Econometrics

Lecture 9

Nathaniel Higgins

JHU

New New material

- We have worked with continuous variables on the left and right side of the equals sign

$$y = \beta_0 + \beta_1 x_1 + u$$

- We have worked with continuous variables on the left side of the equals sign and discrete variables on the right side of the equals sign

$$y = \beta_0 + \beta_1 x_1 + \beta_2 d_1 + u$$

- What's next?

New New material

- We have worked with continuous variables on the left and right side of the equals sign

$$y = \beta_0 + \beta_1 x_1 + u$$

- We have worked with continuous variables on the left side of the equals sign and discrete variables on the right side of the equals sign

$$y = \beta_0 + \beta_1 x_1 + \beta_2 d_1 + u$$

- What's next?
- Discrete variables on the left side of the equals sign

Discrete dependent variables

- AKA *qualitative* dependent variables
- Examples
  - Durables purchases
    - Buying a car
    - Washing machine, refrigerator, dishwasher
  - Voting
    - Vote in a presidential election?
    - Vote yes or no on a particular referendum
  - Participation (in a program)
    - Unemployment insurance
    - School lunch
  - Self-selection into a treatment
    - Go to college
    - Technical training

Binary outcomes
Your own example

- I want you to tell a story about a binary outcome in your own life

Binary outcomes
Your own example

- I want you to tell a story about a binary outcome in your own life
- The binary outcome I want you to tell a story about is . . .

Binary outcomes
Your own example

- I want you to tell a story about a binary outcome in your own life
- The binary outcome I want you to tell a story about is . . .
- How you decided to attend JHU

Binary outcomes
Your own example

- I want you to tell a story about a binary outcome in your own life
- The binary outcome I want you to tell a story about is . . .
- How you decided to attend JHU
- We observe a binary outcome (you all get 1's in the variable *attend* in my dset), but the decision problem is . . . *richer* than what is indicated by a simple 1 or 0

Binary outcomes
Your own example

- I want you to tell a story about a binary outcome in your own life
- The binary outcome I want you to tell a story about is . . .
- How you decided to attend JHU
- We observe a binary outcome (you all get 1's in the variable *attend* in my dset), but the decision problem is . . . *richer* than what is indicated by a simple 1 or 0
- How to create a model of this decision?

Binary outcomes
Your own example

- I want you to tell a story about a binary outcome in your own life
- The binary outcome I want you to tell a story about is . . .
- How you decided to attend JHU
- We observe a binary outcome (you all get 1's in the variable *attend* in my dset), but the decision problem is . . . *richer* than what is indicated by a simple 1 or 0
- How to create a model of this decision?
- I want you to create a model in the form of a narrative
- How did you make the decision?
  - Did you just flip a coin?
  - What was underlying your decision to *attend* JHU vs. *not*

Binary dependent variables
Example

- An individual decides to attend JHU when *the utility of attending is higher than the utility of not attending*

Binary dependent variables
Example

- An individual decides to attend JHU when *the utility of attending is higher than the utility of not attending*
- If . . .
  (happiness if attend) - (cost of attending) >
  (happiness if not attend) - (cost of not attending)
  . . . then individual *i* attends JHU

Binary dependent variables
Example

- An individual decides to attend JHU when *the utility of attending is higher than the utility of not attending*
- If . . .
  (happiness if attend) - (cost of attending) >
  (happiness if not attend) - (cost of not attending)
  . . . then individual *i* attends JHU
- If not, then they don't

Binary dependent variables

- How do we use this idea to make a model?
- A person attends if their *net utility* of attending is positive (if attending is better than not attending)

Binary dependent variables

- How do we use this idea to make a model?
- A person attends if their *net utility* of attending is positive (if attending is better than not attending)
- Let's say that y^* (`yStar`) represents the net happiness of attending (the net happiness is a continuous variable)
- So,
  Net happiness $=$
  $U($ attending $) - U($ not $)$

Binary dependent variables

- How do we use this idea to make a model?
- A person attends if their *net utility* of attending is positive (if attending is better than not attending)
- Let's say that y^* (`yStar`) represents the net happiness of attending (the net happiness is a continuous variable)
- So,
  Net happiness $=$
  $U($ attending $) - U($ not $)$
- Let's say that x is data that we think influences net happiness

Binary dependent variables

- How do we use this idea to make a model?
- A person attends if their *net utility* of attending is positive (if attending is better than not attending)
- Let's say that y^* (`yStar`) represents the net happiness of attending (the net happiness is a continuous variable)
- So,
  Net happiness $=$
  $U($ attending $) - U($ not $)$
- Let's say that x is data that we think influences net happiness
- In terms of our variables:
  $y^* = \beta_0 + x\beta_1 + u$

Binbary dependent variables

- In terms of our variables
  $y^* = \beta_0 + x\beta_1 + u > 0$
- BUT we don't *observe* net happiness. Net happiness is not in our data.
- What *is* in our data:
  - A bunch of stuff that we think might influence net happiness (the X data)
  - Whether or not the person attended (y)
- So:
  $y = 1$ (individual attends) if $y^* > 0$
  $y = 0$ (individual does not attend) if $y^* < 0$
- Or:
  $y = 1$ if $u > -\beta_0 - x\beta_1$
  $y = 0$ if $u < -\beta_0 - x\beta_1$

Introduce binary data by contrast w/ continuous data.
So let's create some continuous data first.

R code

```
# Create a dset with 100 observations
# Generate an independent (RHS) variable x
from a uniform [0,1]
# Set seed

# Generate an independent (RHS) variable x
from a uniform [0,1]

# Specify parameters of the model (beta0 =
-2.5, beta1 = 5
```

Introduce binary data by contrast w/ continuous data.
So let's create some continuous data first.

R code

```
# Create a dset with 100 observations
# Generate an independent (RHS) variable x
from a uniform [0,1]
# Set seed
set.seed(12)
# Generate an independent (RHS) variable x
from a uniform [0,1]

# Specify parameters of the model (beta0 =
-2.5, beta1 = 5
```

Introduce binary data by contrast w/ continuous data.
So let's create some continuous data first.

R code

```
# Create a dset with 100 observations
# Generate an independent (RHS) variable x
from a uniform [0,1]
# Set seed
set.seed(12)
# Generate an independent (RHS) variable x
from a uniform [0,1]
x <- runif(100, min=0, max=1)
# Specify parameters of the model (beta0 =
-2.5, beta1 = 5
```

Introduce binary data by contrast w/ continuous data.
So let's create some continuous data first.

