AFRE 835: Introductory Econometrics

Chapter 17: Limited Dependent Variable Models

Spring 2017

Introduction

- In discussing the various models thus far, we have not put much emphasis on restrictions associated with the dependent variable itself,

  . . . with the exception of the Linear Probability Model used when the dependent variable is a binary one.
- There are many examples of so-called **Limited Dependent Variables (LDV)**, including
  1. binary variables (participation decisions, voting models, etc);
  2. multinomial choice models (housing or occupation choice, recreation site choice, etc.);
  3. categorical variables (e.g., educational achievement, bond ratings, etc.);
  4. count variables (e.g., number of children, number of arrests);
  5. corner solution models (e.g., quantity of beef, tofu and/or chicken to purchase);
  6. censored or truncated data (low income agents, class enrollment, etc.);
  7. sample selection (on-site surveys);
- We will focus our attention on items 1, 4, 5 and 6.

Outline

The Binary Response Model

- In chapter 7, we briefly discussed how to model binary responses using a **Linear Probability Model (LPM)**.
- We saw that, while the LPM can provide useful information on the average marginal impact of a variable, it had several limitations, including
  1. Fitted choice probabilities can lie outside the unit interval; and
  2. It assumes the marginal effect of each variable is constant;
- In this section, we consider alternative models that are nonlinear, but avoid the limitations of the LPM.
- As useful starting point is to note that, in the case of a binary dependent variable y,

$$\begin{aligned} E(y|\boldsymbol{x}) &= 1 \cdot Pr(y=1|\boldsymbol{x}) + 0 \cdot Pr(y=1|\boldsymbol{x}) \\ &= Pr(y=1|\boldsymbol{x}) \end{aligned} \tag{1}$$

- The key step is specifying a functional form for $Pr(y=1|\boldsymbol{x})$.

Linear Probability and Logit Models

- Various models correspond to choosing a functional form for $G(\cdot)$ in the *linear index model*:

$$Pr(y = 1|\boldsymbol{x}) = G(\boldsymbol{x\beta}) = G(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k). \tag{2}$$

- For the Linear Probability, $G(z) = z$, so that

$$Pr(y = 1|\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k. \tag{3}$$

- For the **logit model**, $G(z) = \Lambda(z) \equiv \frac{exp(z)}{1+exp(z)}$, so that

$$Pr(y = 1|\boldsymbol{x}) = \frac{exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}{1 + exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}. \tag{4}$$

  Notice that $Pr(y = 1|\boldsymbol{x}) = G(z) \in (0, 1)$, avoiding one of the LPM limitations.

The Probit Model

- For the **probit model**,

$$G(z) = \Phi(z) \equiv \int_{-\infty}^{z} \phi(v)dv, \tag{5}$$

  where $\Phi(z)$ is the standard normal cumulative distribution function (cdf) and

$$\phi(z) = \Phi'(z) = \frac{1}{\sqrt{2\pi}} exp(-z^2/2) \tag{6}$$

  denotes the standard normal probability density function (pdf).
- As with the logit model, the probit model, by using a cdf for $G(z)$, ensures that $G(z) \in (0, 1)$.
- Moreover, for both the logit and probit models,
  1. $G'(z) > 0$;
  2. $G(z) \to 0$ as $z \to -\infty$;
  3. $G(z) \to 1$ as $z \to \infty$;

The Latent Variable Model Foundation

- Suppose that our binary variable corresponds to

$$y = \begin{cases} 1 & \text{Option A is chosen} \\ 0 & \text{Option B is chosen} \end{cases} \tag{7}$$

- One way to derive the logit and probit model is to start with an underlying **latent** (or unobserved) variable y^*, such that

$$y^* = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + e \tag{8}$$

- Think of y^* as the net gain from choosing option A over option B.
  - y^* can, for example, represent $U(A) - U(B)$, where $U(j)$ denotes the utility the individual perceives for option j.
  - Alternative, y^* could represent $\pi(A) - \pi(B)$, where $\pi(j)$ denotes the individual's expected profits from choosing option j.

The Latent Variable Model Foundation (cont'd)

- The individual is assumed to choose the alternative that makes them best off; e.g., in McFadden's Random Utility Maximization (RUM) framework.
- We do not observe y^*, but instead observe only y, where
- Thus

$$\begin{aligned} y &= \begin{cases} 1 & y^* > 0 \\ 0 & y^* \leq 0 \end{cases} \\ &= 1[y^* > 0] \end{aligned} \tag{9}$$

- The probit model arises if $e \overset{iid}{\sim} \mathcal{N}(0,1)$, since then

$$\begin{aligned} P(y=1|\boldsymbol{x}) = P(y^* > 0|\boldsymbol{x}) &= P(\boldsymbol{x}\boldsymbol{\beta} + e > 0|\boldsymbol{x}) \\ &= P(-e < \boldsymbol{x}\boldsymbol{\beta}|\boldsymbol{x}) \\ &= \Phi(\boldsymbol{x}\boldsymbol{\beta}) \end{aligned} \tag{10}$$

since if $e \overset{iid}{\sim} \mathcal{N}(0,1)$, then $-e \overset{iid}{\sim} \mathcal{N}(0,1)$.

