EC 655
Limited Dependent Variables 1

Justin Smith

Fall 2021

Introduction

- Many economic outcomes are qualitative in nature
  - Grouped into categories
    - Work or not
    - Drive, take bus, cycle to work
    - Number of visits to Doctor
- People sometimes use alternative methods in this context
  - Though linear regression does still work in many cases
- In this section we cover models for these types of variables

Binary Choice

Potential Outcomes and Linear Regression

- Start with the same potential outcomes setup
  - y_1 is the outcome with treatment
  - y_0 is the outcome without treatment
  - w is a binary variable with 1 denoting treatment, and 0 no treatment
- We observe (y, w), where

$$y = y_0 + (y_1 - y_0)w$$

- The *average* difference between treated and control is

$$E(y|w=1) - E(y|w=0)$$

$$= [E(y_1|w=1) - E(y_0|w=1)] + E(y_0|w=1) - E(y_0|w=0)$$

Binary Choice

- Key difference here is that y_1 and y_0 are binary

$$y_0 \in \{0,1\}$$
$$y_1 \in \{0,1\}$$

- You can get treatments effects in the same way we did before
  - Use linear regression
  - Estimate by OLS
- Only the *interpretation* changes
  - Since y is a Bernoulli random variable,

$$E(y|w) = Pr(y=1|w)$$

Binary Choice

- So you can restate the difference in observed means as

$$Pr(y = 1|w = 1) - Pr(y = 1|w = 0)$$

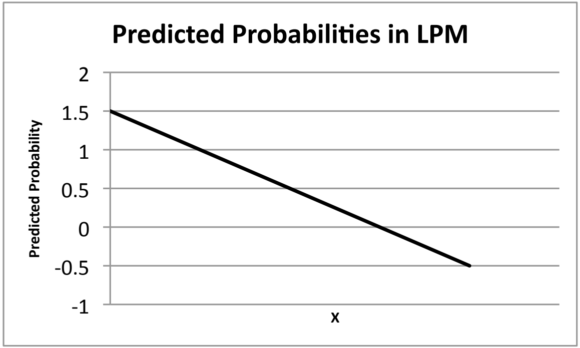
$$= [Pr(y_1 = 1|w = 1) - Pr(y_0 = 1|w = 1)] + Pr(y_0 = 1|w = 1) - Pr(y_0 = 1|w = 0)$$

- You can interpret as the difference in *response probabilities*
- Difference in observed response probabilities will estimate treatment effects if
  - Independence of potential outcomes
  - Mean independence of potential outcomes
  - Conditional mean independence of potential outcomes
- Can also use linear regression with binary or continuous treatment
- Called the **Linear Probability Model**

Binary Choice

Issues with Linear Probability Model

1. Predicted probabilities are not restricted to [0,1] interval
   - Can have nonsense probabilities



Binary Choice

2. Heteroskedasticity

$$Var[y|w] = E[y - E[y|w]|w]$$
$$= E[y^2|w] - (E[y|w])^2$$
$$= Pr[y = 1|w](1 - Pr[y = 1|w])$$

- As noted above $Pr[y = 1|w]$ depends on w
- So the variance of y will depend on w
- We therefore must use robust standard errors
- You could also use GLS

Binary Choice

Nonlinear Models

- In some cases we may want to fix the predicted probability issue
  - If you are doing prediction, for example
- One way to do this is to feed the model through a CDF
- This is often motivated with an index model

$$y_0^* = \beta_0 - \eta$$

$$y_1^* = y_0^* + \beta_1$$

$$y^* = y_0^* + (y_1^* - y_0^*)w$$

$$y^* = \beta_0 + \beta_1 w - \eta$$

Binary Choice

- In this setup, we do not observe y^*
  - It is some underlying continuous outcome driving our decisions
- Instead, we observe the binary y where

$$y = 1\{y^* > 0\}$$

- Plugging the model into this

$$y = 1\{\beta_0 + \beta_1 w - \eta > 0\}$$

- The probability that $y = 1$ is therefore

$$Pr(y = 1|w) = Pr(\beta_0 + \beta_1 w - \eta > 0|w)$$

Binary Choice

- The random component here is η, so rearrange to isolate it

$$Pr(y = 1|w) = Pr(\beta_0 + \beta_1 w > \eta|w)$$
$$Pr(y = 1|w) = Pr(\eta < \beta_0 + \beta_1 w|w)$$