R code

```
# Create a dset with 100 observations
# Generate an independent (RHS) variable x
from a uniform [0,1]
# Set seed
set.seed(12)
# Generate an independent (RHS) variable x
from a uniform [0,1]
x <- runif(100, min=0, max=1)
# Specify parameters of the model (beta0 =
-2.5, beta1 = 5
beta0 <- -2.5
beta1 <- 5
```

Specify the model

R code

```
# Generate yStar = beta0 + x*beta1 + u
```

Specify the model

R code

```
# Generate yStar = beta0 + x*beta1 + u
u <- rnorm(100)
yStar <- beta0 + x*beta1 + u
```

Graph the results

R code

```
# Plot the data (yStar against x) and an OLS
prediction
```

Graph the results

R code

```
# Plot the data (yStar against x) and an OLS prediction
plot(x,yStar)
m1 <- lm(yStar   x)
lines(x,m1$fitted.values)
```

Create binary data corresponding to our story

R code

```
# Create y from yStar

# y = 1 if x*beta1 + u > 0

# Graph the new data
```

Binary dependent variables
vs. continuous data

Create binary data corresponding to our story

R code

```
# Create y from yStar
y <- 0
# y = 1 if x*beta1 + u > 0

# Graph the new data
```

Binary dependent variables
vs. continuous data

Create binary data corresponding to our story

R code

```
# Create y from yStar
y <- 0
# y = 1 if x*beta1 + u > 0
y[x*beta1 + u > 0] <- 1
# Graph the new data
```

Binary dependent variables
vs. continuous data

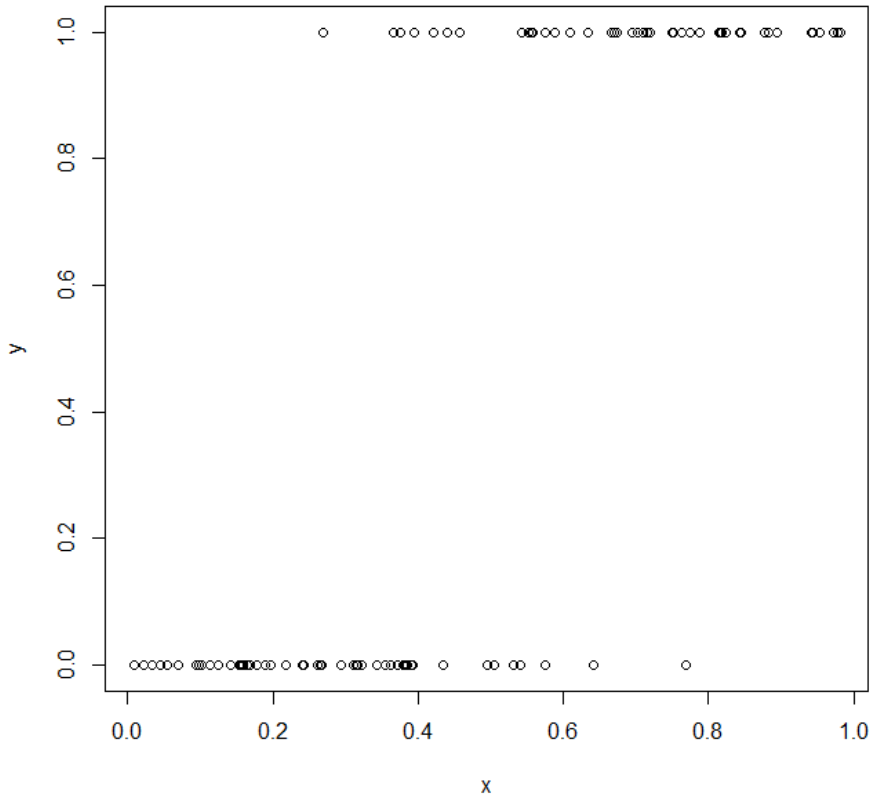
Create binary data corresponding to our story

R code

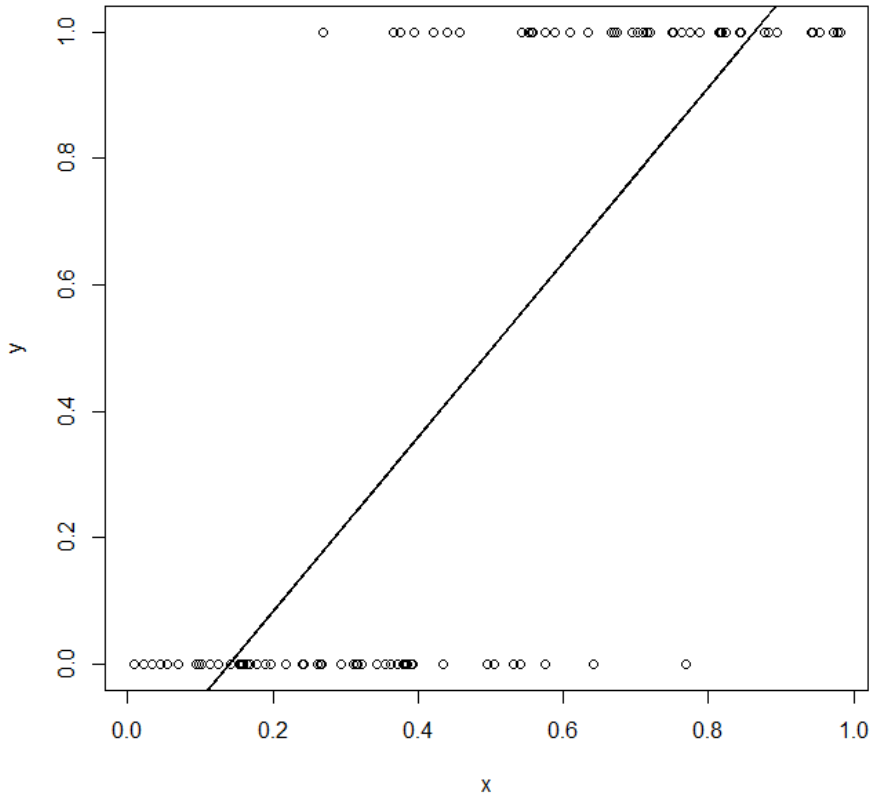
```
# Create y from yStar
y <- 0
# y = 1 if x*beta1 + u > 0
y[x*beta1 + u > 0] <- 1
# Graph the new data
plot(x,y)
```