The Latent Variable Model Foundation (cont'd)

- In a similar fashion, the logit model arises if e follows a standard logistic distribution with cdf $\Lambda(z)$, since

$$\begin{aligned} P(y=1|\boldsymbol{x}) &= P(y^* > 0|\boldsymbol{x}) = P(\boldsymbol{x}\boldsymbol{\beta} + e > 0|\boldsymbol{x}) \\ &= P(-e < \boldsymbol{x}\boldsymbol{\beta}|\boldsymbol{x}) \\ &= \Lambda(\boldsymbol{x}\boldsymbol{\beta}) \end{aligned} \tag{11}$$

Again, this last step applies since if e has a cdf of $\Lambda(z)$, then, because it is symmetrically distributed around zero, the $-e$ has a cdf of $\Lambda(z)$.

Estimating the Logit and Probit Models

- Estimation of the Logit and Probit models is complicated by their nonlinear nature.
- We can write a nonlinear regression model of the form:

$$\begin{aligned} y &= E(y|\boldsymbol{x}) + [y - E(y|\boldsymbol{x})] \\ &= P(y=1|\boldsymbol{x}) + u \\ &= G(\boldsymbol{x}\boldsymbol{\beta}) + u \end{aligned} \tag{12}$$

- OLS and GLS procedures are not applicable.
- However, nonlinear Least Squares can be used, but rarely is.
- Instead **Maximum Likelihood Estimation (MLE)** is used instead (See Appendix 17A).

Maximum Likelihood Estimation (MLE)

- Intuitively, one can think of MLE as trying to find the set of model parameters that maximize the *ex ante* probability of observing the outcomes in your data.
- The probability of observing a given outcome for individual i depends upon $\boldsymbol{x}_i$ and $\boldsymbol{\beta}$, with the density (i.e., *likelihood*) given by:

$$f(y_i|\boldsymbol{x}_i,\boldsymbol{\beta}) = [G(\boldsymbol{x}_i\boldsymbol{\beta})]^{y_i}[1-G(\boldsymbol{x}_i\boldsymbol{\beta})]^{(1-y_i)} \tag{13}$$

- In our setting, with random draws, the joint likelihood function becomes the product of individual likelihood contributions

$$\prod_{i=1}^{n} f(y_i|\boldsymbol{x}_i,\boldsymbol{\beta}) = \prod_{i=1}^{n}[G(\boldsymbol{x}_i\boldsymbol{\beta})]^{y_i}[1-G(\boldsymbol{x}_i\boldsymbol{\beta})]^{(1-y_i)} \tag{14}$$

- MLE chooses the $\boldsymbol{\beta}$ to maximize this, or equivalently its log:

$$\mathcal{L} = \sum_{i=1}^{n}\mathcal{L}_i(\boldsymbol{x}_i,\boldsymbol{\beta}) = \sum_{i=1}^{n}\{y_i log\,[G(\boldsymbol{x}_i\boldsymbol{\beta})] + (1-y_i)log\,[1-G(\boldsymbol{x}_i\boldsymbol{\beta})]\} \tag{15}$$

Maximum Likelihood Estimation (MLE)

- Given that one has specified the model correction, under very general conditions, MLE is
  - Consistent
  - Asymptotically normal
  - Asymptotically efficient
- In Stata, MLE for both probit and logit models are provided, with syntax
  - *probit depvar [indepvars] [if] [in] [weight] [, options]*
  - *logit depvar [indepvars] [if] [in] [weight] [, options]*
- These routines provide, among other things,
  - point estimates
  - (asymptotic) standard errors
  - (asymptotic) t-statistics
  - the value of the log-likelihood function $\mathcal{L}$ evaluated at the estimated parameter value

Testing Multiple Hypothesis

- There are a number of ways to test joint hypotheses (e.g., a series of exclusion restrictions).
- The two most commonly used procedures are
  1. The **Wald Statistic**, which we discussed earlier
     ... This is implemented in Stata using the *test*command;
  2. The **Likelihood Ratio (LR) Test** is based on a comparison of the log-likelihood function $\mathcal{L}$ for the *unrestricted* ($\mathcal{L}_{ur}$) versus the *restricted* ($\mathcal{L}_r$) model, with

$$LR = 2(\mathcal{L}_{ur} - \mathcal{L}_r) \overset{a}{\sim} \chi^2_q \tag{16}$$

     where q denotes the number of restrictions imposed.
  3. Note that $LR \geq 0$ since $\mathcal{L}_{ur} \geq \mathcal{L}_r$

Example #1: Female Labor Force Participation (code)

```
********************************************************************************
*                                                                              *
*       Female Participation in the Labor Force - Estimation                   *
*                                                                              *
********************************************************************************;
reg     inlf educ exper expersq age kidslt6 kidsge6 nwifeinc, robust;
outreg  using "`TableA'", bdec(3) se tex title(Female Labor Force Participation)
        ctitle("", OLS) replace;

logit   inlf educ exper expersq age kidslt6 kidsge6 nwifeinc;
outreg  using "`TableA'", bdec(3) se tex title(Female Labor Force Participation)
        ctitle("", logit) merge;

probit  inlf educ exper expersq age kidslt6 kidsge6 nwifeinc;
outreg  using "`TableA'", bdec(3) se tex title(Female Labor ForceParticipation)
        ctitle("", probit) merge;
```

