ solves the corresponding population moment condition $\mathrm{E}[\sum_{i=1}^{n} \mathbf{h}_i(\boldsymbol{\beta})] = \mathbf{0}$.
- The binary outcome model MLE solves

$$\sum_{i=1}^{N} (y_i - F(\mathbf{x}_i'\boldsymbol{\beta})) \frac{F'(\mathbf{x}_i'\boldsymbol{\beta})}{F(\mathbf{x}_i'\boldsymbol{\beta})(1 - F(\mathbf{x}_i'\boldsymbol{\beta}))} \mathbf{x}_i = \mathbf{0}.$$

  - So a necessary and sufficient condition for consistency is

$$\mathrm{E}[y_i|\mathbf{x}_i] = F(\mathbf{x}_i'\boldsymbol{\beta}).$$

  - Consistent given correct specification of $p_i = \mathrm{E}[y_i|\mathbf{x}_i] = \Pr[y_i = 1|\mathbf{x}_i]$.
- Qualitatively similar to OLS in linear model: need $\mathrm{E}[y_i|\mathbf{x}_i]$ correct.

Asymptotic distribution of MLE

- For correctly specified distribution the MLE

$$\widehat{\boldsymbol{\beta}}_{\text{ML}} \overset{a}{\sim} \mathcal{N}\left[\boldsymbol{\beta},\ \left(-\text{E}\left[\frac{\partial^2 \ln L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}}\right]\right)^{-1}\right]$$

- Specializing to binary outcome MLE

$$\widehat{\boldsymbol{\beta}}_{\text{ML}} \overset{a}{\sim} \mathcal{N}\left[\boldsymbol{\beta},\ \left(\sum_{i=1}^{N} \frac{(F_i')^2}{F_i(1-F_i)}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\right],\ F_i = F(\mathbf{x}_i'\boldsymbol{\beta}),\ F_i' = F'(\mathbf{x}_i'\boldsymbol{\beta}$$

- Default ML standard errors replace $F(\mathbf{x}_i'\boldsymbol{\beta})$ by $F(\mathbf{x}_i'\widehat{\boldsymbol{\beta}})$.
  - For independent cross-section data there is no need for robust se's
  - Reason: For binary data the distribution must be Bernoulli
    The only possibly misspecification is of $\Pr[y_i = 1|\mathbf{x}_i]$
    But then have more serious problem of inconsistency.

Statistical inference

- Consider test of $H_0 : \beta = 0$ against $H_a : \beta \neq 0$.
- Wald test:
  - $w = \widehat{\beta}/se(\widehat{\beta})$ and reject if $|w| > 1.96$
  - chisquared version rejects if $w^2 > \chi^2_{.05}(1) = 3.84$.
- Likelihood ratio test
  - $LR = -2 * [\ln(L_{rest}) - \ln(L_{unrest})]$ and reject if $LR > \chi^2_{.05}(1) = 3.84$.
- LM test or score test
  - used when H_0 model easier to estimate than H_a
  - used less here.
- All three are asymptotically equivalent
  - Wald is most often used.

4. Binary data: Logit, Probit and OLS

- **Logit model** to begin with:

$$p_i = \Pr[y_i = 1|\mathbf{x}_i] = \Lambda(\mathbf{x}_i'\boldsymbol{\beta}) = \frac{e^{\mathbf{x}_i'\boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i'\boldsymbol{\beta}}}.$$

  - $\Lambda(z) = e^z/(1+e^z) = 1/(1+e^{-z})$ is the logistic c.d.f.
  - The derivative $\Lambda'(z) = \Lambda(z)(1-\Lambda(z))$ is the logistic density.
  - For this reason also called **logistic regression** model.

- Logit ML first-order conditions simplify to

$$\sum_{i=1}^{n}(y_i - \Lambda(\mathbf{x}_i'\boldsymbol{\beta}))\mathbf{x}_i = \mathbf{0}.$$

  - Residual $y_i - \Lambda(\mathbf{x}_i'\boldsymbol{\beta})$ is orthogonal to regressors (like OLS).