Binary dependent variables
vs. continuous data

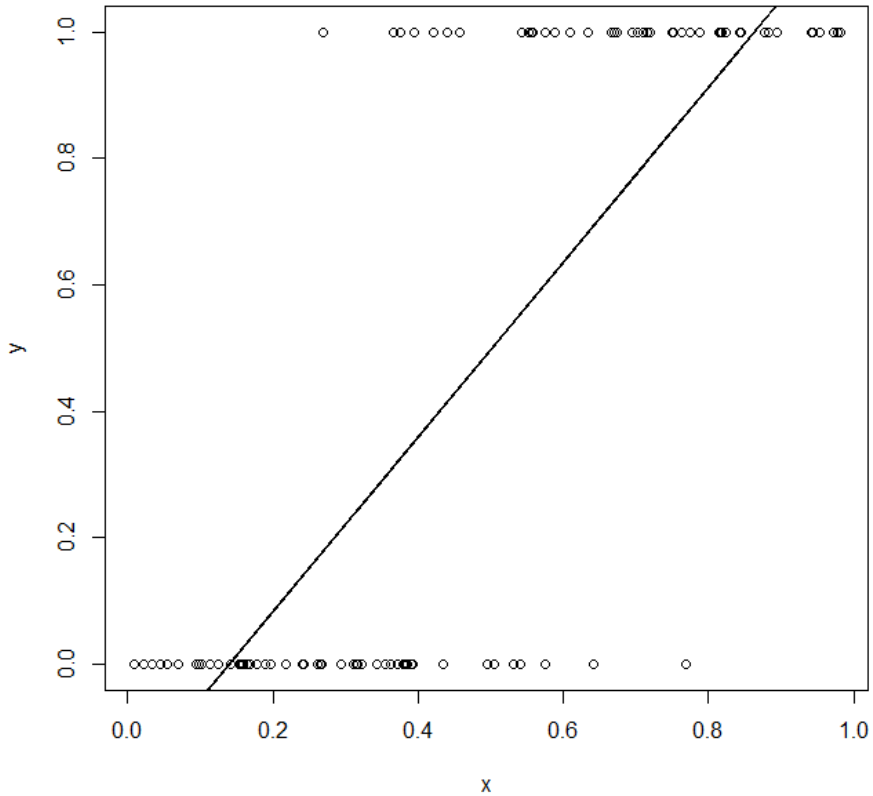
If we are dealing with dummy variables as dependent variables, then the data looks bunched-up



We can still fit a linear model to this data



We can still fit a linear model to this data



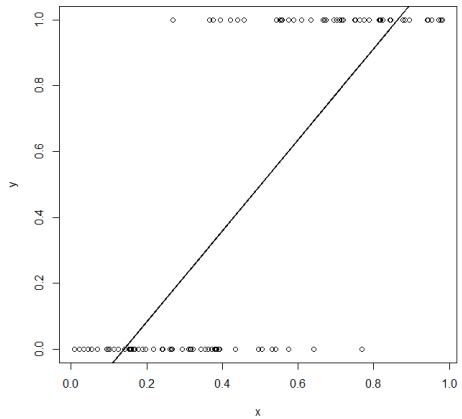
So what's the problem? Why not just use OLS?

- We *can* just use OLS. We call this the Linear Probability Model.

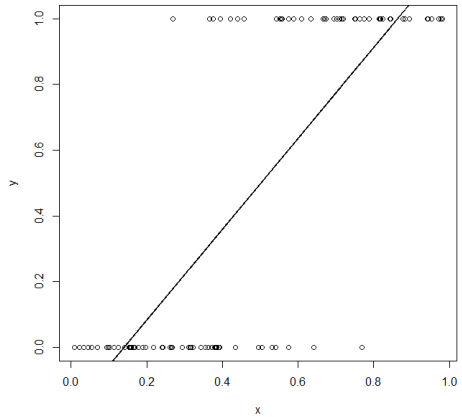
- We *can* just use OLS. We call this the Linear Probability Model.
- Virtues
  - Simple
  - Marginal effects (treatment effects) are easy to interpret

- We *can* just use OLS. We call this the Linear Probability Model.
- Virtues
  - Simple
  - Marginal effects (treatment effects) are easy to interpret
- Shortcomings
  - Predictions can be outside of $[0, 1]$ (i.e. crap)
  - Heteroskedasticity
  - Poor fit to the data by design
  - Constant marginal effects make a lot less sense in a probability model than they do in other cases

- Recall what heteroskedasticity is
  - Heteroskedasticity is when the variability in the errors of our model is not constant over x
  - Example of positive correlation: as x increases, variability of the model increase also
- Remember the graph of our linear model on made-up data?



When x tends to be small, what tends to be true about the errors?



Let's find out

R code

```
# Examine heteroskedasticity
# Fit a LPM to the data

# Get the errors; call them eLPM

# Plot the errors against x

# Does the variability of errors seem
constant?
```

R code

```
# Examine heteroskedasticity
# Fit a LPM to the data
m2 <- lm(y ~ x)
# Get the errors; call them eLPM

# Plot the errors against x

# Does the variability of errors seem
constant?
```

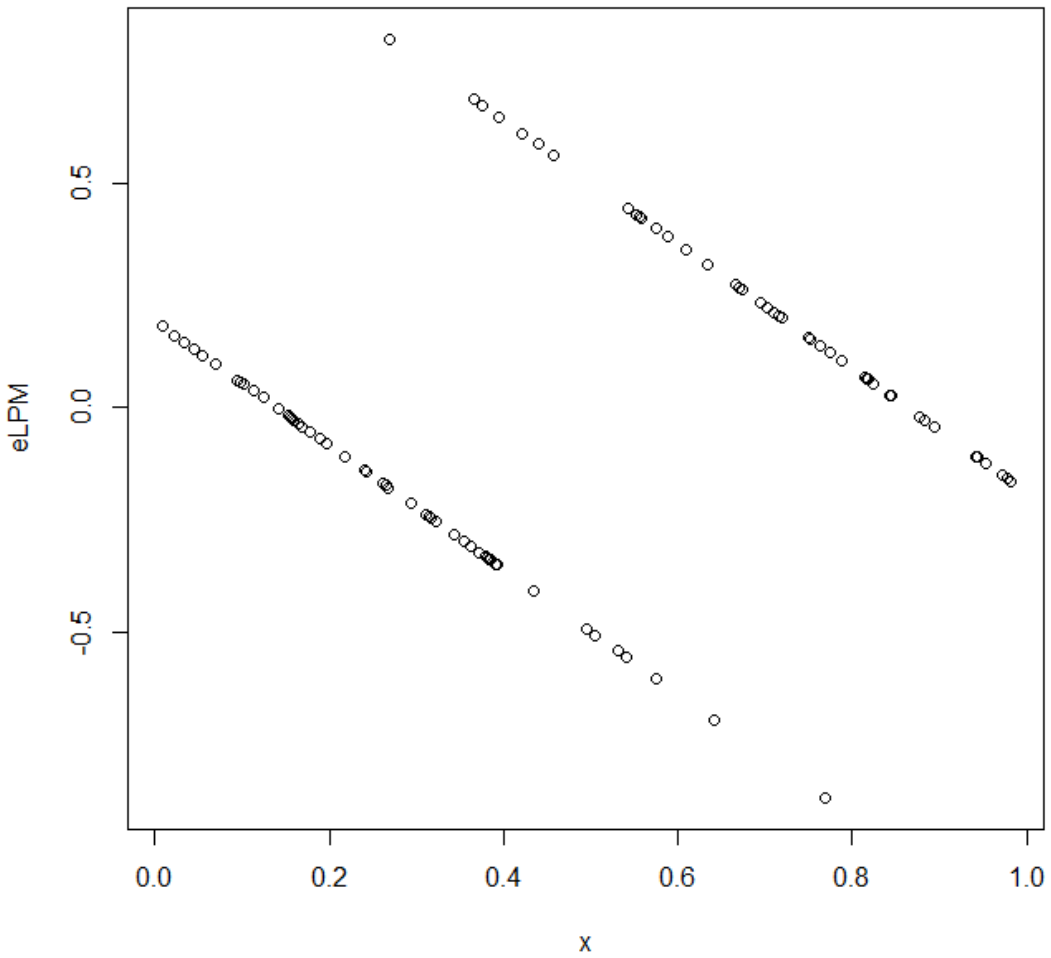
R code

```
# Examine heteroskedasticity
# Fit a LPM to the data
m2 <- lm(y ~ x)
# Get the errors; call them eLPM
eLPM <- m2$residuals
# Plot the errors against x

# Does the variability of errors seem
constant?
```