Example #1: Female Labor Force Participation (mroz.dta)

	OLS	logit	probit
educ	0.038	0.221	0.131
	(0.007)**	(0.043)**	(0.025)**
exper	0.039	0.206	0.123
	(0.006)**	(0.032)**	(0.019)**
expersq	-0.001	-0.003	-0.002
	(0.000)**	(0.001)**	(0.001)**
age	-0.016	-0.088	-0.053
	(0.002)**	(0.015)**	(0.008)**
kidslt6	-0.262	-1.443	-0.868
	(0.032)**	(0.204)**	(0.119)**
kidsge6	0.013	0.060	0.036
	(0.014)	(0.075)	(0.043)
nwifeinc	-0.003	-0.021	-0.012
	(0.002)*	(0.008)*	(0.005)*
_cons	0.586	0.425	0.270
	(0.152)**	(0.860)	(0.509)

Interpreting Logit and Probit Models

- In a linear probability model, computing marginal effects is relatively easy, since

$$\frac{\partial P(y=1|\boldsymbol{x})}{\partial x_j} = \beta_j \tag{17}$$

- Marginal effects are only slightly more complicated if the LPM includes quadratic or logarithmic terms.
- Because both the logit and probit models are nonlinear functions, the marginal effects are generally more complicated.
- Indeed this is true for any choice of $G(z)$ that is nonlinear.

Marginal Effects - Continuous Variables

- For a continuous variable, the partial effect of a change in, say, x_j is given by:
$$\frac{\partial P(y=1|\boldsymbol{x})}{\partial x_j} = g(\boldsymbol{x\beta})\beta_j \quad (18)$$
where $g(z) \equiv G'(z)$.
  - For the probit model, $g(z) = \phi(z)$, the standard normal *pdf*;
  - For the logit model, $g(z) = \Lambda(z)[1-\Lambda(z)] = \frac{exp(z)}{[1+exp(z)]^2}$.
- Some important implications of (19) are:
  - the marginal impact varies with $\boldsymbol{x}$.
  - for both the logit and probit models, this marginal effect peaks at $z=0$, which corresponds to the point where $P(y=1|\boldsymbol{x}) = G(0) = 0.5$.
  - While the parameters do not directly indicate marginal effects, they do indicate their signs (since $g(z) > 0$) *and* relative effects, since
$$\frac{\partial P(y=1|\boldsymbol{x})/\partial x_j}{\partial P(y=1|\boldsymbol{x})/\partial x_k} = \frac{g(\boldsymbol{x\beta})\beta_j}{g(\boldsymbol{x\beta})\beta_k} = \frac{\beta_j}{\beta_k} \quad (19)$$

Marginal Effects - Discrete Variables

- The impact of change in a discrete variable (say x_j) is computed by holding all other factors fixed and changing x_j from 0 to 1;
- Let $\boldsymbol{x}_{-j}$ denote $\boldsymbol{x}$ excluding x_j, the effect of x_j is given by

$$\begin{aligned} &P(y=1|\boldsymbol{x}_{-j}, x_j=1) - P(y=1|\boldsymbol{x}_{-j}, x_j=0) \\ &G(\beta_0+\beta_1x_1+\cdots+\beta_{j-1}x_{j-1}+\beta_j+\beta_{j+1}x_{j+1}+\cdots+\beta_kx_k) \\ -&G(\beta_0+\beta_1x_1+\cdots+\beta_{j-1}x_{j-1}+0+\beta_{j+1}x_{j+1}+\cdots+\beta_kx_k) \quad (20) \end{aligned}$$

- A similar approach is used to compute the impact of a change in an integer variable (such as the number of children) from say c_j to c_j+1; i.e.,

$$\begin{aligned} &P(y=1|\boldsymbol{x}_{-j}, x_j=c_j+1) - P(y=1|\boldsymbol{x}_{-j}, x_j=c_j) \\ &G(\beta_0+\beta_1x_1+\cdots+\beta_{j-1}x_{j-1}+\beta_j(c_j+1)+\beta_{j+1}x_{j+1}+\cdots+\beta_kx_k) \\ -&G(\beta_0+\beta_1x_1+\cdots+\beta_{j-1}x_{j-1}+\beta_j(c_j)+\beta_{j+1}x_{j+1}+\cdots+\beta_kx_k) \end{aligned} \quad (21)$$

Computing and Comparing Marginal Effects

- Because the marginal effects vary with $\boldsymbol{x}$, a natural question is what value(s) of $\boldsymbol{x}$ do we use?
  - One approach is to use $\boldsymbol{x} = \bar{\boldsymbol{x}}$, which yields the **partial effect at the average (PEA)**.
    ...This is usually not a good idea, as $\bar{\boldsymbol{x}}$ may not represent a very interesting case or any reasonable individual in the sample.
  - A more reasonable approach, especially if you have a random sample, is to compute the marginal effect for each individual in the sample and average these marginal effects.
    ...This is know as the **average partial effect (APE)** or **average marginal effect (AME)**
- Direct comparisons among LPM, logit and probit parameters can be misleading, as they involve difference scaling factors.