- Logit estimates for private health insurance (repeats earlier)

```
. 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
------------------------------------------------------------------------------
```

Probit model

- **Probit model** specifies

$$p_i = \Pr[y_i = 1|\mathbf{x}_i] = \Phi(\mathbf{x}_i'\boldsymbol{\beta}).$$

  - $\Phi(z) = \int_{-\infty}^{z} \phi(s)ds$ is the standard normal.
  - The derivative $\Phi'(z) = \phi(z) = (1/\sqrt{2\pi})\exp(-z^2/2)$ is the standard normal density function.

- Probit ML first-order conditions do not simplify, unlike logit case

$$\sum_{i=1}^{N}(y_i - \Phi(\mathbf{x}_i'\boldsymbol{\beta}))\frac{\phi(\mathbf{x}_i'\boldsymbol{\beta})}{\Phi(\mathbf{x}_i'\boldsymbol{\beta})(1-\Phi(\mathbf{x}_i'\boldsymbol{\beta}))}\mathbf{x}_i = \mathbf{0}.$$

- Probit estimates for private health insurance

```
. 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 for binary data

- OLS regression of y_i on $\mathbf{x}_i$.
  - Then we are implicitly setting

$$p_i = \Pr[y_i = 1|\mathbf{x}_i] = \mathbf{x}_i'\boldsymbol{\beta},$$

    which has obvious weakness that $p < 0$ and $p > 1$ is possible.
  - called the linear probability model.
- Asymptotic distribution: use heteroskedastic robust standard errors

$$\widehat{\boldsymbol{\beta}}_{\text{OLS}} \overset{a}{\sim} \mathcal{N}\left[\boldsymbol{\beta}, (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\Sigma\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\right]$$

  - where for $\mathbf{X}'\widehat{\Sigma}\mathbf{X}$ use

$$\sum(y_i - \mathbf{x}_i'\widehat{\boldsymbol{\beta}})^2\mathbf{x}_i\mathbf{x}_i' \qquad \text{or} \qquad \sum\mathbf{x}_i'\widehat{\boldsymbol{\beta}}(1 - \mathbf{x}_i'\widehat{\boldsymbol{\beta}})\mathbf{x}_i\mathbf{x}_i'$$

  - Need this as the error in $y_i = \mathbf{x}_i'\boldsymbol{\beta} + u_i$ is intrinsically heteroskedastic, since $\mathrm{V}[y_i] = p_i(1 - p_i)$ for a Bernoulli random variable.

- OLS estimates for private health insurance

```
. 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
```

ins	Coef.	Robust 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 $\frac{1}{N}\sum_{i=1}^{N} F(\mathbf{x}_i'\widehat{\boldsymbol{\beta}})$ 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 $\widehat{p}_i > 1$.

5. Binary Data: Marginal effects

- Coefficients in different models are not directly comparable!
- Instead compare marginal effects across models

$$\Pr[y = 1|\mathbf{x}] = \mathrm{E}[y = 1|\mathbf{x}] = F(\mathbf{x}'\boldsymbol{\beta}).$$

  with different models having different $F(\cdot)$.
- Marginal effect: $\mathrm{ME}_j = \partial \Pr[y = 1|\mathbf{x}]/\partial x_j = \partial F(\mathbf{x}'\boldsymbol{\beta})/\partial x_j$ is

$$\begin{aligned} \mathrm{ME}_j &= F'(\mathbf{x}'\boldsymbol{\beta}) \times \beta_j && \text{for general } F(\cdot) \\ &= \begin{cases} \Lambda(\mathbf{x}'\boldsymbol{\beta})(1-\Lambda(\mathbf{x}'\boldsymbol{\beta}))\beta_j & \text{logit model} \\ \phi(\mathbf{x}'\boldsymbol{\beta})\beta_j & \text{probit model} \\ \beta_j & \text{OLS} \end{cases} \end{aligned}$$

- The marginal effect depends on
  - the functional form of F and
  - the evaluation point $\mathbf{x}$
  - the parameter $\boldsymbol{\beta}$.

Marginal effects: AME, MEM, and MER

- Consider three different marginal effects
  - **1.** AME: Average Marginal Effect for j^{th} regressor

$$\text{AME}_j = \frac{1}{N}\sum_{i=1}^{N}\text{ME}_j = \frac{1}{N}\sum_{i=1}^{N}F'(\mathbf{x}_i'\widehat{\boldsymbol{\beta}})\times\widehat{\beta}_j$$

    - ★ For population AME compute the sample-weighted AME.
  - **2.** MEM: Marginal Effect at mean value $\mathbf{x} = \overline{\mathbf{x}}$

$$\text{MEM}_j = \text{ME}_j(\mathbf{x}=\overline{\mathbf{x}}) = F'(\overline{\mathbf{x}}'\widehat{\boldsymbol{\beta}})\times\widehat{\beta}_j.$$

  - **3.** MER: Marginal Effect at representative value $\mathbf{x} = \mathbf{x}^*$

$$\text{MER}_j = \text{ME}_j(\mathbf{x}=\mathbf{x}^*) = F'(\mathbf{x}^{*\prime}\widehat{\boldsymbol{\beta}})\times\widehat{\beta}_j.$$

- These differ unless $F(\mathbf{x}'\boldsymbol{\beta}) = \mathbf{x}'\boldsymbol{\beta}$ (the linear model).

Marginal effects (continued)

- Stata 11: MEs computed using new post-estimation command `margins`
  - AME: `margins, dydx(*)`
  - MEM: `margins, dydx(*) atmean`
  - MER: `margins, dydx(*) at()`
- Stata 10: MEs computed using post-estimation commands `mfx` or `margeff`
  - AME: user-written command `margeff`
  - MEM: Stata command `mfx`
  - MER: Stata command `mfx, at()`
- These commands available after most Stata estimation commands
  - use `margins` if you have Stata 11

- AME compared to MEM for logit
  - Stata 11 use `margins, dydx(*)` and `margins, dydx(*) atmean`.