- The probability that y equals 1 depends on
  - Treatment status w
  - The distribution of eta
- Different choices for the η distribution lead to different models

Binary Choice

Probit Model

- Assuming $\eta \sim \mathcal{N}(0, \sigma_\eta^2)$ leads to the **Probit Model**

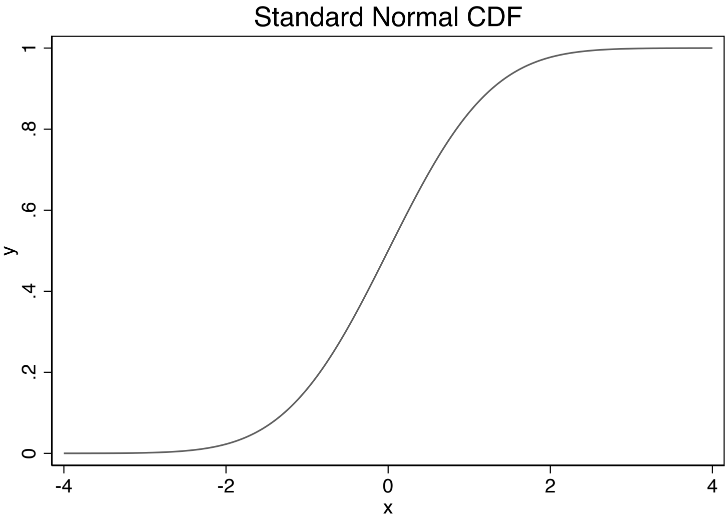
$$Pr(y = 1|w) = \Phi\left(\frac{\beta_0 + \beta_1 w}{\sigma_\eta^2}\right)$$

- where

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{\frac{-v^2}{2}}\, dv$$

- Because $\Phi(.)$ is a CDF, $Pr(y = 1|w)$ is always between 0 and 1
  - This solves the predicted probability problem

Binary Choice



Binary Choice

- The observed difference in probabilities is

$$Pr(y = 1|w = 1) - Pr(y = 1|w = 0)$$
$$= \Phi\left(\frac{\beta_0 + \beta_1}{\sigma_\eta^2}\right) - \Phi\left(\frac{\beta_0}{\sigma_\eta^2}\right)$$

- Notice that β_1 does <u>not</u> equal the difference in response probabilities
  - They are the slopes in the index model
  - The index model parameters are not usually of interest
- To get the difference in response probabilities, feed parameters into the CDF first
- In nonlinear models, parameters are not "marginal effects"
  - You need to separately compute them after estimating the model

Binary Choice

- In models with more variables and where they are continuous

$$Pr(y=1|\mathbf{x}) = \Phi\left(\frac{\mathbf{x}\boldsymbol{\beta}}{\sigma_\eta^2}\right)$$

- The marginal effect for continuous variable x_j

$$\frac{\partial Pr(y=1|\mathbf{x})}{\partial x_j} = \phi\left(\frac{\mathbf{x}\boldsymbol{\beta}}{\sigma_\eta^2}\right)\frac{\beta_j}{\sigma_\eta^2}$$

- This is a function of the entire vector $\mathbf{x}$
  - You need to specify their values to get the marginal effect
  - Normally people hold them at the mean
  - In theory you can get a distribution of marginal effects

Binary Choice

Logit Model

- Assuming $\eta \sim$ Logistic$(0, \sigma_\eta^2)$ leads to the **Logit Model**

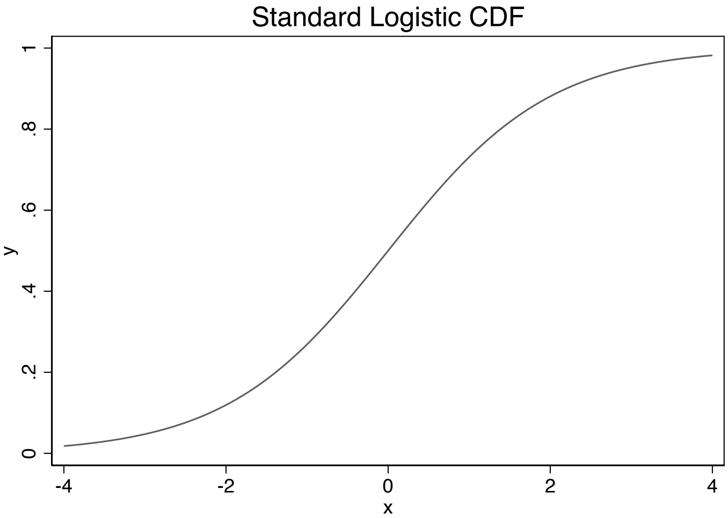
$$Pr(y = 1|w) = \Lambda\left(\frac{\beta_0 + \beta_1 w}{\sigma_\eta^2}\right)$$