  ...Note that $g(\cdot)$ maximizes at roughly 0.4 for the probit model and at 0.25 for the logit model.
- It is better, instead, to compare marginal effects.

Ex. #1: Female Labor Force Participation (parameters)

	OLS	logit	probit
educ	0.038	0.221	0.131
	(0.007)**	(0.043)**	(0.025)**
exper	0.039	0.206	0.123
	(0.006)**	(0.032)**	(0.019)**
expersq	-0.001	-0.003	-0.002
	(0.000)**	(0.001)**	(0.001)**
age	-0.016	-0.088	-0.053
	(0.002)**	(0.015)**	(0.008)**
kidslt6	-0.262	-1.443	-0.868
	(0.032)**	(0.204)**	(0.119)**
kidsge6	0.013	0.060	0.036
	(0.014)	(0.075)	(0.043)
nwifeinc	-0.003	-0.021	-0.012
	(0.002)*	(0.008)*	(0.005)*
_cons	0.586	0.425	0.270
	(0.152)**	(0.860)	(0.509)

Example #1: Female Labor Force Participation (marginal effects code)

```
******************************************************************************
*                                                                            *
*       Female Participation in the Labor Force - marginal effects           *
*                                                                            *
******************************************************************************;
reg     inlf educ exper expersq age kidslt6 kidsge6 nwifeinc, robust;
margins, dydx(*);
outreg  using "`TableB'", stat(b_dfdx se_dfdx) bdec(4) tex
        title(Marginal Effect Female Labor Force Participation)
        ctitle("", LPM) nocons replace;

logit   inlf educ exper expersq age kidslt6 kidsge6 nwifeinc;
margins, dydx(*);
outreg  using "`TableB'", stat(b_dfdx se_dfdx) bdec(4) tex
        title(Marginal Effect Female Labor Force Participation)
        ctitle("", logit) nocons merge;

probit  inlf educ exper expersq age kidslt6 kidsge6 nwifeinc;
margins, dydx(*);
outreg  using "`TableB'", stat(b_dfdx se_dfdx) bdec(4) tex
        title(Marginal Effect Female Labor Force Participation)
        ctitle("", probit) nocons merge;
```


Example #1: Marginal Effects Female Labor Force

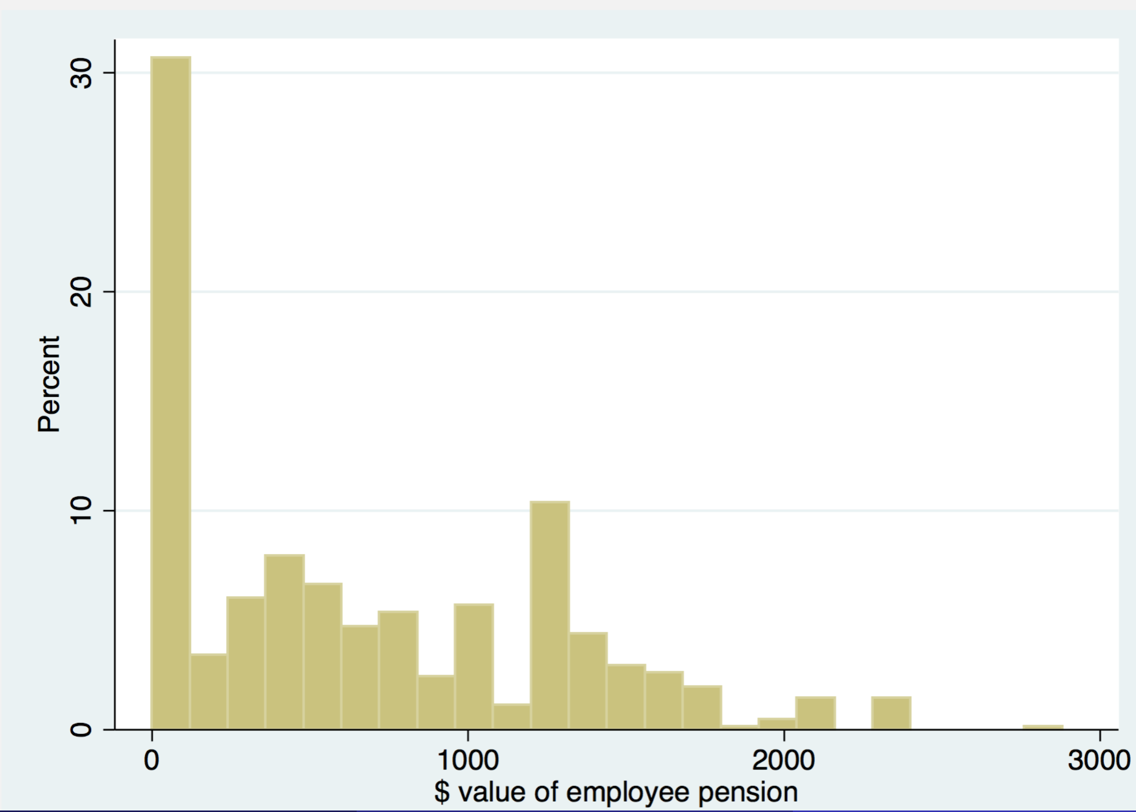
	LPM	logit	probit
educ	0.0380	0.0395	0.0394
	(0.0073)**	(0.0073)**	(0.0072)**
exper	0.0395	0.0368	0.0371
	(0.0058)**	(0.0052)**	(0.0052)**
expersq	-0.0006	-0.0006	-0.0006
	(0.0002)**	(0.0002)**	(0.0002)**
age	-0.0161	-0.0157	-0.0159
	(0.0024)**	(0.0024)**	(0.0024)**
kidslt6	-0.2618	-0.2578	-0.2612
	(0.0318)**	(0.0319)**	(0.0319)**
kidsge6	0.0130	0.0107	0.0108
	(0.0135)	(0.0133)	(0.0131)
nwifeinc	-0.0034	-0.0038	-0.0036
	(0.0015)*	(0.0015)*	(0.0014)*

