# Applied Econometrics

## Lecture 3

Nathaniel Higgins

ERS and JHU

20 September 2010

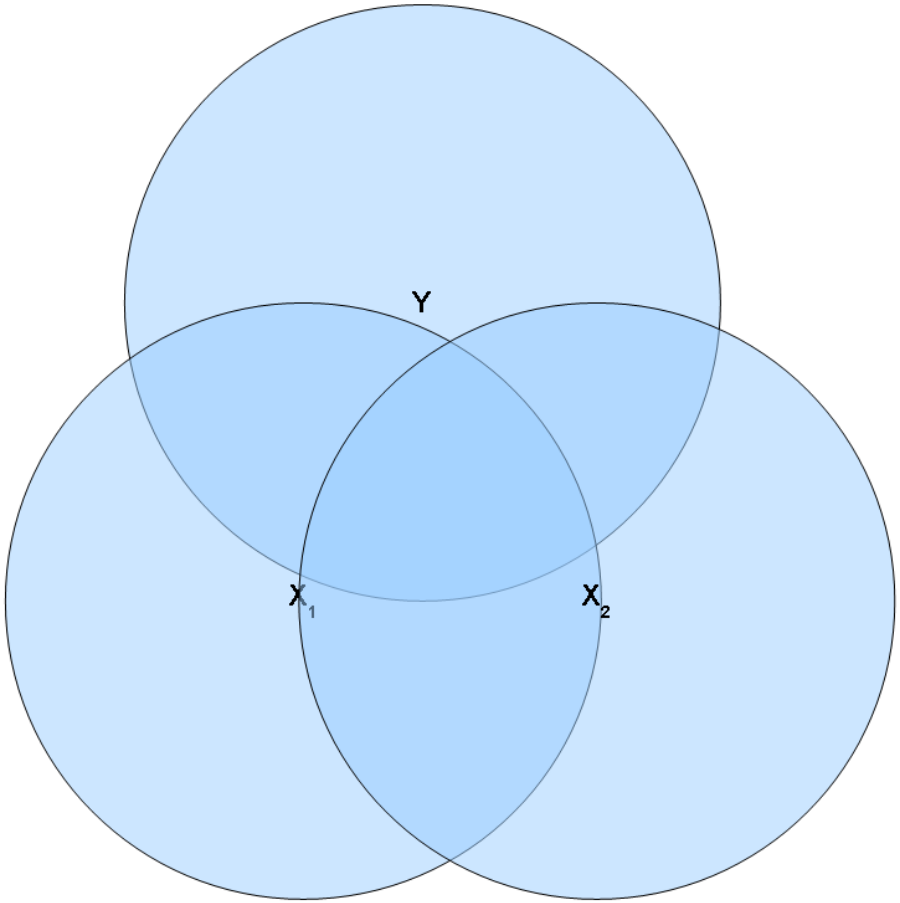
# Outline of today's lecture

- Schedule and Due Dates
- Making OLS make sense
  - Uncorrelated X's
  - Correlated X's
  - Omitted variable bias
- Endogeneity in general
- What are Instrumental Variables?

# Major due dates
## Reminder

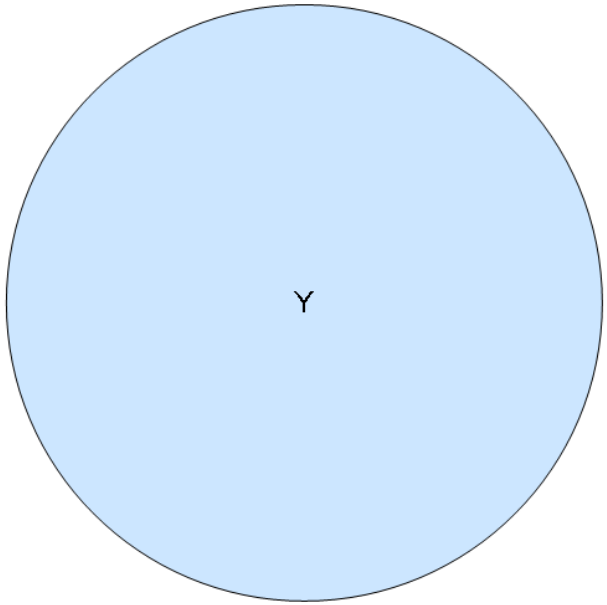
- Replication project: 4a on 2 November
- Final project: Day of final (9-15 December)