- where

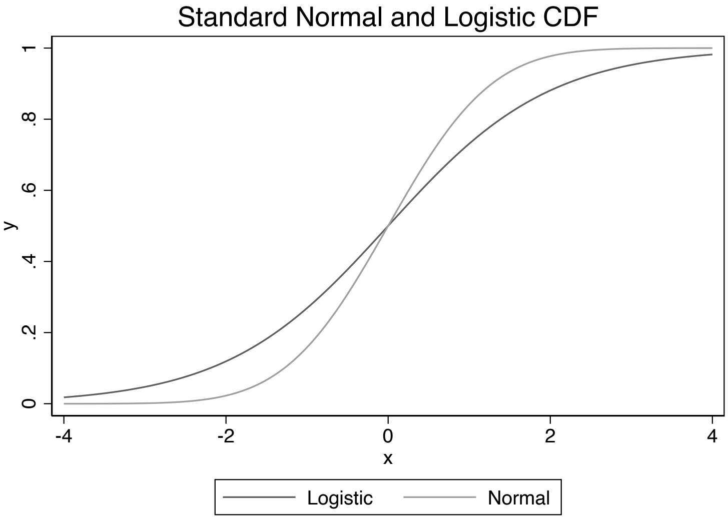
$$\Lambda(z) = \frac{e^z}{1 + e^z}$$

- Again, because $\Lambda(.)$ is a CDF, $Pr(y = 1|w)$ is always between 0 and 1

Binary Choice



Binary Choice



Binary Choice

- The observed difference in probabilities is

$$
\begin{aligned}
&Pr(y=1|w=1) - Pr(y=1|w=0) \\
&= \Lambda\left(\frac{\beta_0+\beta_1}{\sigma_\eta^2}\right) - \Lambda\left(\frac{\beta_0}{\sigma_\eta^2}\right)
\end{aligned}
$$

- In models with more variables and where they are continuous

$$
Pr(y=1|\mathbf{x}) = \Lambda\left(\frac{\mathbf{x}\boldsymbol{\beta}}{\sigma_\eta^2}\right)
$$

- The marginal effect for continuous variable x_j

$$
\frac{\partial Pr(y=1|\mathbf{x})}{\partial x_j} = \Lambda\left(\frac{\mathbf{x}\boldsymbol{\beta}}{\sigma_\eta^2}\right)\left(1-\Lambda\left(\frac{\mathbf{x}\boldsymbol{\beta}}{\sigma_\eta^2}\right)\right)\frac{\beta_j}{\sigma_\eta^2}
$$

Estimation of Probit and Logit

- Both models usually estimated by Maximum Likelihood
- Method maximizes the probability of getting our sample by choosing parameters
  - The**Likelihood Function** is the function of the parameters given the data
- The probability distribution of y_i is

$$f(y_i|\mathbf{x_i};\boldsymbol{\beta}) = F\left(\frac{\mathbf{x_i}\boldsymbol{\beta}}{\sigma_\eta^2}\right)^{y_i}\left(1 - F\left(\frac{\mathbf{x_i}\boldsymbol{\beta}}{\sigma_\eta^2}\right)\right)^{1-y_i}$$

- The joint probability of observing the all the y_i in the data is

$$f(\mathbf{y}|\mathbf{X};\boldsymbol{\beta}) = \Pi_{i=1}^{n} F\left(\frac{\mathbf{x_i}\boldsymbol{\beta}}{\sigma_\eta^2}\right)^{y_i}\left(1 - F\left(\frac{\mathbf{x_i}\boldsymbol{\beta}}{\sigma_\eta^2}\right)\right)^{1-y_i}$$