```
. * Marginal effects for logit: AME differs from MEM
. quietly logit ins retire age hstatusg hhincome educyear married hisp

. margeff

Average marginal effects on Prob(ins==1) after logit
```

ins	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
retire	.0426943	.0181787	2.35	0.019	.0070647	.0783239
age	-.0031693	.0024486	-1.29	0.196	-.0079685	.0016299
hstatusg	.0675283	.0196091	3.44	0.001	.0290951	.1059615
hhincome	.0005002	.0001646	3.04	0.002	.0001777	.0008228
educyear	.0248111	.0029706	8.35	0.000	.0189889	.0306334
married	.1235562	.0191419	6.45	0.000	.0860388	.1610736
hisp	-.1608825	.0339246	-4.74	0.000	-.2273735	-.0943914

```
. mfx

Marginal effects after logit
      y  = Pr(ins) (predict)
         =  .37283542
```

variable	dy/dx	Std. Err.	z	P>|z|	[95% C.I.]		X
retire*	.0457255	.0194	2.36	0.018	.007711	.08374	.624766
age	-.0034129	.00264	-1.29	0.196	-.008585	.001759	66.9139
hstatusg*	.0716613	.02057	3.48	0.000	.031346	.111977	.704616
hhincome	.0005386	.00018	3.02	0.003	.000189	.000888	45.2639
educyear	.0267179	.0033	8.09	0.000	.020245	.033191	11.8986
married*	.1295601	.01974	6.56	0.000	.090862	.168259	.733001
hisp*	-.1677028	.03418	-4.91	0.000	-.23469	-.100715	.072676

Marginal effects: Approximations for logit and probit

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

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

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

Marginal effects: Single-index models

- Single-index model: nonlinear model with

$$\mathrm{E}[y|\mathbf{x}] = \Pr[y = 1|\mathbf{x}] = F(\mathbf{x}'\boldsymbol{\beta}).$$

  - $\mathrm{E}[y|\mathbf{x}]$ is a transformation $F(\cdot)$ of a linear combination of the regressors.
  - Logit and probit are examples.

- All marginal effects are the same multiple of the relevant parameter:

$$\mathrm{ME}_j = \frac{\partial \mathrm{E}[y|\mathbf{x}]}{\partial x_j} = F'(\mathbf{x}'\boldsymbol{\beta})\beta_j.$$

  - **1.** Sign of β_j equals the sign of ME_j if $F(\cdot)$ is monotonic increasing.
  - **2.** If β_j is two times β_k then ME_j is two times ME_k.

$$\frac{\mathrm{ME}_j}{\mathrm{ME}_k} = \frac{F'(\mathbf{x}'\boldsymbol{\beta})\beta_j}{F'(\mathbf{x}'\boldsymbol{\beta})\beta_k} = \frac{\beta_j}{\beta_k}.$$

Marginal effects: Odds ratio interpretation for logit

- Odds ratio: $p/(1-p)$ measures the probability that $y = 1$ relative to the probability that $y = 0$.
  - E.g. $y = 1$ denotes survival and $y = 0$ denotes death.
  - Odds ratio $= 2$ means odds of survival are twice those of death.
- Logit model

$$p = \frac{\exp(\mathbf{x}'\boldsymbol{\beta})}{1+\exp(\mathbf{x}'\boldsymbol{\beta})} \quad \Rightarrow \quad \frac{p}{1-p} = \exp(\mathbf{x}'\boldsymbol{\beta})$$

  - Then $\partial(p/(1-p))/\partial x_j = \exp(\mathbf{x}'\boldsymbol{\beta}) \times \beta_j = (p/(1-p)) \times \beta_j$.
  - So $\widehat{\beta}_j = 0.1$ means a one unit change in x_j increases the odds ratio by a multiple 0.1.
  - More precisely the odds ratio is multiplied by $\exp(\widehat{\beta}_j)$.