# OLS estimator
## Graphically

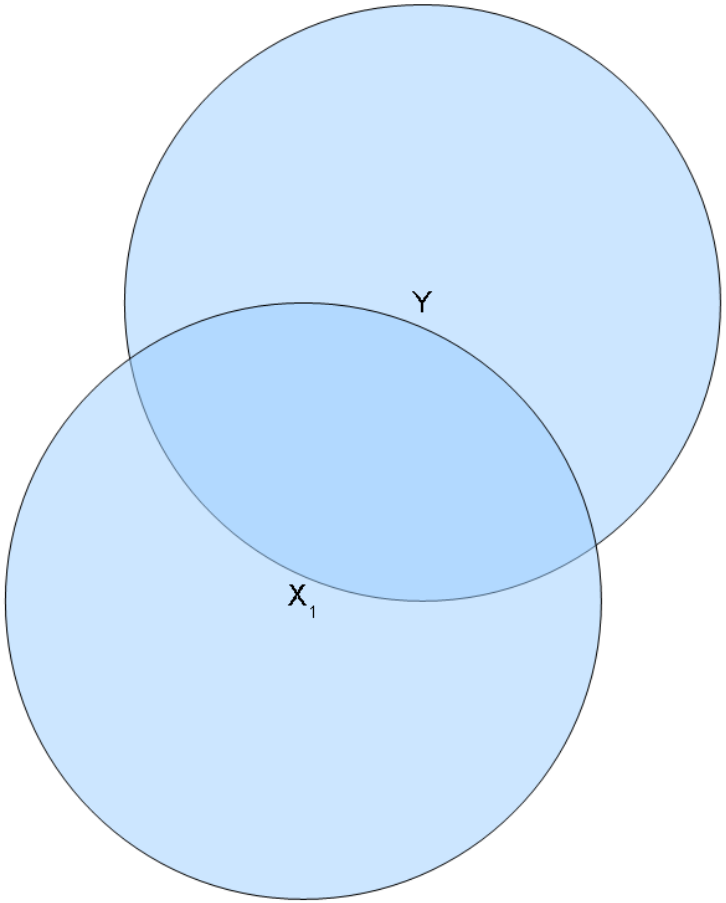


# OLS estimator
## Graphically

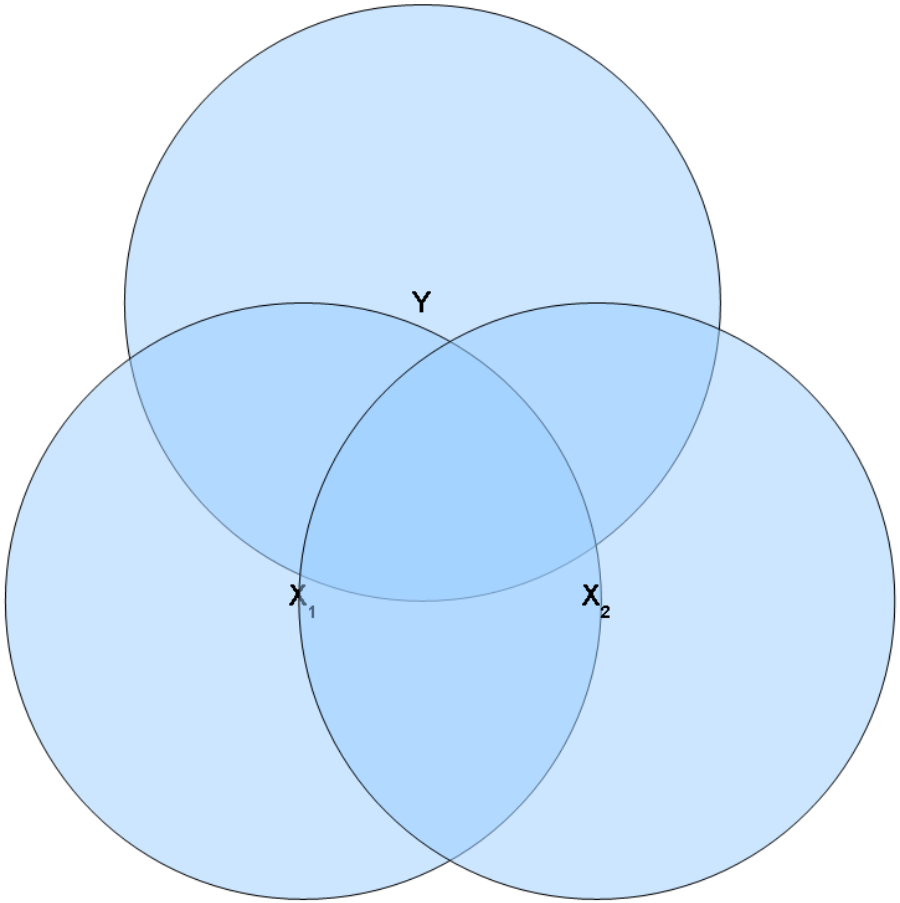
# OLS estimator
## Graphically



# OLS estimator
## Graphically



# OLS estimator
## Graphically



# OLS estimator
## Try it

### Stata code

```
* Set seed
set seed 12345
* Create a matrix of correlations
matrix C = (1, 0.2, 0.2 \ 0.2, 1, 0.2 \ ///
0.2, 0.2, 1)
* Create a matrix of means
matrix m = (3,2,2)
* Create a matrix of standard deviations
matrix sd = (0.5,2,1)
* Draw three random variable from a
* multivariate distribution
drawnorm x1 x2 x3, n(100) means(m) ///
sds(sd) corr(C)
* Draw some "unobservable" stuff
gen eps = rnormal()
```

# OLS estimator
## Try it

### Stata code

```
* Create a dependent variable y
gen y = 5 + 2*x1 - 3*x2 + eps
* Regress y on x1 (by itself)
regress y x1
```

# OLS estimator
## Try it

### Stata code

```
* Create a dependent variable y
gen y = 5 + 2*x1 - 3*x2 + eps
* Regress y on x1 (by itself)
regress y x1
* Regress y on x2 (by itself)
regress y x2
```

# OLS estimator
## Try it

### Stata code

```
* Create a dependent variable y
gen y = 5 + 2*x1 - 3*x2 + eps
* Regress y on x1 (by itself)
regress y x1
* Regress y on x2 (by itself)
regress y x2
* Regress y on x1 and x2
regress y x1 x2
```

# Omitted variables

- When we regress $y$ on $x_1$ alone, what happens?

# Omitted variables

- When we regress $y$ on $x_1$ alone, what happens?

```
      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  1,    98) =    0.00
       Model |  .096434032     1  .096434032           Prob > F      =  0.9596
    Residual |  3655.80887    98  37.3041721           R-squared     =  0.0000
-------------+------------------------------           Adj R-squared = -0.0102
       Total |   3655.9053    99  36.9283364           Root MSE      =  6.1077

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |  -.0682766   1.342874    -0.05   0.960    -2.733166    2.596613
       _cons |   5.790433   4.098127     1.41   0.161    -2.342166    13.92303
------------------------------------------------------------------------------
```

# Omitted variables

- When we regress $y$ on $x_1$ alone, what happens?

```
      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  1,    98) =    0.00
       Model |  .096434032     1  .096434032           Prob > F      =  0.9596
    Residual |  3655.80887    98  37.3041721           R-squared     =  0.0000
-------------+------------------------------           Adj R-squared = -0.0102
       Total |   3655.9053    99  36.9283364           Root MSE      =  6.1077

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |  -.0682766   1.342874    -0.05   0.960    -2.733166    2.596613
       _cons |   5.790433   4.098127     1.41   0.161    -2.342166    13.92303
------------------------------------------------------------------------------
```

- When we regress $y$ on $x_2$ alone, what happens?

# Omitted variables

- When we regress $y$ on $x_1$ alone, what happens?

```
      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  1,    98) =    0.00
       Model |  .096434032     1  .096434032           Prob > F      =  0.9596
    Residual |  3655.80887    98  37.3041721           R-squared     =  0.0000
-------------+------------------------------           Adj R-squared = -0.0102
       Total |   3655.9053    99  36.9283364           Root MSE      =  6.1077

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |  -.0682766   1.342874    -0.05   0.960    -2.733166    2.596613
       _cons |   5.790433   4.098127     1.41   0.161    -2.342166    13.92303
------------------------------------------------------------------------------
```

- When we regress $y$ on $x_2$ alone, what happens?

```
      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  1,    98) = 2033.12
       Model |  3487.78809     1  3487.78809           Prob > F      =  0.0000
    Residual |  168.117208    98  1.71548171           R-squared     =  0.9540
-------------+------------------------------           Adj R-squared =  0.9535
       Total |   3655.9053    99  36.9283364           Root MSE      =  1.3098

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x2 |  -2.894991   .0642045   -45.09   0.000    -3.022403    -2.76758
       _cons |   10.83503   .1752564    61.82   0.000     10.48724    11.18282
------------------------------------------------------------------------------
```

