Econometrics

Lecture 5

Nathaniel Higgins

ERS and JHU

3 October 2011

Administrative

- What you are “responsible” for
  - Lecture material
  - Book material
  - Homework material
  - NOT Stata
  - We will review during lecture 7 (on 17 October)
- TA session
  - Multivariate regression
  - How to produce Stata output in tables (very useful)
  - Homework questions
- Burning questions on homework, now that you have seen my answers?

Multivariate regression

- So far have been doing this:

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

- we are bored with it now
- Almost never do bivariate regression in practical work

Multivariate regression

- So far have been doing this:

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

- we are bored with it now
- Almost never do bivariate regression in practical work
- Now we are doing this:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

- or this:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$$

Multivariate regression

- The difference between executing a bivariate and a multivariate regression in Stata?
- Nill

Multivariate regression

- The difference between executing a bivariate and a multivariate regression in Stata?
- Nill
- The difference between running the regression

$$birthWeight = \beta_0 + \beta_1 cigsPerDay + u$$

and

$$birthWeight = \beta_0 + \beta_1 cigsPerDay + \beta_2 methConsumption + u?$$

Multivariate regression

- The difference between executing a bivariate and a multivariate regression in Stata?
- Nill
- The difference between running the regression

$$birthWeight = \beta_0 + \beta_1 cigsPerDay + u$$

and

$$birthWeight = \beta_0 + \beta_1 cigsPerDay + \beta_2 methConsumption + u?$$

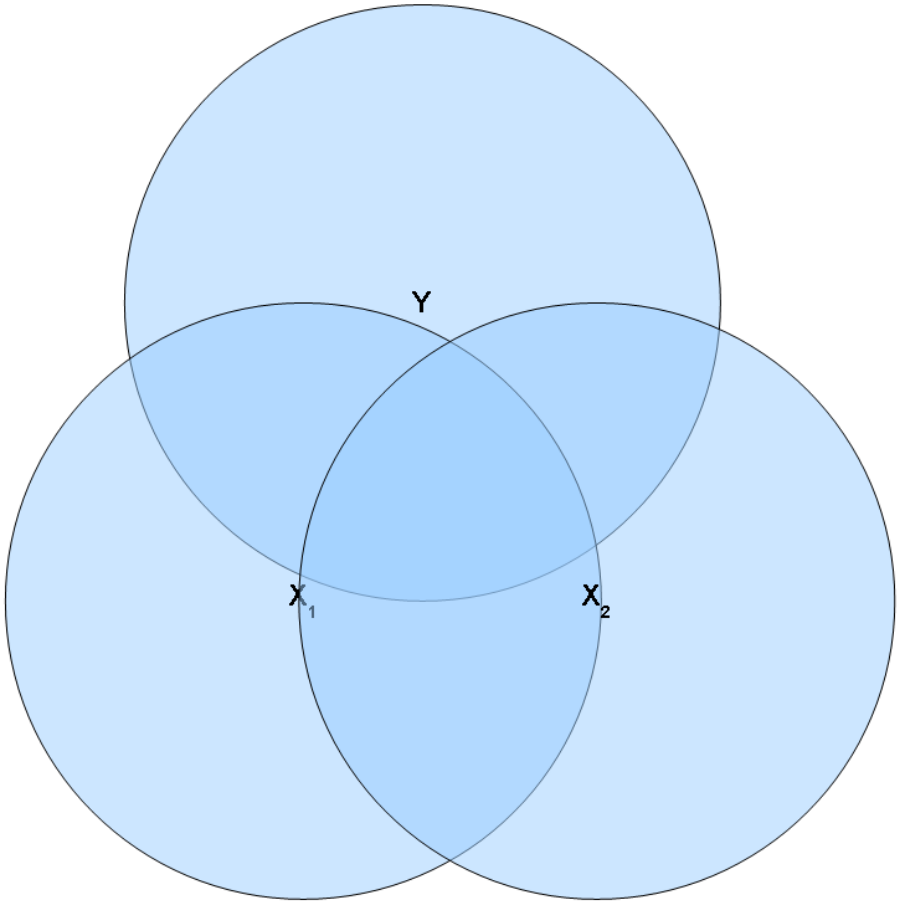
Stata code

```
reg birthWeight cigsPerDay
reg birthWeight cigsPerDay methConsumption
```