R code

```
# Examine heteroskedasticity
# Fit a LPM to the data
m2 <- lm(y ~ x)
# Get the errors; call them eLPM
eLPM <- m2$residuals
# Plot the errors against x
plot(x,eLPM)
# Does the variability of errors seem
constant?
```


eLPM
0.5
0.0
-0.5
0.0
0.2
0.4
0.6
0.8
1.0
x

R code

```
# Summarize the errors when x < 0.2

# Summarize the errors when 0.8 >= x > 0.2

# Summarize the errors when x >= 0.8
```

R code

```
# Summarize the errors when x < 0.2
sd(eLPM[x < 0.2])
# Summarize the errors when 0.8 >= x > 0.2
sd(eLPM[(x >= 0.2) | (x < 0.8)])
# Summarize the errors when x >= 0.8
sd(eLPM[x >= 0.8])
```

- Think about this graphically or mathematically — whichever you prefer
- There's a straightfoward way to think about this mathematically if you want

- Think about this graphically or mathematically — whichever you prefer
- There's a straightfoward way to think about this mathematically if you want
- Since y is either 1 or 0, u is either $1 - \beta_0 - x\beta_1$ or $0 - \beta_0 - x\beta_1$

- Think about this graphically or mathematically — whichever you prefer
- There's a straightfoward way to think about this mathematically if you want
- Since y is either 1 or 0, u is either $1 - \beta_0 - x\beta_1$ or $0 - \beta_0 - x\beta_1$
- What is the variance (in general)?

- Think about this graphically or mathematically — whichever you prefer
- There's a straightfoward way to think about this mathematically if you want
- Since y is either 1 or 0, u is either $1 - \beta_0 - x\beta_1$ or $0 - \beta_0 - x\beta_1$
- What is the variance (in general)? It is the squared expected deviation from the mean.

- Think about this graphically or mathematically — whichever you prefer
- There's a straightfoward way to think about this mathematically if you want
- Since y is either 1 or 0, u is either $1 - \beta_0 - x\beta_1$ or $0 - \beta_0 - x\beta_1$
- What is the variance (in general)? It is the squared expected deviation from the mean.
- The mean of u is . . .

- Think about this graphically or mathematically — whichever you prefer
- There's a straightfoward way to think about this mathematically if you want
- Since y is either 1 or 0, u is either $1 - \beta_0 - x\beta_1$ or $0 - \beta_0 - x\beta_1$
- What is the variance (in general)? It is the squared expected deviation from the mean.
- The mean of u is . . . 0. That hasn't changed.

- Think about this graphically or mathematically — whichever you prefer
- There's a straightfoward way to think about this mathematically if you want
- Since y is either 1 or 0, u is either $1 - \beta_0 - x\beta_1$ or $0 - \beta_0 - x\beta_1$
- What is the variance (in general)? It is the squared expected deviation from the mean.
- The mean of u is . . . 0. That hasn't changed.
- When $y = 1$, $u = 1 - \beta_0 - x\beta_1$; what is the probability that $y = 1$?

- Think about this graphically or mathematically — whichever you prefer
- There's a straightfoward way to think about this mathematically if you want
- Since y is either 1 or 0, u is either $1 - \beta_0 - x\beta_1$ or $0 - \beta_0 - x\beta_1$
- What is the variance (in general)? It is the squared expected deviation from the mean.
- The mean of u is . . . 0. That hasn't changed.
- When $y = 1$, $u = 1 - \beta_0 - x\beta_1$; what is the probability that $y = 1$? $\beta_0 + x\beta_1$

- Think about this graphically or mathematically — whichever you prefer
- There's a straightfoward way to think about this mathematically if you want
- Since y is either 1 or 0, u is either $1 - \beta_0 - x\beta_1$ or $0 - \beta_0 - x\beta_1$
- What is the variance (in general)? It is the squared expected deviation from the mean.
- The mean of u is . . . 0. That hasn't changed.
- When $y = 1$, $u = 1 - \beta_0 - x\beta_1$; what is the probability that $y = 1$? $\beta_0 + x\beta_1$
- When $y = 0$, $u = -\beta_0 - x\beta_1$; what is the probability that $y = 0$?

- Think about this graphically or mathematically — whichever you prefer
- There's a straightfoward way to think about this mathematically if you want
- Since y is either 1 or 0, u is either $1 - \beta_0 - x\beta_1$ or $0 - \beta_0 - x\beta_1$
- What is the variance (in general)? It is the squared expected deviation from the mean.
- The mean of u is . . . 0. That hasn't changed.
- When $y = 1$, $u = 1 - \beta_0 - x\beta_1$; what is the probability that $y = 1$? $\beta_0 + x\beta_1$
- When $y = 0$, $u = -\beta_0 - x\beta_1$; what is the probability that $y = 0$? $1 - \beta_0 - x\beta_1$

- The variance of u conditional on x is

$$Var(u|x) = (1 - \beta_0 - x\beta_1)^2 * \Pr(y = 1) + \\ +(-\beta_0 - x\beta_1)^2 * \Pr(y = 0)$$

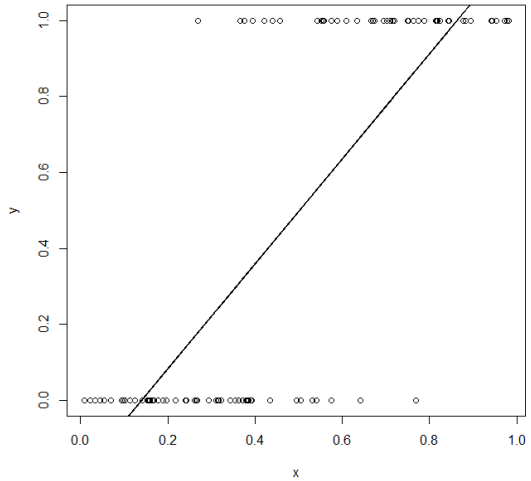
- Whenever the variance of u is anything but constant, we have heteroskedasticity
- So? Consequences:
  - We still get an unbaised estimate of our coefficients
  - We can no longer trust the standard errors
  - So a result that appears statistically significant might not be

- The variance of u conditional on x is

$$Var(u|x) = (1 - \beta_0 - x\beta_1)^2 * \Pr(y = 1) +$$
$$+(-\beta_0 - x\beta_1)^2 * \Pr(y = 0)$$
$$= (\beta_0 + x\beta_1)(1 - \beta_0 - x\beta)$$