# Omitted variables

- When we regress $y$ on $x_1$ alone, what happens?

```
      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  1,    98) =    0.00
       Model |  .096434032     1  .096434032           Prob > F      =  0.9596
    Residual |  3655.80887    98  37.3041721           R-squared     =  0.0000
-------------+------------------------------           Adj R-squared = -0.0102
       Total |   3655.9053    99  36.9283364           Root MSE      =  6.1077

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |  -.0682766   1.342874    -0.05   0.960    -2.733166    2.596613
       _cons |   5.790433   4.098127     1.41   0.161    -2.342166    13.92303
------------------------------------------------------------------------------
```

- When we regress $y$ on $x_2$ alone, what happens?

```
      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  1,    98) = 2033.12
       Model |  3487.78809     1  3487.78809           Prob > F      =  0.0000
    Residual |  168.117208    98  1.71548171           R-squared     =  0.9540
-------------+------------------------------           Adj R-squared =  0.9535
       Total |   3655.9053    99  36.9283364           Root MSE      =  1.3098

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x2 |  -2.894991   .0642045   -45.09   0.000    -3.022403    -2.76758
       _cons |   10.83503   .1752564    61.82   0.000     10.48724    11.18282
------------------------------------------------------------------------------
```

- When we regress $y$ on $x_1$ and $x_2$ together, what happens?

# Omitted variables

- When we regress $y$ on $x_1$ alone, what happens?

```
      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  1,    98) =    0.00
       Model |  .096434032     1  .096434032           Prob > F      =  0.9596
    Residual |  3655.80887    98  37.3041721           R-squared     =  0.0000
-------------+------------------------------           Adj R-squared = -0.0102
       Total |   3655.9053    99  36.9283364           Root MSE      =  6.1077

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |  -.0682766   1.342874    -0.05   0.960    -2.733166    2.596613
       _cons |   5.790433   4.098127     1.41   0.161    -2.342166    13.92303
------------------------------------------------------------------------------
```

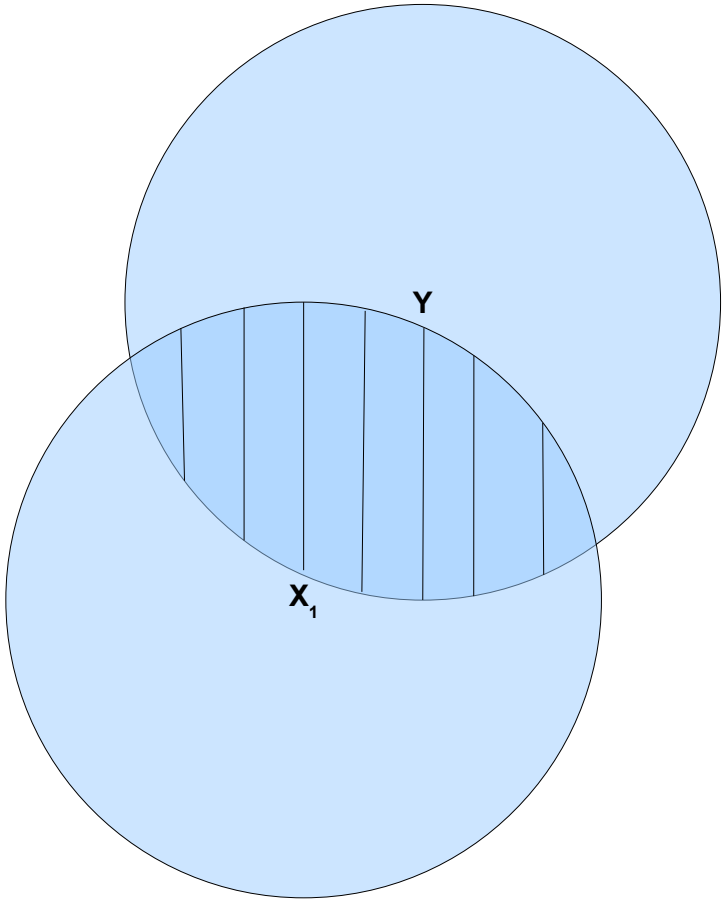
- When we regress $y$ on $x_2$ alone, what happens?

```
      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  1,    98) = 2033.12
       Model |  3487.78809     1  3487.78809           Prob > F      =  0.0000
    Residual |  168.117208    98  1.71548171           R-squared     =  0.9540
-------------+------------------------------           Adj R-squared =  0.9535
       Total |   3655.9053    99  36.9283364           Root MSE      =  1.3098

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x2 |  -2.894991   .0642045   -45.09   0.000    -3.022403    -2.76758
       _cons |   10.83503   .1752564    61.82   0.000     10.48724    11.18282
------------------------------------------------------------------------------
```

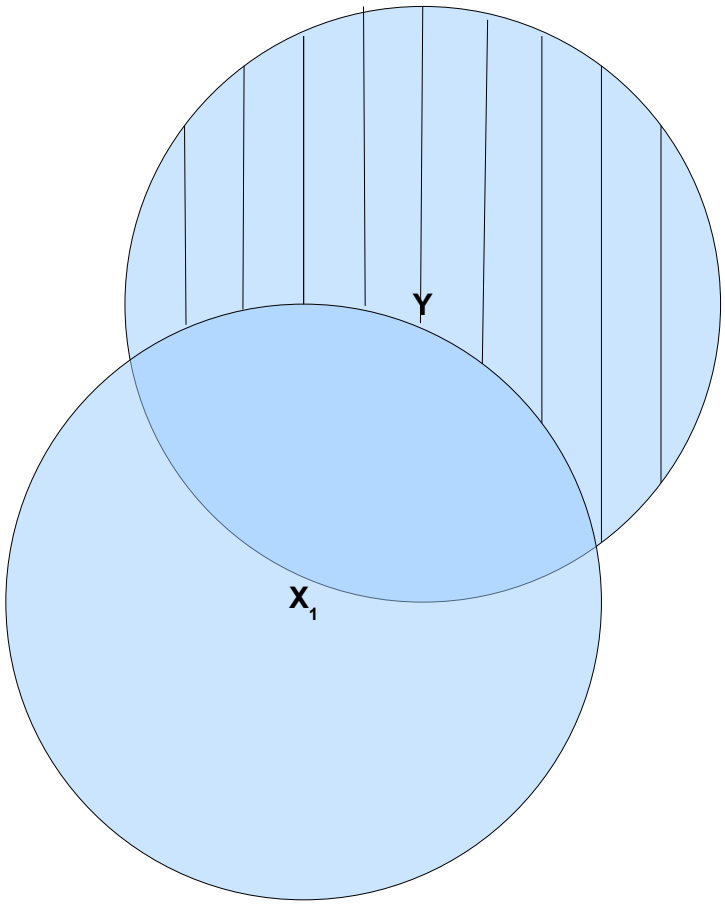
- When we regress $y$ on $x_1$ and $x_2$ together, what happens?

```
      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  2,    97) = 2106.00
       Model |  3573.6071     2  1786.80355           Prob > F      =  0.0000
    Residual |  82.2982026    97  .848435078           R-squared     =  0.9775
-------------+------------------------------           Adj R-squared =  0.9770
       Total |   3655.9053    99  36.9283364           Root MSE      =  .92111

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   2.063415   .2051654    10.06   0.000     1.656218    2.470611
          x2 |  -2.968643   .0457425   -64.90   0.000    -3.059429   -2.877857
       _cons |   4.741896   .6182503     7.67   0.000      3.51484    5.968952
------------------------------------------------------------------------------
```