* $p < 0.05$; ** $p < 0.01$

Corner Solutions

- The second type of limited dependent variable we want to discuss in this chapter arises when there are corner solutions.
- For example, in a model of meat demand, a portion of the population will have strictly zero consumption in a given period, due to either budgetary issues or perferences.
- While one can still model $E(y|\boldsymbol{x})$ as a linear function, such a model will potentially yield negative predictions for y for a substantial portion of the population.
- This is analogous to a similar problem for the LPM in a discrete choice setting.

Example: Employee Pension Contributions



The Tobit Model

- A traditional approach to the corner solutions problem is the **Tobit Model**, which uses an underlying latent variable:

$$y^* = \boldsymbol{x\beta} + u \tag{22}$$

where $u|\boldsymbol{x} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$.

- The observed response variable is then given by $y = max(0, y^*)$.
- This structure allows for a discrete mass of observations at $y = 0$, while allowing for a continuum of outcomes above zero.

Estimating the Tobit Model

- There are several ways to estimate the Tobit model, but the most common in maximum likelihood estimation (MLE).
- For observations at the corner, we have:

$$\begin{aligned}
P(y = 0|\boldsymbol{x}) &= P(y^* < 0|\boldsymbol{x}) \\
&= P(\boldsymbol{x\beta} + u < 0|\boldsymbol{x}) \\
&= P(u < -\boldsymbol{x\beta}|\boldsymbol{x}) \\
&= P\left(\frac{u}{\sigma} < \frac{-\boldsymbol{x\beta}}{\sigma}|\boldsymbol{x}\right) \\
&= \Phi\left(\frac{-\boldsymbol{x\beta}}{\sigma}\right) \\
&= 1 - \Phi\left(\frac{\boldsymbol{x\beta}}{\sigma}\right) \qquad (23)
\end{aligned}$$

where $\Phi(\cdot)$ is the standard normal cdf.

Estimating the Tobit Model (continued)

- For those individual with $y_i > 0$, their likelihood function is given by

$$\frac{1}{\sigma}\phi\left(\frac{y_i - \boldsymbol{x}_i\boldsymbol{\beta}}{\sigma}\right) \tag{24}$$

- The corresponding log-likelihood function is given by $\mathcal{L} = \sum_{i=1}^{n} \mathcal{L}_i$ where:

$$\begin{aligned}\mathcal{L}_i = {} & 1(y_i = 0) log\left(\left[1 - \Phi\left(\frac{-\boldsymbol{x}\boldsymbol{\beta}}{\sigma}\right)\right]\right) \\ & + 1(y_i > 0) log\left[\frac{1}{\sigma}\phi\left(\frac{y_i - \boldsymbol{x}_i\boldsymbol{\beta}}{\sigma}\right)\right]\end{aligned} \tag{25}$$

- Test statistics (t, F, and LR) are formed in the usual way, though typically without using robust standard errors.

Interpreting Tobit Estimates

- Again, given the nonlinearity of the model, interpreting the Tobit Estimates becomes more difficult.
- We may be interested in how $E(y|\boldsymbol{x})$ changes with changes in x_j.