Binary Choice

- The Likelihood Function recasts as a function of the parameters given the data

$$\mathcal{L}(\boldsymbol{\beta}|\mathbf{y},\mathbf{X}) = \Pi_{i=1}^{n} F\left(\frac{\mathbf{x_i}\boldsymbol{\beta}}{\sigma_\eta^2}\right)^{y_i}\left(1 - F\left(\frac{\mathbf{x_i}\boldsymbol{\beta}}{\sigma_\eta^2}\right)\right)^{1-y_i}$$

- Both models usually estimated by Maximum Likelihood
- Method maximizes the joint probability of y values conditional on $\mathbf{x}$
  - This is called the **Likelihood Function**
- In the case of a binary y, the likelihood function is

$$\begin{aligned}
f[y|\mathbf{x};\boldsymbol{\beta}] &= P[Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n|\mathbf{x_i};\boldsymbol{\beta}] \\
&= \Pi_{y_i=1} F\left(\frac{\mathbf{x_i}\boldsymbol{\beta}}{\sigma_\eta^2}\right)\Pi_{y_i=0}\left[1 - F\left(\frac{\mathbf{x_i}\boldsymbol{\beta}}{\sigma_\eta^2}\right)\right] \\
&= \Pi_{i=1}^{n} F\left(\frac{\mathbf{x_i}\boldsymbol{\beta}}{\sigma_\eta^2}\right)^{y_i}\left[1 - F\left(\frac{\mathbf{x_i}\boldsymbol{\beta}}{\sigma_\eta^2}\right)\right]^{1-y_i}
\end{aligned}$$

Binary Choice

- Researchers usually focus on the log of the likelihood instead
  - It is a monotonic (increasing) transformation of the likelihood
  - The same parameter vector solves both versions
  - Log likelihoods are easier to work with
- Log Likelihood

$$ln\mathcal{L}(\boldsymbol{\beta}|\mathbf{y},\mathbf{X}) = \sum_{i=1}^{N}\{y_i lnF\left(\frac{\mathbf{x_i}\boldsymbol{\beta}}{\sigma_\eta^2}\right) + (1-y_i)ln(1-F\left(\frac{\mathbf{x_i}\boldsymbol{\beta}}{\sigma_\eta^2}\right))\}$$

- To solve this equation, you need numerical methods
  - A grid search algorithm that finds the maximum value

Binary Choice

- In ML environments, the estimated variance of $\hat{\boldsymbol{\beta}}$ is estimated as the negative of the expected value of the Hessian (information matrix)

$$V\hat{a}r(\hat{\boldsymbol{\beta}}) = -E[(\frac{\partial^2 lnL}{\partial\hat{\boldsymbol{\beta}}\partial\hat{\boldsymbol{\beta}}'})^{-1}]$$

$$= (\sum_{i=1}^{n} \frac{f(\mathbf{x_i}'\hat{\boldsymbol{\beta}})^2}{F(\mathbf{x_i}'\hat{\boldsymbol{\beta}})(1 - F(\mathbf{x_i}'\hat{\boldsymbol{\beta}}))} \mathbf{x_i}\mathbf{x_i}')^{-1}$$

- where $F(.)$ is either the Normal or Logistic CDF, and $f(.)$ is the associated PDF

Hypothesis Testing in Probit and Logit

- Simple tests for coefficient significance is done by the usual t-test method
  - Assume $\hat{\beta}$ has normal distribution (asymptotically)
  - Test statistic $Z = \frac{\hat{\beta}_k}{S\hat{E}(\hat{\beta}_k)}$

Binary Choice

- More complicated tests done using one of 3 methods:
  1. Likelihood Ratio (LR) Test
     - Test statistic $LR = 2[ln\hat{L}_U - ln\hat{L}_R]$
     - $ln\hat{L}_R$ is log likelihood evaluated at restricted parameter vector
  2. Wald (W) Test
     - Test statistic $W = \hat{g}'[\hat{G}V\hat{a}r(\hat{\boldsymbol{\beta}})\hat{G}']\hat{g}$
     - $\hat{g}$ is a vector of restrictions evaluated at $\hat{\boldsymbol{\beta}}$
     - $\hat{G}$ is the derivative of a vector of restrictions evaluated at $\hat{\boldsymbol{\beta}}$
  3. Lagrange Multiplier (LM) Test
     - Test statistic $LM = \hat{d}'[V\hat{a}r(\hat{\boldsymbol{\beta}})]\hat{d}$
     - $\hat{d}$ is the derivative of lnL evaluated at restricted $\hat{\boldsymbol{\beta}}$
- Restrictions can be linear or non-linear