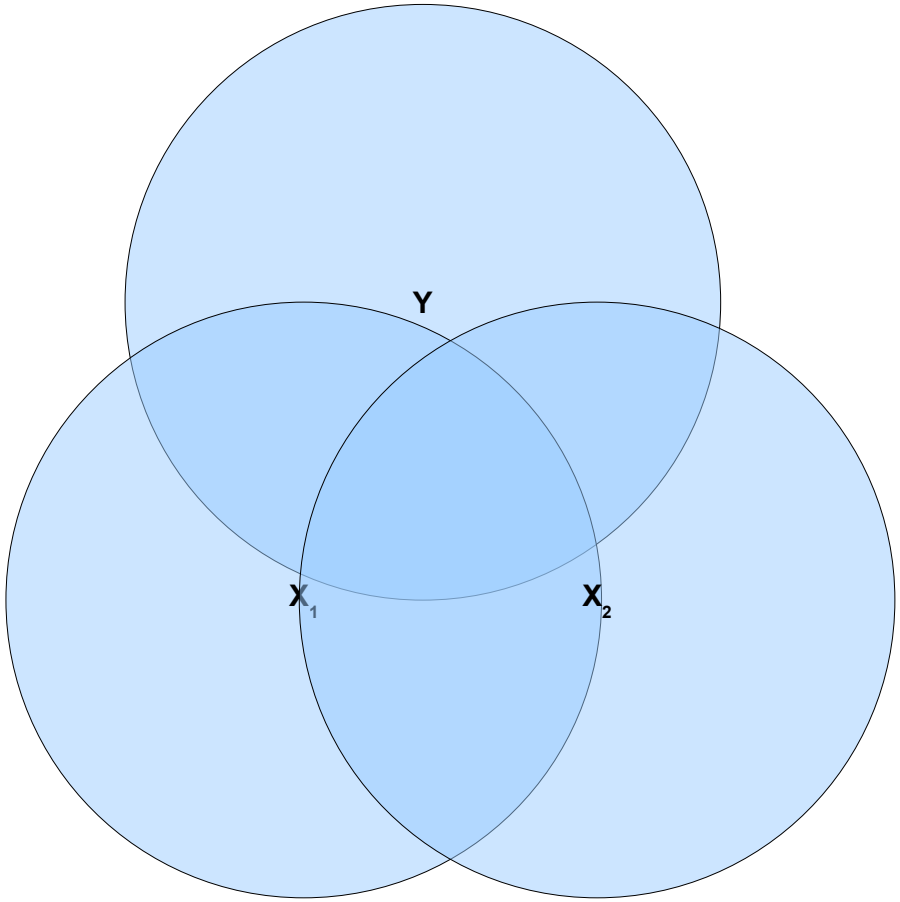
# What the OLS estimator does



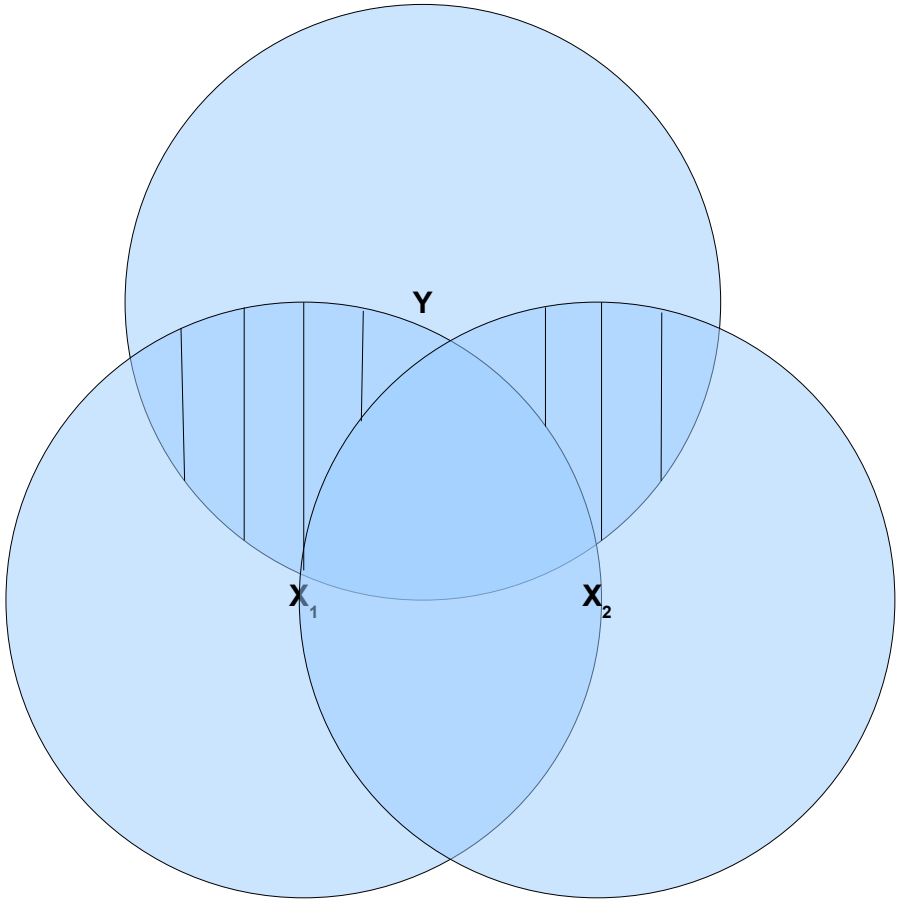
# What the OLS estimator does



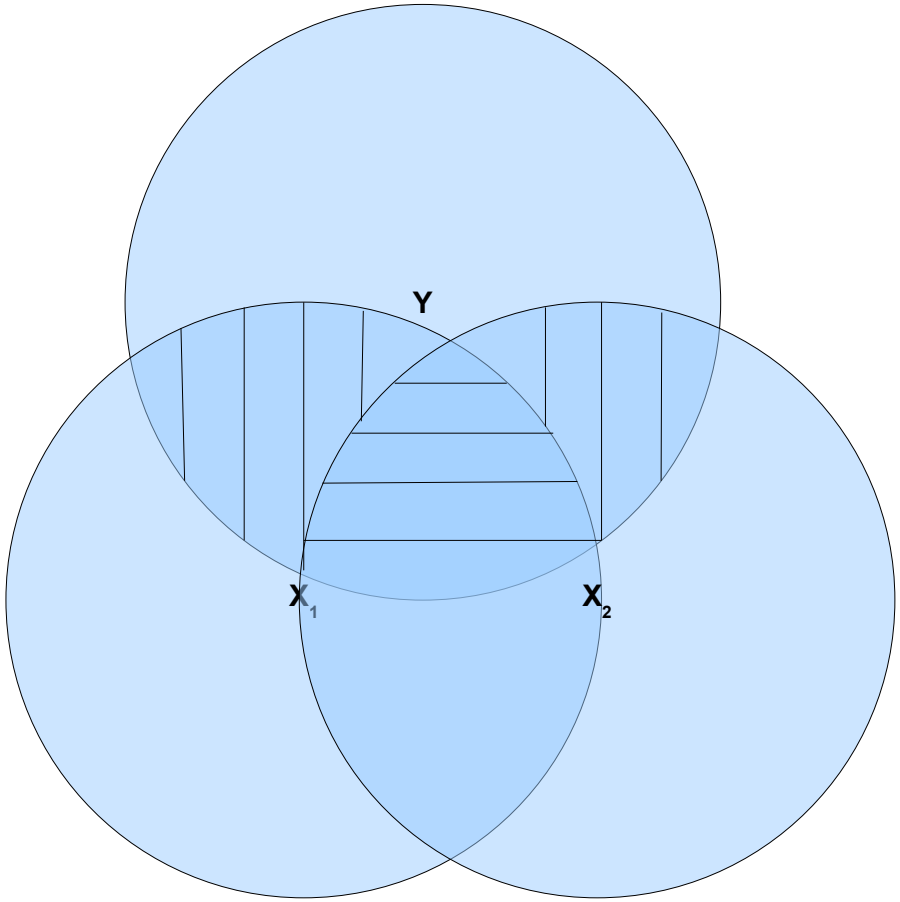
# What the OLS estimator does



# What the OLS estimator does



# What the OLS estimator does



# Try it

## Stata code

```
clear all
* Set seed
set seed 12345
* Create a matrix of correlations
matrix C = (1, 0, 0 \ 0, 1, 0 \ ///
0, 0, 1)
* Create a matrix of means
matrix m = (3,2,2)
* Create a matrix of standard deviations
matrix sd = (0.5,2,1)
* Draw three random variable from a
* multivariate distribution
drawnorm x1 x2 x3, n(100) means(m) ///
sds(sd) corr(C)
* Draw some "unobservable" stuff
gen eps = rnormal()
```

# OLS estimator
## Try it

### Stata code

```
* Create a dependent variable y
gen y = 5 + 2*x1 - 3*x2 + eps
* Regress y on x1 (by itself)
regress y x1
```

# OLS estimator
## Try it

### Stata code

```
* Create a dependent variable y
gen y = 5 + 2*x1 - 3*x2 + eps
* Regress y on x1 (by itself)
regress y x1
* Regress y on x2 (by itself)
regress y x2
```

# OLS estimator
## Try it

### Stata code

```
* Create a dependent variable y
gen y = 5 + 2*x1 - 3*x2 + eps
* Regress y on x1 (by itself)
regress y x1
* Regress y on x2 (by itself)
regress y x2
* Regress y on x1 and x2
regress y x1 x2
```

# Omitted variables

- When we regress $y$ on $x_1$ alone, what happens?

# Omitted variables

- When we regress $y$ on $x_1$ alone, what happens?

```
      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  1,    98) =    2.91
       Model |  112.960597     1  112.960597           Prob > F      =  0.0913
    Residual |  3806.29417    98  38.8397364           R-squared     =  0.0288
-------------+------------------------------           Adj R-squared =  0.0189
       Total |  3919.25476    99   39.588432           Root MSE      =  6.2322

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   2.336792   1.370234     1.71   0.091    -.3823925    5.055976
       _cons |  -1.412461   4.181623    -0.34   0.736    -9.710755    6.885833
------------------------------------------------------------------------------
```

# Omitted variables

- When we regress $y$ on $x_1$ alone, what happens?

```