  ...in which case it helps to know that

$$\begin{aligned} E(y|\boldsymbol{x}) &= P(y > 0|\boldsymbol{x})E(y|y > 0, \boldsymbol{x}) \\ &= \Phi\left(\frac{\boldsymbol{x}\boldsymbol{\beta}}{\sigma}\right)[\boldsymbol{x}\boldsymbol{\beta} + \sigma\lambda(\boldsymbol{x}\boldsymbol{\beta}/\sigma)] \\ &= \Phi\left(\frac{\boldsymbol{x}\boldsymbol{\beta}}{\sigma}\right)\boldsymbol{x}\boldsymbol{\beta} + \sigma\phi\left(\frac{\boldsymbol{x}\boldsymbol{\beta}}{\sigma}\right) \end{aligned} \tag{26}$$

where $\lambda(z) \equiv \phi(a)/\Phi(z)$ is known as the **inverse Mills ratio** and

$$E(y|y > 0, \boldsymbol{x}) = \boldsymbol{x}\boldsymbol{\beta} + \sigma\lambda(\boldsymbol{x}\boldsymbol{\beta}/\sigma). \tag{27}$$

Interpreting Tobit Estimates (cont'd)

- Using this finding,
  1. One can then show that
  $$\frac{\partial E(y|y>0,\boldsymbol{x})}{\partial x_j} = \beta_j \tau\left(\frac{\boldsymbol{x}\boldsymbol{\beta}}{\sigma}\right) \tag{28}$$
  where $\tau(c) \in (0,1)$, with
  $$\tau(c) = \{1 - \lambda(c)[c + \lambda(c)]\}. \tag{29}$$
  2. Moreover:
  $$\frac{\partial E(y|\boldsymbol{x})}{\partial x_j} = \beta_j \Phi\left(\frac{\boldsymbol{x}\boldsymbol{\beta}}{\sigma}\right) \in (0, \beta_j) \tag{30}$$
- The adjustment $\Phi(c)$ reflects the fact that there is a nonzero probability that the individual's outcome will be at the corner and not impacted by a change in x_j.

Average Partial Effects

- The results from the previous slide indicate that the Average Partial Effect (APE) is given by
$$APE = \beta_j \bar{\Phi} \tag{31}$$
where
$$\bar{\Phi} = \frac{1}{n}\sum_{i=1}^{n} \Phi\left(\frac{\boldsymbol{x}_i\boldsymbol{\beta}}{\sigma}\right) = \frac{1}{n}\sum_{i=1}^{n} P(y_i > 0|\boldsymbol{x}_i) \tag{32}$$

Specification Issues

- One limitation associated with the Tobit model, other than the normality assumption, is that it both the probability of not being at the corner and the level of usage given you are not at the corner are driven by the same index; i.e., $\boldsymbol{x}_i\boldsymbol{\beta}$.
- This assumption, however, may not hold.
- Suppose we are modeling fire damage as a function of house age
  - We would expect the probability of a fire to increase with age,
  - but the fire damage (in monetary terms) to decrease with age.
- There are models (e.g., *hurdle* and *two-part* models) that allow the corner probability and conditional usage to have differing parameters and even regressors.

Count Data Models

- Count data models arise when the dependent variable can take on nonnegative integer values.
- Examples include
  - Number of fatalities from mule kicks in Prussian army;
  - Number of children;
  - Number of arrests;
  - Number of marriages;
  - Number of doctor visits in a year;
  - Number of recreation trips to Michigan Lakes; and
  - Number of patents filed by a firm.
- While we can model these outcomes using a linear regression model, the model will potentially yield predictions that are negative for some observations.

The Poisson Model

- The most popular model in this setting is based on the **Poisson distribution**, where
$$P(y=h)=\frac{e^{-\mu}\mu^{h}}{y!} \tag{33}$$
- A key feature of the Poisson distribution is that
$$E(y)=Var(y)=\mu, \tag{34}$$
a property known as *equidispersion*.
- The **Poisson Regression Model** results if we assume
$$\mu=E(y|\boldsymbol{x})=exp(\beta_0+\beta_1x_1+\cdots+\beta_kx_k) \tag{35}$$
- MLE is used for estimation, with the log-likelihood function given by
$$\mathcal{L}(\boldsymbol{y},\boldsymbol{X},\beta)=\sum_{i=1}^{n}[y_i\boldsymbol{x}_i\boldsymbol{\beta}-exp(\boldsymbol{x}_i\boldsymbol{\beta})] \tag{36}$$

Model Interpretation

- As with all nonlinear regression models, the parameters do not indicate the marginal impact of an explanatory variable
- Given the exponential conditional mean; i.e.
$$E(y_i|\boldsymbol{x}_i)=exp(\boldsymbol{x}_i\boldsymbol{\beta}), \tag{37}$$
we have that
$$\frac{\partial E(y_i|\boldsymbol{x}_i)}{\partial x_{ik}}=\beta_k exp(\boldsymbol{x}_i\boldsymbol{\beta}) \tag{38}$$
- Elasticities have a convenient form
$$\frac{\partial E(y_i|\boldsymbol{x}_i)}{\partial x_{ik}}\frac{x_{ik}}{E(y_i|\boldsymbol{x}_i)}=\beta_k x_{ik} \tag{39}$$

The Average Partial Effect

- The corresponding average partial effect (APE) is then given by

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\partial E(y_i|\boldsymbol{x}_i)}{\partial x_{ik}} = \frac{1}{N}\sum_{i=1}^{N}\beta_k exp(\boldsymbol{x}_i\boldsymbol{\beta}) = \beta_k\left[\frac{1}{N}\sum_{i=1}^{N}exp(\boldsymbol{x}_i\boldsymbol{\beta})\right] \quad (40)$$

- For the Poisson ML estimates, when there is a constant in the model, this reduces considerably, since the model is *mean fitting*, with

$$\frac{1}{N}\sum_{i=1}^{N}exp(\boldsymbol{x}_i\hat{\boldsymbol{\beta}}) = \frac{1}{N}\sum_{i=1}^{N}y_i = \bar{y} \quad (41)$$

$$\Rightarrow$$

$$\widehat{APE} = \frac{1}{N}\sum_{i=1}^{N}\frac{\partial E(y_i|\boldsymbol{x}_i)}{\partial x_{ik}} = \hat{\beta}_k\bar{y} \quad (42)$$

where $\hat{\boldsymbol{\beta}}$ denotes the MLE parameter estimates.