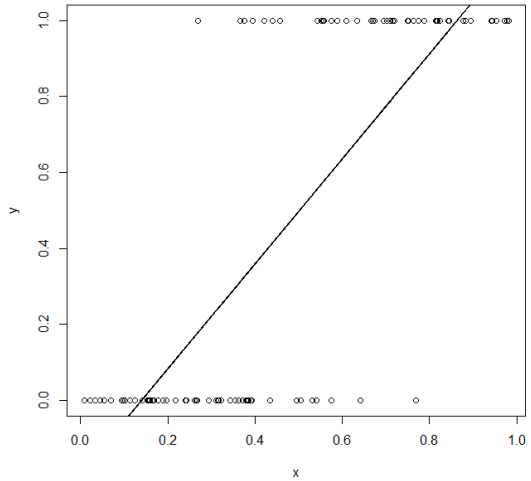
- Whenever the variance of u is anything but constant, we have heteroskedasticity
- So? Consequences:
  - We still get an unbaised estimate of our coefficients
  - We can no longer trust the standard errors
  - So a result that appears statistically significant might not be

The LPM doesn't "bend" (duh — it's linear).



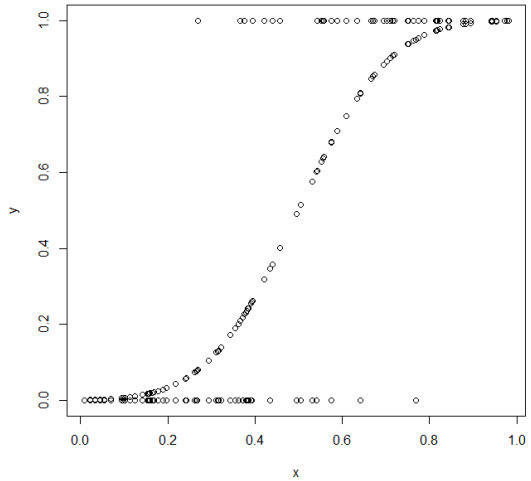
Sick of this graph yet?

The LPM doesn't "bend" (duh — it's linear).



Sick of this graph yet?
Takeaway? The model would fit the data a lot better if there was some bend to it.

I'm not going to tell you how I did it yet, but check this out.



Better.

OK. So if the LPM stinks so bad, where do we go from here?

- Let's put together a model using the data we have
- We observe whether an individual attends JHU or not
- We know this happens when $u > -\beta_0 - x\beta_1$

- Let's put together a model using the data we have
- We observe whether an individual attends JHU or not
- We know this happens when $u > -\beta_0 - x\beta_1$
- If we know the probability that a random draw u was greater than a particular value, then we could write down an equation in terms of:
  1. the data that we DO observe
  2. unknown parameters

- Let's put together a model using the data we have
- We observe whether an individual attends JHU or not
- We know this happens when $u > -\beta_0 - x\beta_1$
- If we know the probability that a random draw u was greater than a particular value, then we could write down an equation in terms of:
  1. the data that we DO observe
  2. unknown parameters
- If we know something about how u is distributed (or we assume something), then we have a model with a functional form

Binary dependent variables

- We know that we observe $y = 1$ for a particular individual whenever the u draw is greater than some function of our data (the x variables we observe) and unknown parameters (β_0 and β_1)
- What is the probability that a random variable u is less than a particular value?

Binary dependent variables

- We know that we observe $y = 1$ for a particular individual whenever the u draw is greater than some function of our data (the x variables we observe) and unknown parameters (β_0 and β_1)
- What is the probability that a random variable u is less than a particular value?
- It is the C.D.F. (statistics!) of the distribution of u
- We usually denote the C.D.F. by a capital F:

$$Pr(u < something) = F(something)$$

Binary dependent variables

- We know that we observe $y = 1$ for a particular individual whenever the u draw is greater than some function of our data (the x variables we observe) and unknown parameters (β_0 and β_1)
- What is the probability that a random variable u is less than a particular value?
- It is the C.D.F. (statistics!) of the distribution of u
- We usually denote the C.D.F. by a capital F:

$$Pr(u < something) = F(something)$$

- If we know the probability that a random variable is less than *something*, what is the probability that u is greater than that same *something*?

Binary dependent variables

- We know that we observe $y = 1$ for a particular individual whenever the u draw is greater than some function of our data (the x variables we observe) and unknown parameters (β_0 and β_1)
- What is the probability that a random variable u is less than a particular value?
- It is the C.D.F. (statistics!) of the distribution of u
- We usually denote the C.D.F. by a capital F:

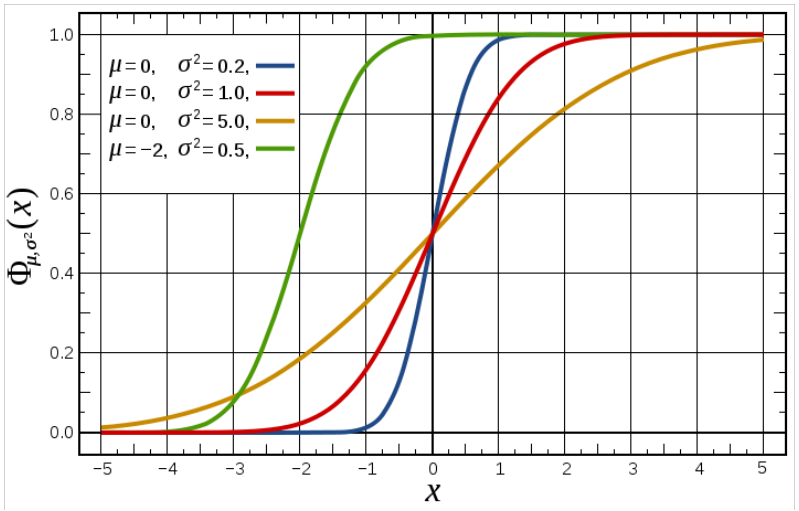
$$Pr(u < something) = F(something)$$

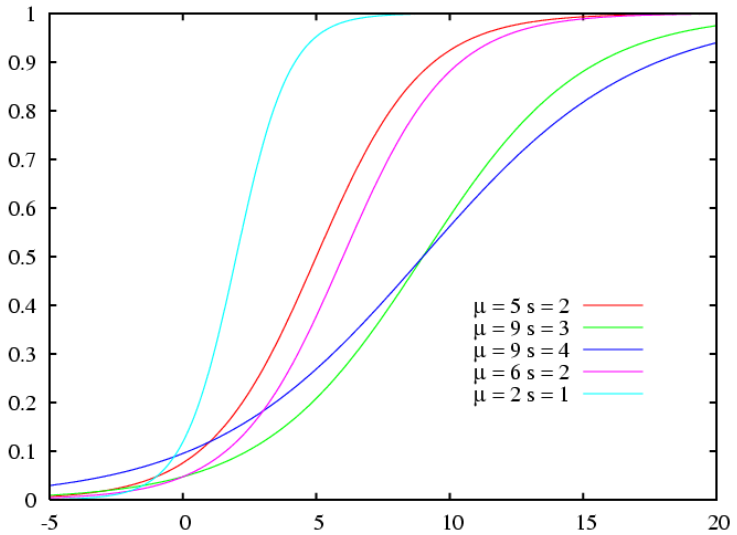
- If we know the probability that a random variable is less than *something*, what is the probability that u is greater than that same *something*?