    - ★ reason: If x_j increases by 1 then $p/(1-p) = \exp(\mathbf{x}'\boldsymbol{\beta} + 1 \times \beta_j) = \exp(\beta_j)\exp(\mathbf{x}'\boldsymbol{\beta})$.
  - Stata command `logistic` reports exponentiated coefficients.

6. Binary Data: Which binary choice model?

- Theoretically it depends on the unknown data generating process.
- Key choice is of $p_i = F(\mathbf{x}_i'\boldsymbol{\beta})$.
- Unlike other ML applications the distribution is determined solely by p_i, so this is only possible misspecification.
- If $F(\cdot)$ in $p_i = F(\mathbf{x}_i'\boldsymbol{\beta})$ is misspecified then MLE is inconsistent.
- But provided p_i is still of single-index form $p_i = F(\mathbf{x}_i'\boldsymbol{\beta})$, then choosing the wrong function F effects all slope parameters equally, and the ratio of slope parameters is constant across the models.

- Logit: binary model most often used by statisticians.
  - Logit generalizes simply to multinomial data ($>$ two outcomes).
- Probit: binary model most often used by economists.
  - Probit is motivated by a latent normal random variable.
  - Probit generalizes to Tobit models and multinomial probit.
- Empirically: logit or probit are similar
  - give similar predictions and marginal effects
  - greatest difference is in prediction of probabilities close to 0 or 1.
- Complementary log-odds model:
  - also a possibility when most outcomes are 0 or 1.
- OLS: can be useful for preliminary data analysis
  - for individual level prediction should use probit or logit
  - for computing average marginal effects Angrist and Pischke (2009) argue that OLS is okay.

7. Binary Data: Model Diagnostics

- Diagnostics for nonlinear model are not so clear cut.
  - There are several measures of model adequacy.
  - Many are very specific to binary outcome models.
  - There is no single best measure.
  - See Amemiya (1981) and Maddala (1983).
- Approaches detailed below:
  - **1.** R-squared measures
  - **2.** Compare $\widehat{y}$ with y.
  - **3.** Compare predicted $\widehat{\Pr}[y = 1]$ with actual $\Pr[y = 1]$.

Model Diagnostics: McFadden's R-Squared

- There are many R-squareds for binary models as R^2 in linear model has many interpretations.
- Best is a measure due to McFadden (1974)

$$R^2 = 1 - \frac{\ln L_{fit}}{\ln L_0},$$

  - $\ln L_{fit}$ = log-likelihood in the fitted model
  - $\ln L_0$ is the log-likelihood in the intercept-only model
- This R^2 should be only used for discrete choice models.
  - Aside: In other nonlinear models instead use

$$R^2_{\text{RG}} = 1 - \frac{\ln L_{\max} - \ln L_{fit}}{\ln L_{\max} - \ln L_0} = \frac{\ln L_{fit} - \ln L_0}{\ln L_{\max} - \ln L_0},$$

    where $\mathcal{L}_{\max}$ is the maximum possible value of the log-likelihood.
  - For binary outcome models $\ln L_{\max} = 0$, so R^2_{RG} = McFadden's R^2.
- For easy interpretation use $\text{Cor}[y, \widehat{p}]$ where $\widehat{p}_i = F(\mathbf{x}_i'\widehat{\boldsymbol{\beta}})$.

Model Diagnostics: Correct Prediction that y=1

- Many measures compare predicted $\widehat{y}$ with y.
- The problem is in defining a rule for when $\widehat{y} = 1$.