Interpretation of Coefficient (cont'd)

- As we have seen in other nonlinear regression models, one typically does not want to rely on the response as the average characteristic

$$\left.\frac{\partial E(y_i|\boldsymbol{x}_i)}{\partial x_{ik}}\right|_{x_i=\bar{x}} = \beta_k exp(\boldsymbol{x}_i\boldsymbol{\beta}) < \frac{1}{N}\sum_{i=1}^{N}\frac{\partial E(y_i|\boldsymbol{x}_i)}{\partial x_{ik}} \quad (43)$$

- One can also compute average elasticities relatively easily, with

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\partial lnE(y_i|\boldsymbol{x}_i)}{\partial ln x_{ik}} = \frac{1}{N}\sum_{i=1}^{N}\beta_k x_{ik} = \beta_k\left[\frac{1}{N}\sum_{i=1}^{N}x_{ik}\right] = \beta_k\bar{x}_k \quad (44)$$

Example: OLS vs Poisson in Arrest Model Parameters

	OLS	Poisson
pcnv	-0.132	-0.402
	(0.034)**	(0.101)**
avgsen	-0.011	-0.024
	(0.014)	(0.024)
tottime	0.012	0.024
	(0.013)	(0.021)
ptime86	-0.041	-0.099
	(0.007)**	(0.022)**
qemp86	-0.051	-0.038
	(0.014)**	(0.034)
inc86	-0.001	-0.008
	(0.000)**	(0.001)**
black	0.327	0.661
	(0.058)**	(0.099)**
hispan	0.194	0.500
	(0.040)**	(0.092)**
[illegible]	[illegible]	[illegible]
	(0.032)	(0.081)
cons	0.577	-0.600

Example: OLS vs Poisson in Arrest Model APE's

	OLS	Poisson
pcnv	-0.132	-0.162
	(0.034)**	(0.040)**
avgsen	-0.011	-0.010
	(0.014)	(0.010)
tottime	0.012	0.010
	(0.013)	(0.008)
ptime86	-0.041	-0.040
	(0.007)**	(0.009)**
qemp86	-0.051	-0.015
	(0.014)**	(0.014)
inc86	-0.001	-0.003
	(0.000)**	(0.001)**
black	0.327	0.267
	(0.058)**	(0.042)**
hispan	0.194	0.202
	(0.040)**	(0.038)**
[illegible]	[illegible]	[illegible]
	(0.032)	(0.033)

Limitations of the Poisson Regression Model

- A key limitation of the Poisson Regression Model is that it assumes equidispersion; i.e.,

$$E(y|\boldsymbol{x}) = Var(y|\boldsymbol{x}). \tag{45}$$

- Note that this is *conditional* equidispersion, which is not the same as *unconditional* equidispersion.
- In practice, most data exhibit overdispersion; i.e.,

$$E(y|\boldsymbol{x}) < Var(y|\boldsymbol{x}). \tag{46}$$

- A rule of thumb is that overdisperion is likely to be a problem if $Var(y) > 2E(y)$.
- Fortunately, in the case of the Poisson Regression model, the MLE parameter estimates will remain consistent as long as $E(y|\boldsymbol{x}) = exp(\boldsymbol{x}\boldsymbol{\beta})$, even if equidispersion is violated.

  ...However, the MLE standard errors must be corrected.