Binary Choice

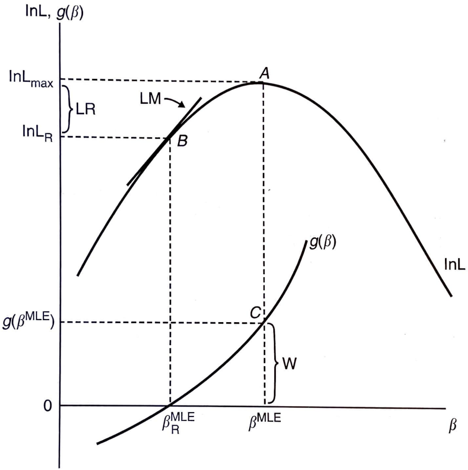


Figure: Source: Kennedy (2008)

Binary Choice

Goodness of Fit in Probit and Logit

1. Confusion Matrix

	Actual		
Predicted	0	1	Total
0	# Correct 0	# Incorrect 1	# Pred 0
1	# Incorrect 0	#Correct 1	# Pred 1
Total	# True 0	# True 1	

2. Pseudo-R^2
   - McFadden $\rightarrow R^2 = 1 - \frac{ln\hat{L}_U}{ln\hat{L}_0}$
     - $ln\hat{L}_0$ is log-likelihood with no explanatory variables
   - Others are possible, but goodness of fit is not incredibly important

Limited Dependent Variable Models

Introduction

- Arises when a continuous dependent variable is limited in its range
  - Censoring
    - Income is top-coded at some level for privacy
  - Corner Solutions
    - Spending on consumer durables limited below by 0
  - Incidental Truncation
    - Wage is not observed for people who do not work
- You can sometimes use OLS depending on context
- We will cover models for censoring and corner solutions

Limited Dependent Variable Models

Potential Outcomes

- We will again appeal to a latent variable model

$$y_0^* = \beta_0 - \eta$$

$$y_1^* = y_0^* + \beta_1$$

$$y^* = y_0^* + (y_1^* - y_0^*) * w$$

$$y^* = \beta_0 + \beta_1 * w - \eta$$

- In this case, the observed y is

$$y = max\{0, y^*\}$$

Limited Dependent Variable Models

- Because it is either zero or positive, the expected value of y is

$$E[y|w] = E[y|y > 0, w]Pr[y > 0|w]$$

- Taking the difference in the observed y, we get

$$E[y|w = 1] - E[y|w = 0]$$

$$= E[y|y > 0, w = 1]Pr[y > 0|w = 1] - E[y|y > 0, w = 0]Pr[y > 0|w = 0]$$

$$= (Pr[y > 0|w = 1] - Pr[y > 0|w = 1])E[y|y > 0, w = 1]$$

$$+(E[y|y > 0, w = 1] - E[y|y > 0, w = 0])Pr[y > 0|w = 0]$$

Limited Dependent Variable Models

- There are two key pieces
  - Participation effect
  - Conditional on Positive effect
- In terms of the potential outcomes

$$E[y|w=1] - E[y|w=0]$$

$$= E[y_1|w=1] - E[y_0|w=0]$$

$$= E[y_1|w=1] - E[y_0|w=1] + E[y_0|w=1] - E[y_0|w=0]$$

- The limited dependent variable does not change causal interpretation
  - As long as potential outcome y_0 is mean independent of treatment

Limited Dependent Variable Models

Conditional on Positive

- In some contexts people run regressions with just the positive outcomes
  - If you wanted to analyze participation decision separately
- Difference in observed y for this group is biased if under random assignment

$$E[y|y>0,w=1]-E[y|y>0,w=0]$$
$$=E[y_1|y_1>0]-E[y_0|y_0>0]$$
$$=E[y_1|y_1>0]-E[y_0|y_1>0]+E[y_0|y_1>0]-E[y_0|y_0>0]$$

- The treatment changes who has positive values of potential outcomes
  - More subtle form of bias
- You cannot interpret conditional on positive effects as causal

Limited Dependent Variable Models

Tobit Model

- This model is used in the context of censoring and corner solutions
- We have data on a random sample
- The outcome is limited in its range
- There is a mass of observations at 1 or more values
  - Usually zeroes
  - Sometimes some upper amount, like income
- Using OLS may be a bad strategy depending on your goals
  - Can produce predicted values outside the limited range