      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  1,    98) =    2.91
       Model |  112.960597     1  112.960597            Prob > F      =  0.0913
    Residual |  3806.29417    98  38.8397364            R-squared     =  0.0288
-------------+------------------------------           Adj R-squared =  0.0189
       Total |  3919.25476    99   39.588432           Root MSE      =  6.2322

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   2.336792   1.370234     1.71   0.091    -.3823925    5.055976
       _cons |  -1.412461   4.181623    -0.34   0.736    -9.710755    6.885833
------------------------------------------------------------------------------
```

- When we regress $y$ on $x_2$ alone, what happens?

# Omitted variables

- When we regress $y$ on $x_1$ alone, what happens?

```
      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  1,    98) =    2.91
       Model |  112.960597     1  112.960597           Prob > F      =  0.0913
    Residual |  3806.29417    98  38.8397364           R-squared     =  0.0288
-------------+------------------------------           Adj R-squared =  0.0189
       Total |  3919.25476    99   39.588432           Root MSE      =  6.2322

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   2.336792   1.370234     1.71   0.091    -.3823925    5.055976
       _cons |  -1.412461   4.181623    -0.34   0.736    -9.710755    6.885833
------------------------------------------------------------------------------
```

- When we regress $y$ on $x_2$ alone, what happens?

```
      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  1,    98) = 2128.61
       Model |  3746.75642     1  3746.75642           Prob > F      =  0.0000
    Residual |  172.498349    98  1.76018724           R-squared     =  0.9560
-------------+------------------------------           Adj R-squared =  0.9555
       Total |  3919.25476    99   39.588432           Root MSE      =  1.3267

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x2 |  -2.977827   .0645433   -46.14   0.000    -3.105911   -2.849743
       _cons |   10.98567   .1761551    62.36   0.000     10.6361    11.33525
------------------------------------------------------------------------------
```