$$Pr(u > something) = 1 - F(something)$$

Binary dependent variables

two *F*'s we use a lot





$$Pr(y = 1) = F(\beta_0 + x\beta_1)$$

- If u is distributed $N(0, 1)$, then we have a probit model

$$F(\beta_0 + x\beta_1) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{\beta_0 + x\beta_1}{2}}$$

- If u is distributed $N(0,1)$, then we have a probit model

$$F(\beta_0 + x\beta_1) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{\beta_0 + x\beta_1}{2}}$$

- if u is distributed as a logistic, then we have a logit model

$$F(\beta_0 + x\beta_1) = \frac{1}{1 + e^{-(\beta_0 + x\beta_1)}}$$

$$Pr(y = 1) = F(\beta_0 + x\beta_1)$$

- So how do we estimate the parameters of our model (the β's)?
- Well, in most cases we use *maximum likelihood*
- In short, we select the β parameters that are *most likely*, given the data we observe
- So, in terms of our model, our estimates of the parameters are the estimates that maximize the *likelihood function*

Maximum likelihood

- Four balls in a hat
- Pull balls out of the hat N times (generate N pieces of data)
- We get data: $(y_1, y_2, \ldots, y_N)$
- Suppose (just to have a concrete example) that the data looked like this:

$$(red, red, red, black, red, black, black, red, red, red)$$

- $N = 10$ (10 draws)
- Number of red balls drawn $= 7$
- Number of black balls drawn $= 3$
- We want to know how many red balls are in the hat (the number of red balls is equal to β, the parameter to be estimated)

Maximum likelihood

- If there had been one red ball and three black balls in the hat:
- The probability of observing the data we observed would be:

$$(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{3}{4}, \frac{1}{4}, \frac{3}{4}, \frac{3}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}) \quad (1)$$

- Or, more compactly:

$$\left(\frac{1}{4}\right)^7 \times \left(\frac{3}{4}\right)^3$$

- This is the probability of the first draw being a red ball, the second draw being a red ball, etc., so that we get *exactly* the draws in the exact order that we got them in (1) above
- (Note that this is not quite the same thing as the probability of getting 7 red balls and 3 black balls)

Maximum likelihood

- If, instead, if there had been two red balls and two black balls in the hat:
- The probability of observing

$$(\textit{red}, \textit{red}, \textit{red}, \textit{black}, \textit{red}, \textit{black}, \textit{black}, \textit{red}, \textit{red}, \textit{red})$$

would be:

$$(\frac{2}{4}, \frac{2}{4}, \frac{2}{4}, \frac{2}{4}, \frac{2}{4}, \frac{2}{4}, \frac{2}{4}, \frac{2}{4}, \frac{2}{4}, \frac{2}{4})$$

or

$$\left(\frac{2}{4}\right)^7 \times \left(\frac{2}{4}\right)^3$$

Maximum likelihood

- If there were three red balls and one black ball in the hat then the probability of getting the data we observed would be:

$$(\frac{3}{4}, \frac{3}{4}, \frac{3}{4}, \frac{1}{4}, \frac{3}{4}, \frac{1}{4}, \frac{1}{4}, \frac{3}{4}, \frac{3}{4}, \frac{3}{4})$$

or

$$\left(\frac{3}{4}\right)^7 \times \left(\frac{1}{4}\right)^3$$

Maximum likelihood

- Compare the probability of observing

$$(red, red, red, black, red, black, black, red, red, red)$$

- If there had been one red ball in the hat:

$$\left(\frac{1}{4}\right)^7 \times \left(\frac{3}{4}\right)^3 = 0.0000257492$$

- If there had been two red balls in the hat:

$$\left(\frac{2}{4}\right)^7 \times \left(\frac{2}{4}\right)^3 = 0.0009765625$$

- Or if there had been three red balls in the hat:

$$\left(\frac{3}{4}\right)^7 \times \left(\frac{1}{4}\right)^3 = 0.002085686$$

Maximum likelihood

- If there had only been $\beta = 1$ red ball in the hat, the chance of observing the data that we did is about 0.003%
- If there had been $\beta = 2$ red balls in the hat, the chance of observing the data that we did is approximately 0.1%, i.e. a little better
- Finally, if there had been $\beta = 3$ red balls in the hat, the chance of observing the data that we did is about 0.2%, twice as likely as if there had been two red balls!

Maximum likelihood

- What is our *maximum likelihood* estimate of β?
- Of the choices 1, 2, or 3 red balls, which one is most likely to have produced the data we observed?
- We pick the choice that results in the highest likelihood

$$\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{3}{4}, \frac{1}{4}, \frac{3}{4}, \frac{3}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right) = 0.0000257492$$

$$\left(\frac{2}{4}, \frac{2}{4}, \frac{2}{4}, \frac{2}{4}, \frac{2}{4}, \frac{2}{4}, \frac{2}{4}, \frac{2}{4}, \frac{2}{4}, \frac{2}{4}\right) = 0.0009765625$$

$$\left(\frac{3}{4}, \frac{3}{4}, \frac{3}{4}, \frac{1}{4}, \frac{3}{4}, \frac{1}{4}, \frac{1}{4}, \frac{3}{4}, \frac{3}{4}, \frac{3}{4}\right) = 0.002085686$$

- We choose $\widehat{\beta}^{MLE} = 3$

- OK. Leave the balls-in-a-hat example behind.
- Speaking generally:
- The maximum likelihood concept is very "portable" — the basic concept works the same in all cases
- The question we seek to answer is always the same: What βs make the data that we observed the most likely?
- We select the parameters β that are most likely to have generated the data we observe $(y_1, y_2, \ldots, y_N)$ and call those βs our estimates $\widehat{\beta}^{MLE}$
- To do this we only need to be able to form the likelihood function

- The likelihood function always takes the same basic form
- To form the likelihood we need to be able to express the probability of each "event," or piece of data y_i occurring
- We need to be able to write down the probability that each observation turns out as it does, that: $Y_i = y_i$

- We write this probability in terms of our data and call it $p(y_i|x_i, \beta)$
  - This is what we did in the balls-in-a-hat example when we wrote down the probability that a given draw $y_i = 1$ (red ball) $= \frac{\beta}{4}$ and the probability that a given draw $y_i = 0$ (black ball) $= 1 - \frac{\beta}{4}$
  - Those two things were expressions of $p(y_i = 1|x_i, \beta)$ and $p(y_i = 0|x_i, \beta)$, respectively

- The likelihood function always takes the same basic form
- To form the likelihood we need to be able to express the probability of each "event," or piece of data y_i occurring
- We need to be able to write down the probability that each observation turns out as it does, that: $Y_i = y_i$
- We write this probability in terms of our data and call it $p(y_i|x_i, \beta)$
  - This is what we did in the balls-in-a-hat example when we wrote down the probability that a given draw $y_i = 1$ (red ball) $= \frac{\beta}{4}$ and the probability that a given draw $y_i = 0$ (black ball) $= 1 - \frac{\beta}{4}$
  - Those two things were expressions of $p(y_i = 1|x_i, \beta)$ and $p(y_i = 0|x_i, \beta)$, respectively

- We write down each and every likelihood function in exactly the same way:

$$L(y_1, \ldots, y_N) = \prod_{i=1}^{N} p(y_i | x_i, \beta)$$

- How to write the likelihood function for a binary choice model?
- How do we write down the probability of each occurrence in a binary model?

$$L(y_1, \ldots, y_N) = \prod_{i=1}^{N} p(y_i | x_i, \beta)$$

- Start with $p(y_i | x_i, \beta)$ — what are the probabilities that we'd like to express?