Censored Normal Regression Model

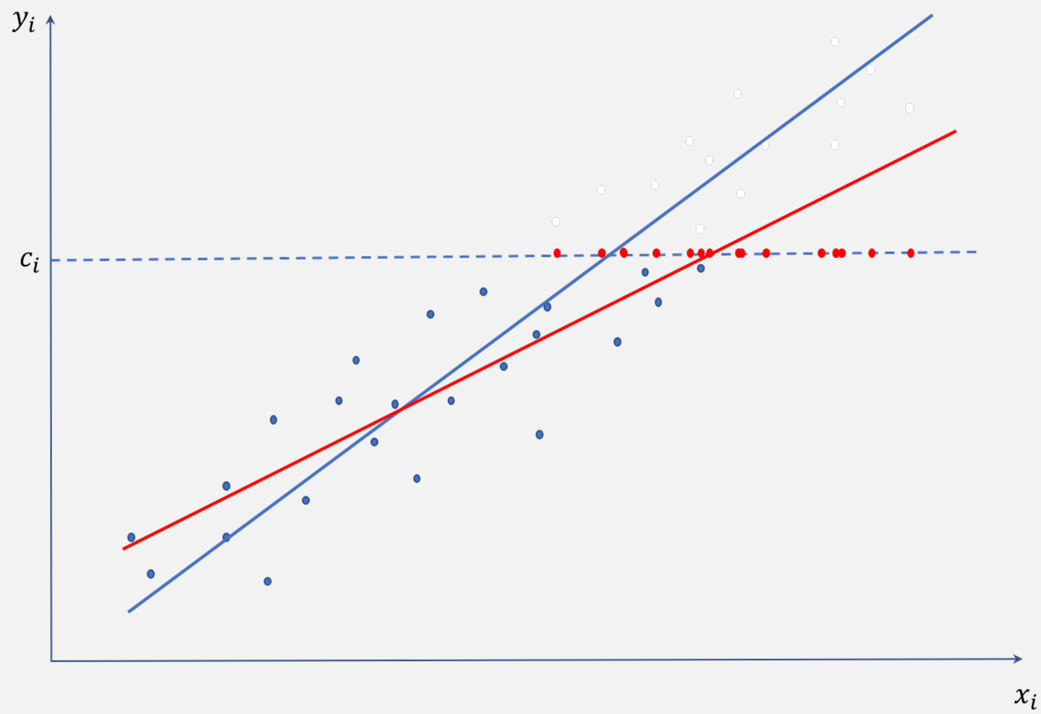
- There are many settings in which we do not observe the variable of interest, but instead a *censored* version of it.
- Examples of censoring including
  - top coding in surveys;
  - capacity constraints (possibly);
  - minimum wage restrictions.
- Note: This is a data collection problem, unlike in the Tobit model.
- Here we focus on the **censored normal regression model**.
- With *censoring from above*, we observe $w_i = min(y_i, c_i)$, where

$$y_i = \boldsymbol{x}_i\boldsymbol{\beta} + u_i \tag{47}$$

  with $u_i|\boldsymbol{x}_i, c_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$ and c_i denotes the point of censoring.

  ...In this case, we only observe y_i if it is below the censoring value, otherwise we observe c_i.

Attenuation Bias with Censoring



Estimation

- Estimation of the censored normal regression model typically proceeds using maximum likelihood estimation.
- As with any MLE, we need to form the likelihood function (i.e., density function) for our observed variable w_i.
- For the uncensored observations, $w_i = y_i$, so the density is the same.
- For the censored observations, we have

$$\begin{aligned} P(w_i = c_i | \boldsymbol{x}_i) &= P(y_i \geq c_i | \boldsymbol{x}_i) \\ &= P(u_i \geq c_i - \boldsymbol{x}\boldsymbol{\beta} | \boldsymbol{x}_i) \\ &= 1 - \Phi\left(\frac{c_i - \boldsymbol{x}\boldsymbol{\beta}}{\sigma}\right) \end{aligned}$$

Estimation (cont'd)

- The corresponding log-likelihood becomes

$$\mathcal{L}(\boldsymbol{w}, \boldsymbol{x}, \beta) = \sum_{i=1}^{n} \left\{ 1(w_i = c_i) log \left[1 - \Phi\left(\frac{c_i - \boldsymbol{x}\boldsymbol{\beta}}{\sigma} \right) \right] \right.$$
$$\left. + 1(w_i < c_i) log \left[\frac{1}{\sigma} \phi \left(\frac{w_i - \boldsymbol{x}\boldsymbol{\beta}}{\sigma} \right) \right] \right\} \tag{48}$$

- Unlike our earlier corner solutions setting (e.g., Tobit models), the coefficients (i.e., the β_j's) do indicate the marginal effects for the *uncensored* population.

Truncated Regression Models

- A similar problem emerges in the **truncated regression model**.
- The difference is that:

  . . . whereas in the case of censoring we still observe the covariates (i.e., the $\boldsymbol{x}_i$'s) for the censored observations,

  . . . with truncation, we have no information about the affected observations (i.e., both y_i and $\boldsymbol{x}_i$ are missing).
- This is again a data gathering problem.
- In the case of the **truncated normal regression model**, we have

$$y_i = \boldsymbol{x}_i\boldsymbol{\beta} + u_i, u_i|\boldsymbol{x}_i, c_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2). \tag{49}$$

- We do not have a random sample, but instead observe y_i and its covariates only if y_i falls within a given range.

Estimation

- Estimation typically proceeds using MLE.
- For example, with truncation from above (leaving $y_i < c_i$), the log-likelihood function becomes

$$\mathcal{L}(\boldsymbol{y}, \boldsymbol{x}, \beta) = \sum_{i=1}^{n} \left\{ log \left[\frac{1}{\sigma} \phi \left(\frac{y_i - \boldsymbol{x}\boldsymbol{\beta}}{\sigma} \right) \right] - log \left[\Phi \left(\frac{c_i - \boldsymbol{x}\boldsymbol{\beta}}{\sigma} \right) \right] \right\} \tag{50}$$

- Truncation, like censoring, leads to attenuation bias.
- As in the case of censoring, the coefficients (i.e., the β_j's) do indicate the marginal effects for the *untruncated* population.