# Omitted variables

- When we regress $y$ on $x_1$ alone, what happens?

```
      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  1,    98) =    2.91
       Model |  112.960597     1  112.960597           Prob > F      =  0.0913
    Residual |  3806.29417    98  38.8397364           R-squared     =  0.0288
-------------+------------------------------           Adj R-squared =  0.0189
       Total |  3919.25476    99   39.588432           Root MSE      =  6.2322

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   2.336792   1.370234     1.71   0.091    -.3823925    5.055976
       _cons |  -1.412461   4.181623    -0.34   0.736    -9.710755    6.885833
------------------------------------------------------------------------------
```

- When we regress $y$ on $x_2$ alone, what happens?

```
      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  1,    98) = 2128.61
       Model |  3746.75642     1  3746.75642           Prob > F      =  0.0000
    Residual |  172.498349    98  1.76018724           R-squared     =  0.9560
-------------+------------------------------           Adj R-squared =  0.9555
       Total |  3919.25476    99   39.588432           Root MSE      =  1.3267

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x2 |  -2.977827   .0645433   -46.14   0.000    -3.105911   -2.849743
       _cons |   10.98567   .1761551    62.36   0.000     10.6361    11.33525
------------------------------------------------------------------------------
```

- When we regress $y$ on $x_1$ and $x_2$ together, what happens?

# Omitted variables

- When we regress $y$ on $x_1$ alone, what happens?

```
      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  1,    98) =    2.91
       Model |  112.960597     1  112.960597           Prob > F      =  0.0913
    Residual |  3806.29417    98  38.8397364           R-squared     =  0.0288
-------------+------------------------------           Adj R-squared =  0.0189
       Total |  3919.25476    99   39.588432           Root MSE      =  6.2322

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   2.336792   1.370234     1.71   0.091    -.3823925    5.055976
       _cons |  -1.412461   4.181623    -0.34   0.736    -9.710755    6.885833
------------------------------------------------------------------------------
```

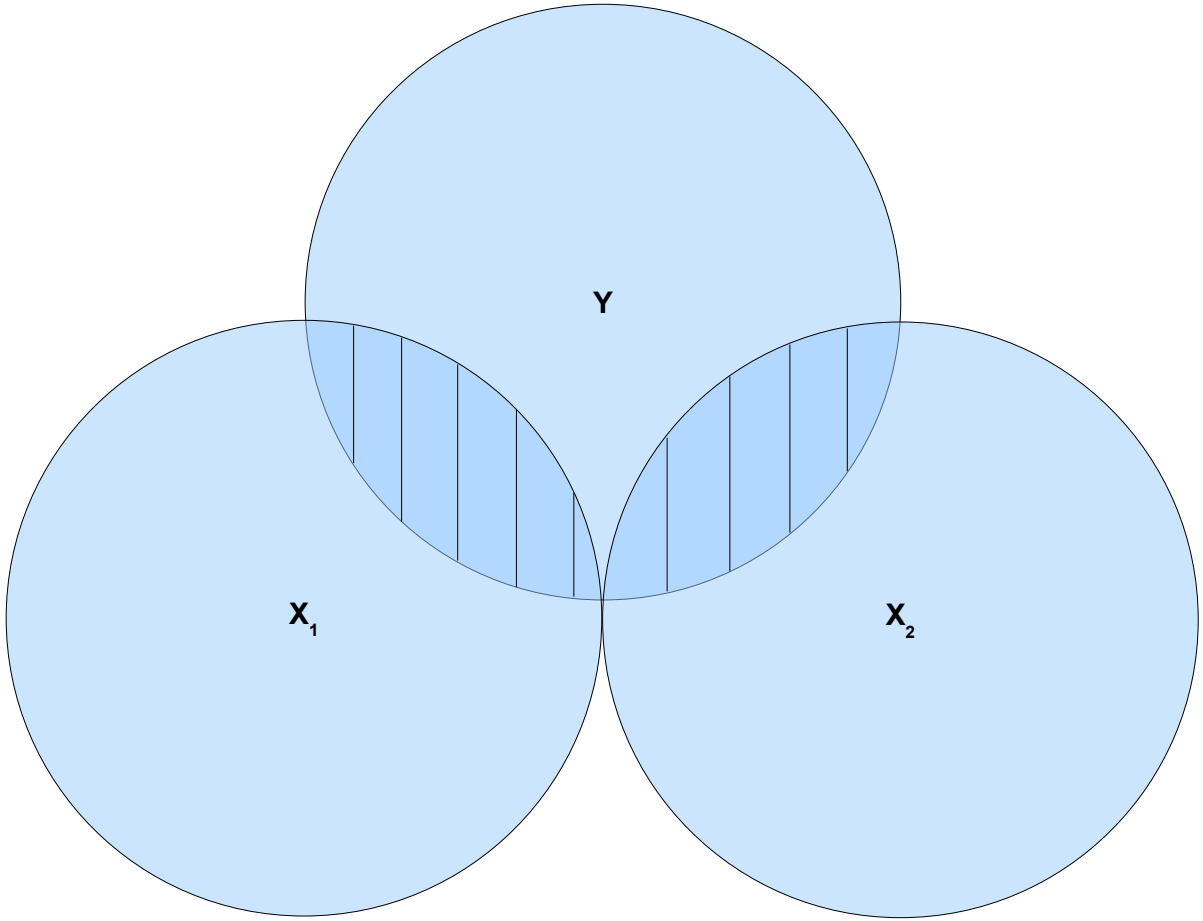
- When we regress $y$ on $x_2$ alone, what happens?

```
      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  1,    98) = 2128.61
       Model |  3746.75642     1  3746.75642           Prob > F      =  0.0000
    Residual |  172.498349    98  1.76018724           R-squared     =  0.9560
-------------+------------------------------           Adj R-squared =  0.9555
       Total |  3919.25476    99   39.588432           Root MSE      =  1.3267

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x2 |  -2.977827   .0645433   -46.14   0.000    -3.105911   -2.849743
       _cons |   10.98567   .1761551    62.36   0.000     10.6361    11.33525
------------------------------------------------------------------------------
```

- When we regress $y$ on $x_1$ and $x_2$ together, what happens?

```
      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  2,    97) = 2261.20
       Model |  3836.95656     2  1918.47828           Prob > F      =  0.0000
    Residual |   82.298202    97  .848435072           R-squared     =  0.9790
-------------+------------------------------           Adj R-squared =  0.9786
       Total |  3919.25476    99   39.588432           Root MSE      =  .92111

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |     2.0885   .2025537    10.31   0.000     1.686487    2.490513
          x2 |  -2.969277   .0448183   -66.25   0.000    -3.058229   -2.880325
       _cons |   4.667907   .6248163     7.47   0.000     3.427819    5.907994
------------------------------------------------------------------------------
```

# What the OLS estimator does



# Tricks with OLS

- See the difference?