  - Obvious is $\widehat{y} = 1$ when $\widehat{p} = F(\mathbf{x}'\widehat{\boldsymbol{\beta}}) > 0.5$.
  - But this can yield $\widehat{y} = 1$ all the time if most of the sample has $y = 1$ (or $\widehat{y} = 0$ all the time if most of the sample has $y = 0$).
- The receiver operator characteristics (ROC) curve does this for different thresholds
  - for $0 \leq c \leq 1$ recompute $\widehat{y}_i(c) = 1$ when $\widehat{p}_i = F(\mathbf{x}_i'\widehat{\boldsymbol{\beta}}) > c$ and $\widehat{y}_i(c) = 0$ otherwise.
  - plot the fraction of true positives against false positives
  - also called plot sensitivity against (1 - specificity).
  - departures from a 45 degree line are preferred.

Model Diagnostics: Correct Prediction of Pr[y=1]

- Can compare predicted $\widehat{\Pr}[y=1]$ with y.
- Doing this on average is not helpful over the entire sample.
  - For logit:

$$\begin{aligned} & \textstyle\sum_{i=1}^{N}(y_i - \Lambda(\mathbf{x}_i'\widehat{\boldsymbol{\beta}}))\mathbf{x}_i = \mathbf{0} && \text{f.o.c. for MLE} \\ \Rightarrow \quad & \textstyle\sum_{i=1}^{N}(y_i - \Lambda(\mathbf{x}_i'\widehat{\boldsymbol{\beta}})) = 0 && \text{if regressors include intercept} \\ \Rightarrow \quad & \tfrac{1}{N}\textstyle\sum_{i=1}^{N}\widehat{p}_i = \bar{y} && \text{where } \widehat{p}_i = \Lambda(\mathbf{x}_i'\widehat{\boldsymbol{\beta}}). \end{aligned}$$

  - And probit in practice comes close to this.
- More useful for comparisons with subsamples or out of sample
  - Do a generalized chi-square goodness-of-fit test.
  - Stata post-estimation command `gof`.

8. Binary data: Index function model

- Index function model
  - gives a way to interpret the function $F(\cdot)$ in a binary model.
  - generalizes to ordered multinomial models and Tobit models.
- Specify a regression model for latent variable y^*

$$y^* = \mathbf{x}'\boldsymbol{\beta} + u.$$

- This cannot be estimated as y^* is not observed. Instead we observe

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

  - The choice of 0 as the threshold is a normalization.

- Examples:
  - y^* is tendency to work - we observe only whether or not work ($y = 1$)
  - y^* is a propensity to commute by public transit - we observe only whether or not the public transit is used ($y = 1$).

Index function model: resulting binary outcome

- Outcome probability: Suppressing conditioning on $\mathbf{x}$:

$$\begin{aligned} \Pr[y=1] &= \Pr[y^* > 0] \\ &= \Pr[\mathbf{x}'\boldsymbol{\beta} + u > 0] \\ &= \Pr[-u < \mathbf{x}'\boldsymbol{\beta}] \\ &= F(\mathbf{x}'\boldsymbol{\beta}), \end{aligned}$$

  - F is the c.d.f. of $-u$ (equals c.d.f. of u if density symmetric about 0).
- Probit model: Assume $u \sim \mathcal{N}[0,1]$. Then

$$\Pr[y=1] = \Phi(\mathbf{x}'\boldsymbol{\beta}).$$

- Logit model: Assume $u \sim$ logistic. Then

$$\Pr[y=1] = \Lambda(\mathbf{x}'\boldsymbol{\beta}) = \exp(\mathbf{x}'\boldsymbol{\beta})/[1+\exp(\mathbf{x}'\boldsymbol{\beta})].$$

- Probit simulation: $y^* = 1 + x + u$, $u \sim \mathcal{N}[0,1]$, $x \sim \mathcal{N}[0,1]$. And $y = 1$ if $y^* > 0$ and $y = 0$ otherwise. $N = 200$.