The likelihood equation — a rough explanation with our model

- The probability of a 1 (attending JHU) is $F(\beta_0 + x_i\beta_1)$
- The probability of a 0 (not attending JHU) is $1 - F(\beta_0 + x_i\beta_1)$

The likelihood equation — a rough explanation with our model

- The probability of a 1 (attending JHU) is $F(\beta_0 + x_i\beta_1)$
- The probability of a 0 (not attending JHU) is $1 - F(\beta_0 + x_i\beta_1)$
- Say we had two observations: we observed one person who attended $y_1 = 1$ and one person who did not $y_2 = 0$

The likelihood equation — a rough explanation with our model

- The probability of a 1 (attending JHU) is $F(\beta_0 + x_i\beta_1)$
- The probability of a 0 (not attending JHU) is $1 - F(\beta_0 + x_i\beta_1)$
- Say we had two observations: we observed one person who attended $y_1 = 1$ and one person who did not $y_2 = 0$
- What is the likelihood of our data? Hint: use the *exact* same procedure you did with the balls in the hat

The likelihood equation — a rough explanation with our model

- The probability of a 1 (attending JHU) is $F(\beta_0 + x_i\beta_1)$
- The probability of a 0 (not attending JHU) is $1 - F(\beta_0 + x_i\beta_1)$
- Say we had two observations: we observed one person who attended $y_1 = 1$ and one person who did not $y_2 = 0$
- What is the likelihood of our data? Hint: use the *exact* same procedure you did with the balls in the hat
- Probability of observing y_1 times the probability of observing y_2 gives us the *likelihood* of our data:

The likelihood equation — a rough explanation with our model

- The probability of a 1 (attending JHU) is $F(\beta_0 + x_i\beta_1)$
- The probability of a 0 (not attending JHU) is $1 - F(\beta_0 + x_i\beta_1)$
- Say we had two observations: we observed one person who attended $y_1 = 1$ and one person who did not $y_2 = 0$
- What is the likelihood of our data? Hint: use the *exact* same procedure you did with the balls in the hat
- Probability of observing y_1 times the probability of observing y_2 gives us the *likelihood* of our data:

$$L = F(\beta_0 + x_1\beta_1)(1 - F(\beta_0 + x_2\beta_1))$$

The likelihood equation — a rough explanation with our model

- The probability of a 1 (attending JHU) is $F(\beta_0 + x_i\beta_1)$
- The probability of a 0 (not attending JHU) is $1 - F(\beta_0 + x_i\beta_1)$
- Say we had two observations: we observed one person who attended $y_1 = 1$ and one person who did not $y_2 = 0$
- What is the likelihood of our data? Hint: use the *exact* same procedure you did with the balls in the hat
- Probability of observing y_1 times the probability of observing y_2 gives us the *likelihood* of our data:

$$L = F(\beta_0 + x_1\beta_1)(1 - F(\beta_0 + x_2\beta_1))$$

- In general, the likelihood of a whole bunch of data is:

The likelihood equation — a rough explanation with our model

- The probability of a 1 (attending JHU) is $F(\beta_0 + x_i\beta_1)$
- The probability of a 0 (not attending JHU) is $1 - F(\beta_0 + x_i\beta_1)$
- Say we had two observations: we observed one person who attended $y_1 = 1$ and one person who did not $y_2 = 0$
- What is the likelihood of our data? Hint: use the *exact* same procedure you did with the balls in the hat
- Probability of observing y_1 times the probability of observing y_2 gives us the *likelihood* of our data:

$$L = F(\beta_0 + x_1\beta_1)(1 - F(\beta_0 + x_2\beta_1))$$

- In general, the likelihood of a whole bunch of data is:

$$L = \prod_i F(\beta_0 + x_i\beta_1) \prod_j (1 - F(\beta_0 + x_j\beta_1))$$

Remember what we are trying to do: we are trying to figure out how to choose $\hat{\beta}$

- We select the $\hat{\beta}$ that maximized the *likelihood* of our data
- That's as technical as we'll get about it in this class
- R (or Stata or SAS, etc.) find the coefficients that maximize the likelihood function for us
- Let's try it

Let's try a probit / logit model with the synthetic data

R code

```
# First, calculate the LPM again

# Get the predictions ("yhatOLS")

# Calculate probit model

# Get the predictions ("yhatProbit")
```

Let's try a probit / logit model with the synthetic data

R code

```
# First, calculate the LPM again
m2 <- lm(y ~ x)
# Get the predictions ("yhatOLS")

# Calculate probit model

# Get the predictions ("yhatProbit")
```

Let's try a probit / logit model with the synthetic data

R code

```
# First, calculate the LPM again
m2 <- lm(y ~ x)
# Get the predictions ("yhatOLS")
yhatOLS <- m2$fitted.values
# Calculate probit model

# Get the predictions ("yhatProbit")
```

Let's try a probit / logit model with the synthetic data

R code

```
# First, calculate the LPM again
m2 <- lm(y ~ x)
# Get the predictions ("yhatOLS")
yhatOLS <- m2$fitted.values
# Calculate probit model
m3 <- glm(y ~ x, family=binomial(link=probit))
# Get the predictions ("yhatProbit")
```

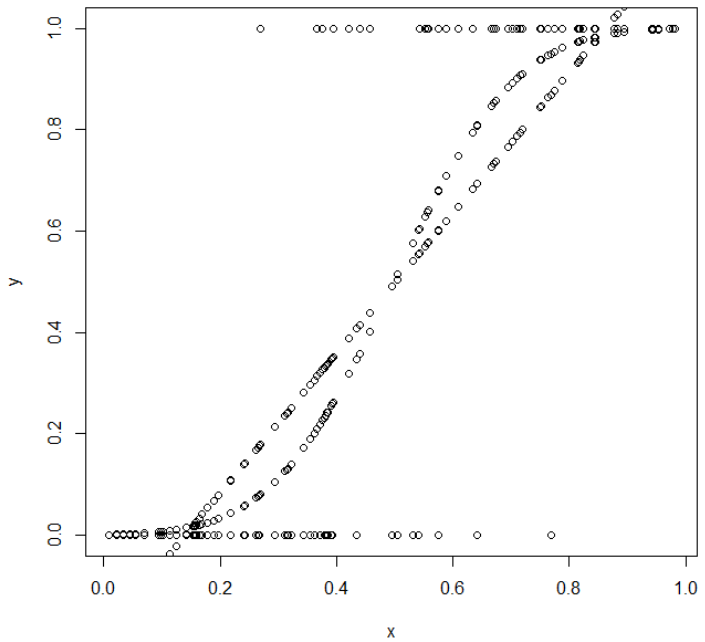
Let's try a probit / logit model with the synthetic data