# Tricks with OLS

- See the difference?
- Back to the correlated case
- I hope you're saving your work in a do-file so that you don't have to type everything again!

## Stata code

```
* Set seed
set seed 12345
* Create a matrix of correlations
matrix C = (1, 0.2, 0.2 \ 0.2, 1, 0.2 \ ///
0.2, 0.2, 1)
* Create a matrix of means
matrix m = (3,2,2)
* Create a matrix of standard deviations
matrix sd = (0.5,2,1)
* Draw three random variable from a
* multivariate distribution
drawnorm x1 x2 x3, n(100) means(m) ///
sds(sd) corr(C)
* Draw some "unobservable" stuff
gen eps = rnormal()
```

# Tricks with OLS

- Now we're going to learn how to *isolate* the effect of $x_1$ on $y$, even if we don't include $x_2$ in the regression

# Tricks with OLS

- Now we're going to learn how to *isolate* the effect of $x_1$ on $y$, even if we don't include $x_2$ in the regression
- We're sort of going to cheat, but that's OK
- This particular type of cheating is going to provide us with lots of intuition when we start using instrumental variables

# Tricks with OLS

- I want you to do something that may seem strange at first:

# Tricks with OLS

- I want you to do something that may seem strange at first:
- I want you to regress $x_1$ on $x_2$
- Do it now

# Tricks with OLS

- I want you to do something that may seem strange at first:
- I want you to regress $x_1$ on $x_2$
- Do it now
- You have just separated $x_1$ into two parts:
  1. The part of $x_1$ that can be explained by $x_2$
  2. The part of $x_1$ that cannot be explained by $x_2$

# Tricks with OLS

- I want you to do something that may seem strange at first:
- I want you to regress $x_1$ on $x_2$
- Do it now
- You have just separated $x_1$ into two parts:
  1. The part of $x_1$ that can be explained by $x_2$
  2. The part of $x_1$ that cannot be explained by $x_2$
- Which part is used to estimate the true $\beta_1$?
- Think back to the Ballantine if you're not sure

## Stata code

```
* Regress x1 on x2
reg x1 x2
* Obtain predictions

* Use those predictions to obtain the ///
unexplained variation in x1

* Regress y on x1 and x2, so we know ///
what we're shooting for

* Regress y on x1hat

* Ta-da!
```

## Stata code

```
* Regress x1 on x2
reg x1 x2
* Obtain predictions
predict x1hat, xb
* Use those predictions to obtain the ///
unexplained variation in x1

* Regress y on x1 and x2, so we know ///
what we're shooting for

* Regress y on x1hat

* Ta-da!
```

## Stata code

```
* Regress x1 on x2
reg x1 x2
* Obtain predictions
predict x1hat, xb
* Use those predictions to obtain the ///
unexplained variation in x1
gen x1u = x1 - x1hat
* Regress y on x1 and x2, so we know ///
what we're shooting for

* Regress y on x1hat

* Ta-da!
```

## Stata code

```
* Regress x1 on x2
reg x1 x2
* Obtain predictions
predict x1hat, xb
* Use those predictions to obtain the ///
unexplained variation in x1
gen x1u = x1 - x1hat
* Regress y on x1 and x2, so we know ///
what we're shooting for
reg y x1 x2
* Regress y on x1hat

* Ta-da!
```

## Stata code

```
* Regress x1 on x2
reg x1 x2
* Obtain predictions
predict x1hat, xb
* Use those predictions to obtain the ///
unexplained variation in x1
gen x1u = x1 - x1hat
* Regress y on x1 and x2, so we know ///
what we're shooting for
reg y x1 x2
* Regress y on x1hat
reg y x1u
* Ta-da!
```

# Tricks with OLS
## Making them useful

- We have just learned how to isolate variation in our independent variables
- Turns out this is pretty useful
- The technique really is at the heart of the technique we use most often to solve common endogeneity problems in econometrics
- Wait, what the heck is endogeneity?

# Endogeneity
## What is it?

- Technically, endogeneity is pretty simple
  - A regressor is *endogenous* when it is correlated with the unobservables
  - That is: $cov(x, \epsilon) \neq 0$

# Endogeneity
## What is it?

- Technically, endogeneity is pretty simple
  - A regressor is *endogenous* when it is correlated with the unobservables
  - That is: $cov(x, \epsilon) \neq 0$
- But that technical definition doesn't give you the intuition you need to be good applied econometricians
- To do that, let's consider the most common types of endogeneity

# Endogeneity
## Common flavors

1. Reverse causality (simultaneity)
2. Omitted variables
3. Measurement error
4. Sample selection

# Reverse causality

Reverse causality

$$y = x\beta + \epsilon_y \tag{1a}$$

$$x = y\alpha + \epsilon_x \tag{1b}$$

- If $x$ causes $y$ (implied by (1a)), and at the same time $y$ causes $x$ (implied by (1b)), then $x$ is correlated with $\epsilon_y$ and $y$ is correlated with $\epsilon_x$.
- Think of reverse causality like a feedback loop

# Reverse causality

- Suppose we want to estimate (1a) and we are specifically interested in finding out what $\beta$ is
- Ordinary least squares will work just fine if $x$ is uncorrelated with $\epsilon_y$
- But if we think that there is some reverse causality, we think that $x$ is determined, in part, by $y$
- Look again at (1b). Substitute the expression for $y$ in (1a) into (1b) to get

$$\begin{aligned} x &= y\alpha + \epsilon_x \\ &= (x\beta + \epsilon_y)\alpha + \epsilon_x. \end{aligned} \tag{2}$$

# Reverse causality

- Now ask yourself again: is $x$ uncorrelated with $\epsilon_y$ in (1a)?
- No!
- In fact, $\epsilon_y$ is in the equation determining $x$, as we can see in (2)
- So reverse causality can be thought of as a general problem of endogeneity, just as we claimed above.

# Omitted variables

- Omitted variables is a straightforward case of correlation between a regressor and the unobservables
- And we've already discussed it in some detail. So we'll make this brief.
- Suppose that $y$ is caused by $x1$, $x2$, and $x3$

$$y = \beta_0 + x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + \epsilon \tag{3}$$

# Omitted variables

- Suppose we are specifically interested in the coefficient $\beta_1$, i.e. the effect of $x1$ on $y$
- Suppose that $x_3$ is not in our dataset
- For all intents and purposes $x_3$ becomes part of the unobservable vector $\epsilon$
- If $x_3$ is uncorrelated with $x_1$, we've got no problem
- If, instead, $x_3$ is correlated with $x_1$, then when we regress

$$y = \beta_0 + x_1\beta_1 + x_2\beta_2 + \epsilon \tag{4}$$

we again have a situation where $cov(x_3, \epsilon) \neq 0$,i.e. a problem of endogeneity.