```
. * Index function model to generate probit
. clear

. quietly set obs 200

. quietly generate x = rnormal(0,1)

. quietly generate ystar = 1 + 1*x + rnormal(0,1)

. quietly generate y = ystar > 0

. summarize y x
```

Variable	Obs	Mean	Std. Dev.	Min	Max
y	200	.71	.4549007	0	1
x	200	-.1005735	1.029603	-2.830635	2.679533

```
. probit y x, nolog

Probit regression                                 Number of obs   =        200
                                                  LR chi2(1)      =      78.06
                                                  Prob > chi2     =     0.0000
Log likelihood = -81.398245                       Pseudo R2       =     0.3241
```

y	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
x	1.099966	.1538048	7.15	0.000	.7985144	1.401418
_cons	.9806923	.1403632	6.99	0.000	.7055855	1.255799

9. Binary Data: Additive Random Utility Model

- The additive random utility model (ARUM)
  - generalizes to unordered multivariate models.
- Consumer choice: consumer selects alternative with highest utility.
- Specify the utilities of alternatives 0 and 1 to be

$$U_0 = V_0 + \varepsilon_0$$
$$U_1 = V_1 + \varepsilon_1$$

  - V_0 and V_1 are deterministic components of utility. (The dependence on regressors is detailed below).
  - ε_0 and ε_1 are random components of utility.
- We observe $y = 1$ if $U_1 > U_0$ and $y = 0$ if $U_1 \leq U_0$.

ARUM: Binary outcome

- Outcome probability: Suppressing dependence on **x**

$$\begin{aligned} \Pr[y=1] &= \Pr[U_1 > U_0] \\ &= \Pr[V_1 + \varepsilon_1 > V_0 + \varepsilon_0] \\ &= \Pr[\varepsilon_0 - \varepsilon_1 < V_1 - V_0] \\ &= F(V_1 - V_0), \end{aligned}$$

  where F is the c.d.f. of $(\varepsilon_0 - \varepsilon_1)$.
- Binary probit: ε_0 and ε_1 are joint normal with $\mathrm{V}[\varepsilon_0 - \varepsilon_1] = 1$.
- Binary logit: ε_0 and ε_1 are type I extreme value distributed with $f(\varepsilon) = e^{-\varepsilon}\exp(-e^{-\varepsilon})$, as then $(\varepsilon_0 - \varepsilon_1)$ is logistic distributed.
- The random component ε in the utility model is needed.
  - Otherwise, choice is deterministic with e.g. alternative 1 always chosen if $V_1 > V_0$.

ARUM: Regressors

- Distinguish between two types of regressors
  - $\mathbf{z}_{ij}$ alternative-varying regressors e.g. price may vary over alternatives
  - $\mathbf{w}_i$ case-specific regressors (or alternative-invariant) e.g. race or gender.
- Then deterministic component of utility:

$$V_{ij} = \mathbf{z}_{ij}'\boldsymbol{\alpha} + \mathbf{w}_i'\gamma_j, \quad j = 0, 1,$$

  where coefficients
  - $\boldsymbol{\alpha}$ does not vary with alternative
  - γ_j does vary with the alternatives.
- Outcome probability:

$$\begin{aligned} \Pr[y_i = 1] &= F(V_{i1} - V_{i0}) \\ &= F((\mathbf{z}_{i1} - \mathbf{z}_{i0})'\boldsymbol{\alpha} + \mathbf{w}_i'(\gamma_1 - \gamma_0)). \end{aligned}$$

  - This is earlier model with $\mathbf{x}_i'\boldsymbol{\beta} = (\mathbf{z}_{i1} - \mathbf{z}_{i0})'\boldsymbol{\alpha} + \mathbf{w}_i'(\gamma_1 - \gamma_0)$.
  - Case-specific regressors: only difference $(\gamma_1 - \gamma_0)$ can be identified.

10. Some References

- The material is covered in graduate level texts including
  - CT(2005) MMA chapter 14 and CT(2009) MUS chapter 14
  - Wooldridge, J.M. (2002), *Econometric Analysis of Cross Section and Panel Data*, MIT Press.
  - Greene, W.H. (2007), *Econometric Analysis*, Prentice-Hall, Sixth edition.
- A classic book is
  - Maddala, G.S. (1986), *Limited-Dependent and Qualitative Variables in Econometrics*, Cambridge University Press.