R code

```
# First, calculate the LPM again
m2 <- lm(y ~ x)
# Get the predictions ("yhatOLS")
yhatOLS <- m2$fitted.values
# Calculate probit model
m3 <- glm(y ~ x, family=binomial(link=probit))
# Get the predictions ("yhatProbit")
yhatProbit <- m3$fitted.values
```

Now graph it. Scatter the true y values against x, the LPM $\hat{y}$ against x, and the Probit $\hat{y}$ against x.

Now graph it. Scatter the true y values against x, the LPM $\hat{y}$ against x, and the Probit $\hat{y}$ against x.



Let's try a probit / logit model with the Mroz data

R code

```
# Load the Mroz data


# First, calculate the LPM

# Get the predictions ("yhatOLS")

# Calculate probit model


# Get the predictions ("yhatProbit")
```

Let's try a probit / logit model with the Mroz data

R code

```
# Load the Mroz data
mroz <- read.csv("Labor-supply-Mroz-1987.csv",
stringsAsFactors=F)
# First, calculate the LPM

# Get the predictions ("yhatOLS")

# Calculate probit model


# Get the predictions ("yhatProbit")
```

Let's try a probit / logit model with the Mroz data

R code

```
# Load the Mroz data
mroz <- read.csv("Labor-supply-Mroz-1987.csv",
stringsAsFactors=F)
# First, calculate the LPM
m4 <- lm(lfp ~ wa + wa2 + we, data=mroz)
# Get the predictions ("yhatOLS")

# Calculate probit model



# Get the predictions ("yhatProbit")
```

Let's try a probit / logit model with the Mroz data

R code

```
# Load the Mroz data
mroz <- read.csv("Labor-supply-Mroz-1987.csv",
stringsAsFactors=F)
# First, calculate the LPM
m4 <- lm(lfp ~ wa + wa2 + we, data=mroz)
# Get the predictions ("yhatOLS")
yhatOLS <- m4$fitted.values
# Calculate probit model



# Get the predictions ("yhatProbit")
```

Let's try a probit / logit model with the Mroz data

R code

```
# Load the Mroz data
mroz <- read.csv("Labor-supply-Mroz-1987.csv",
stringsAsFactors=F)
# First, calculate the LPM
m4 <- lm(lfp ~ wa + wa2 + we, data=mroz)
# Get the predictions ("yhatOLS")
yhatOLS <- m4$fitted.values
# Calculate probit model
m5 <- glm(lfp ~ wa + wa2 + we, data=mroz,
family=binomial(link=probit))
# Get the predictions ("yhatProbit")
```

Let's try a probit / logit model with the Mroz data

R code

```
# Load the Mroz data
mroz <- read.csv("Labor-supply-Mroz-1987.csv",
stringsAsFactors=F)
# First, calculate the LPM
m4 <- lm(lfp ~ wa + wa2 + we, data=mroz)
# Get the predictions ("yhatOLS")
yhatOLS <- m4$fitted.values
# Calculate probit model
m5 <- glm(lfp ~ wa + wa2 + we, data=mroz,
family=binomial(link=probit))
# Get the predictions ("yhatProbit")
yhatProbit <- m5$fitted.values
```

- Look at the regression results
- Compare the coefficients of the LPM to the coefficients of the Probit model
- What do you notice?

- Pay attention especially to the coefficient on `we`

- Look at the regression results
- Compare the coefficients of the LPM to the coefficients of the Probit model
- What do you notice?
- You should notice that they look quite a bit different
- Pay attention especially to the coefficient on `we`

- Look at the regression results
- Compare the coefficients of the LPM to the coefficients of the Probit model
- What do you notice?
- You should notice that they look quite a bit different
- Pay attention especially to the coefficient on `we`
- Interpret the coefficient on `we` in the LPM
- Interpret the coefficient on `we` in the Probit model

- What is the marginal effect of `we` in a probit model?
- Well, it's not really fair to ask you, because the real functional form of a probit is pretty complicated, and you haven't seen it written-out yet
- But let's just say that it is *some* function G
- So $y = G(we\beta_{we} + x\beta_{other})$
- What is the marginal effect?

- What is the marginal effect of `we` in a probit model?
- Well, it's not really fair to ask you, because the real functional form of a probit is pretty complicated, and you haven't seen it written-out yet
- But let's just say that it is *some* function G
- So $y = G(we\beta_{we} + x\beta_{other})$
- What is the marginal effect?
- $\frac{dy}{dwe} = \frac{dG}{dwe}\beta_{we}$

- What is the marginal effect of `we` in a probit model?
- Well, it's not really fair to ask you, because the real functional form of a probit is pretty complicated, and you haven't seen it written-out yet
- But let's just say that it is *some* function G
- So $y = G(we\beta_{we} + x\beta_{other})$
- What is the marginal effect?
- $\frac{dy}{dwe} = \frac{dG}{dwe}\beta_{we}$
- Punchline?
- The coefficient β_{we} is no longer the marginal effect

Binary dependent variables
Model interpretation

- So how do we interpet the coefficient?

Binary dependent variables
Model interpretation

- So how do we interpet the coefficient?
- We don't!

Binary dependent variables
Model interpretation

- So how do we interpet the coefficient?
- We don't!
- We still recognize the t-stat on the coefficient as representing the statistical significance of the independent variable on the dependent variable (so causal interpretation is sitll there)

- So how do we interpet the coefficient?
- We don't!
- We still recognize the t-stat on the coefficient as representing the statistical significance of the independent variable on the dependent variable (so causal interpretation is sitll there)
- But if we want to understand the *marginal effect*, we must obtain $\frac{dG}{dwe}\beta_{we}$
- How? If you guessed that "R does it for me," then you're right. 10 points for you.

The choice between probit and logit will not usually be consequential.

1. Models are similar (see the figures on previous pages)
2. In practice, models are equally easy to run in R (or any other software)
3. Often we will run both $\rightarrow$ compare coefficents $\rightarrow$ if they are not very different, just pick a model (flip a coin) and make a footnote about the similarity of the models

Homework

- Just a few problems on this new stuff (you will need data that I will send to you)
- Continue making progress on your replication project
- You now know how to collect data and how to merge it on to the data set you have already assembled
- Collect all of your data! (not by next time)
- If you have your data set collected by the end of the month (we have class on 30 April) you will have plenty of time to run regressions and interpret results!