# Omitted variables

- WARNING: Including more variables in an equation does *not* necessarily solve OVB
- If you are missing a single variable from the "true" equation, and you find the data on this missing variable and include it in a regression, you are all set
- But how often does that happen?
- If you're missing a bunch of stuff, and you just include one more, you aren't necessarily making the problem any better
- Try it if you don't believe me — it's easy
  - Take your dset with $x_1$, $x_2$, and $x_3$
  - Create a y that is dependent on all three variables
  - Run a regression of $y$ on $x_1$ by itself
  - Then include $x_2$
  - Then include $x_3$

# Measurement error

Measurement error

$$y = x\beta + \epsilon \tag{5}$$

- Suppose $x$ is measured with error
- We have data on $x$, but it isn't exactly equal to the true $x$
- So let's say that the data we have is actually $\tilde{x}$, where $\tilde{x} = x + \eta$
- The error is unobservable — it is measurement error that is embedded in the data
- Our observations of $x$ are just . . . noisy.

# Measurement error

- We want to estimate $\beta$ in (5), but because our $x$ is measured with error, we are really working with the model

$$y = \tilde{x}\beta + \epsilon \tag{6}$$

# Measurement error

- We want to estimate $\beta$ in (5), but because our $x$ is measured with error, we are really working with the model

$$\begin{aligned} y &= \tilde{x}\beta + \epsilon \\ &= (x + \eta)\beta + \epsilon \\ &= x\beta + \eta\beta + \epsilon \\ &= x\beta + (\eta\beta + \epsilon) \end{aligned} \tag{6}$$

- Notice that everything grouped together in the final expression of (6), i.e. $(\eta\beta + \epsilon)$, is unobservable

# Measurement error

- We want to estimate $\beta$ in (5), but because our $x$ is measured with error, we are really working with the model

$$\begin{aligned} y &= \tilde{x}\beta + \epsilon \\ &= (x + \eta)\beta + \epsilon \\ &= x\beta + \eta\beta + \epsilon \\ &= x\beta + (\eta\beta + \epsilon) \end{aligned} \tag{6}$$

- Notice that everything grouped together in the final expression of (6), i.e. $(\eta\beta + \epsilon)$, is unobservable
- If we regress $y$ on the data that we *do* have, i.e if we regress $y$ on $\tilde{x}$, we have an endogeneity problem since $\tilde{x}$ is correlated with the unobservables: $cov(\tilde{x}, \eta\beta + \epsilon) \neq 0$

# Sample selection

- We'll deal with this using a different technique
- It's not so different from IV conceptually, but the mechanics are different
- Preview: Heckman selection model
- Hint: This is the hospital example
- But we'll leave this for later, since we need to get to IV now

# Endogeneity example

- An example of endogeneity
  - earnings & schooling
  - political success & campaign donations
  - health & smoking
  - crime & police

# Instrumental variables

"The method of instrumental variables is a signature technique in the econometrics toolkit." – Angrist and Krueger (2001)

- Instrumental variables are variables that are explicitly excluded from some equations and included in others, and therefore are correlated with some outcomes only through their effect on other variables.

  – Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin (1996). "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 91(434): 444-455.

# Instrument variables

$$y = x\beta + \epsilon$$
$$x = z\alpha + \eta$$

# Instrument variables

$$y = x\beta + \epsilon$$
$$x = z\alpha + \eta$$

$$\hat{\alpha} = (z'z)^{-1}z'x$$

$$\begin{aligned}\hat{x} &= z\hat{\alpha} \\ &= z(z'z)^{-1}z'x\end{aligned}$$

$$\begin{aligned}\hat{\beta}^{IV} &= (\hat{x}'\hat{x})^{-1}\hat{x}'y \\ &= (x'z(z'z)^{-1}z'z(z'z)^{-1}z'x)^{-1}x'z(z'z)^{-1}z'y \\ &= (x'z(z'z)^{-1}z'x)^{-1}x'z(z'z)^{-1}z'y\end{aligned}$$

## Stata code

```
clear all
* Set seed
set seed 12345
* Create a matrix of means
matrix m = (2,2)
* Create a matrix of std.  devs.
matrix sd = (2,1)
* Draw xy and x3 independently
drawnorm xy x3, n(100) means(m) sds(sd)
* Construct x1 out of x2, x3, and some
unobservable stuff
gen epsx = rnormal()
gen x1 = xy + x3 + epsx
```

# Instrumental variables
## Try it

### Stata code

```
* Construct y out of x1, x2, and some ///
unobservable stuff
gen epsy = rnormal()
gen y = xy + x1 + epsy
* Regress y on x1
regress y x1
* Now try it with a clean version of x1



* You can actually do this in one step if you
want
```

## Stata code

```
* Construct y out of x1, x2, and some ///
unobservable stuff
gen epsy = rnormal()
gen y = xy + x1 + epsy
* Regress y on x1
regress y x1
* Now try it with a clean version of x1
regress x1 x3
predict x1clean
regress y x1clean
* You can actually do this in one step if you
want
```

## Stata code

```
* Construct y out of x1, x2, and some ///
unobservable stuff
gen epsy = rnormal()
gen y = xy + x1 + epsy
* Regress y on x1
regress y x1
* Now try it with a clean version of x1
regress x1 x3
predict x1clean
regress y x1clean
* You can actually do this in one step if you
want
ivregress 2sls y (x1 = x3)
```

# IV: In practice

- IV should be used with large samples
  - IV is consistent but not unbiased
- Use robust standard errors
- Learn how to use IV in Stata by reading C & T

# Homework
## Book readings

- Read C & T
  - Read 6.1, 6.2, and up to and including 6.3.3 (i.e. don't read 6.3.4–6.3.8 yet)
  - Read 6.4.1 and 6.4.2

# Homework
## Journal readings

- Butcher, Kristin F., and Anne Case (1994). "The Effect of Sibling Sex Composition on Women's Education and Earnings," *The Quarterly Journal of Economics*, 109(3): 531-563.
- Angrist, Joshua D., and Alan B. Krueger (1991). "Does Compulsory School Attendance Affect Schooling and Earnings?," *The Quarterly Journal of Economics*, 106(4): 979-1014.

# Homework

- Due in TWO WEEKS (4 October), but get a jump
- Posted on the website