Limited Dependent Variable Models

- The Tobit model starts with the latent variable model

$$y^* = \mathbf{x}\boldsymbol{\beta} - \eta, \text{ where } \eta \sim N(0, \sigma^2)$$

$$y = \max(0, y^*)$$

- The conditional expectation of interest depends on context
  - Censored data
    - $E[y^*|\mathbf{x}]$
    - y^* usually has meaning when data are censored
  - Corner Solutions
    - $E[y|\mathbf{x}]$ and $E[y|y > 0, \mathbf{x}]$
    - y_i^* usually has no meaning for corner solutions
    - This is the most common situation

Limited Dependent Variable Models

- Estimate this model by maximum likelihood
- The likelihood function has two pieces
  - When $y_i = 0$

$$
\begin{aligned}
Pr(y_i = 0|\mathbf{x}) &= Pr(y_i^* < 0|\mathbf{x}) \\
&= Pr(\mathbf{x}\boldsymbol{\beta} - \eta < 0|\mathbf{x}) \\
&= Pr(\eta > \mathbf{x}\boldsymbol{\beta}|\mathbf{x}) \\
&= Pr(\frac{\eta}{\sigma_\eta} > \frac{\mathbf{x}\boldsymbol{\beta}}{\sigma_\eta}|\mathbf{x}) \\
&= 1 - \Phi\left(\frac{\mathbf{x}\boldsymbol{\beta}}{\sigma_\eta}\right)
\end{aligned}
$$

Limited Dependent Variable Models

- likelihood function continued...
  - When $y_i > 0$

$$f(y_i|y_i > 0, \mathbf{x}) = \frac{1}{\sigma_\eta}\phi\left(\frac{y_i - \mathbf{x}\boldsymbol{\beta}}{\sigma_\eta}\right)$$

  - Combining terms, we can form the likelihood function

$$\mathcal{L}(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}) = \Pi_{y_i=0}\left(1 - \Phi\left(\frac{\mathbf{x}\boldsymbol{\beta}}{\sigma_\eta}\right)\right)\Pi_{y_i>0}\left(\frac{1}{\sigma_\eta}\phi\left(\frac{y_i - \mathbf{x}\boldsymbol{\beta}}{\sigma_\eta}\right)\right)$$

  - The log likelihood is

$$ln\mathcal{L}(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}) = \sum_{y_i=0} ln\left(1 - \Phi\left(\frac{\mathbf{x}\boldsymbol{\beta}}{\sigma_\eta}\right)\right) + \sum_{y_i>0} ln\left(\frac{1}{\sigma_\eta}\phi\left(\frac{y_i - \mathbf{x}\boldsymbol{\beta}}{\sigma_\eta}\right)\right)$$

- Maximize this function by choosing the parameter vector $\boldsymbol{\beta}$ and σ_η

Limited Dependent Variable Models

- Marginal effects will depend on the context of our estimation
  - Censored data

$$\frac{\partial E[y^*|\mathbf{x}]}{\partial x_k} = \beta_k$$

    - In a Tobit with censored data, you can interpret the slope directly
  - Corner Solutions
  - $\frac{\partial E[y|\mathbf{x}]}{\partial x_k} = \Phi(\frac{\mathbf{x}\boldsymbol{\beta}}{\sigma})\beta_k$
  - $\frac{\partial E[y|y>0,\mathbf{x}]}{\partial x_k} = \{1 - \lambda(\frac{\mathbf{x}\boldsymbol{\beta}}{\sigma})[\frac{\mathbf{x}\boldsymbol{\beta}}{\sigma} + \lambda(\frac{\mathbf{x}\boldsymbol{\beta}}{\sigma})]\}\beta_k$
    - With corner solutions it depends on what you want
    - You may want slope for random person, or conditional on $y > 0$

Limited Dependent Variable Models

Issues with Tobit Model

1. It must be possible for the dependent variable to take values near the limit
   - Example: not the case with consumer durables
   - You either spend zero or a large amount
2. Intensive and Extensive margins have same parameters
   - Means the model is relatively inflexible
   - Can be solved by modelling each separately
3. Normality assumption
4. Care must be taken in interpreting the coefficients
   - Do we care about the effect of x_k on y or y^*?