Multivariate regression

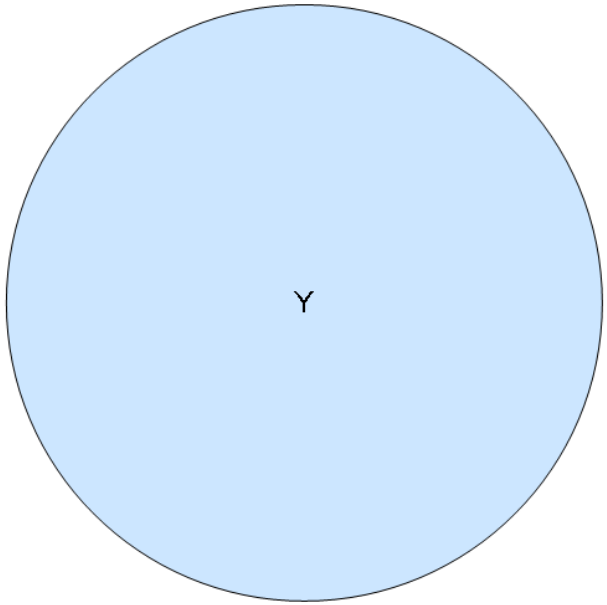
- So *doing* multivariate regression is easy (this is true)
- But it brings complications
- So this lecture is going to be about what those complications are
- First thing is understanding (conceptually) how each x-variable *individually* relates to the single y-variable
- Ballantine time!

Multivariate OLS, graphically
Why is it called the Ballantine?

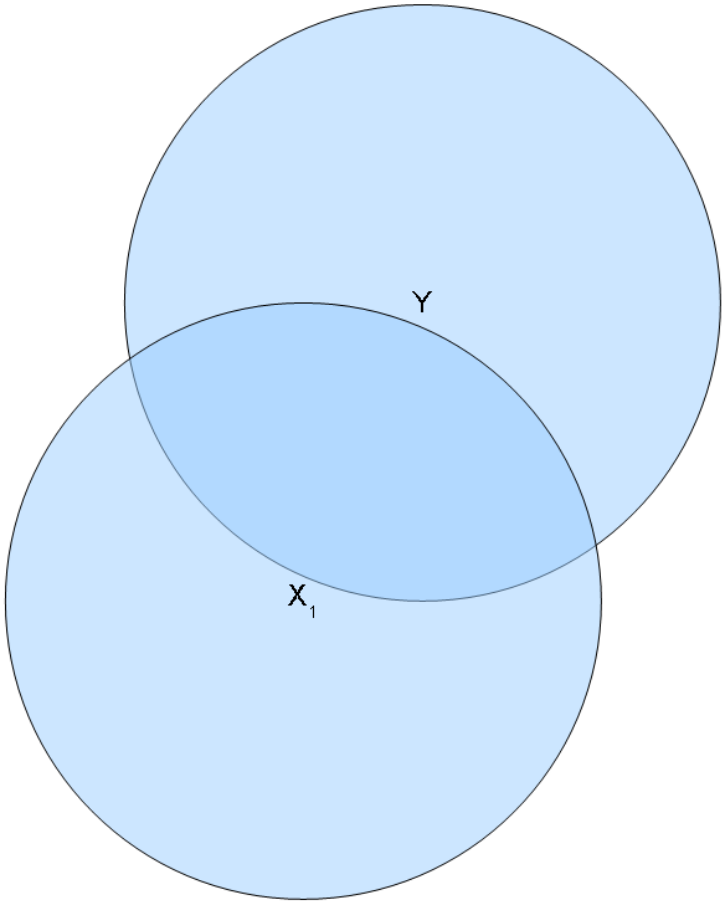


Multivariate OLS, graphically
Why is it called the Ballantine?

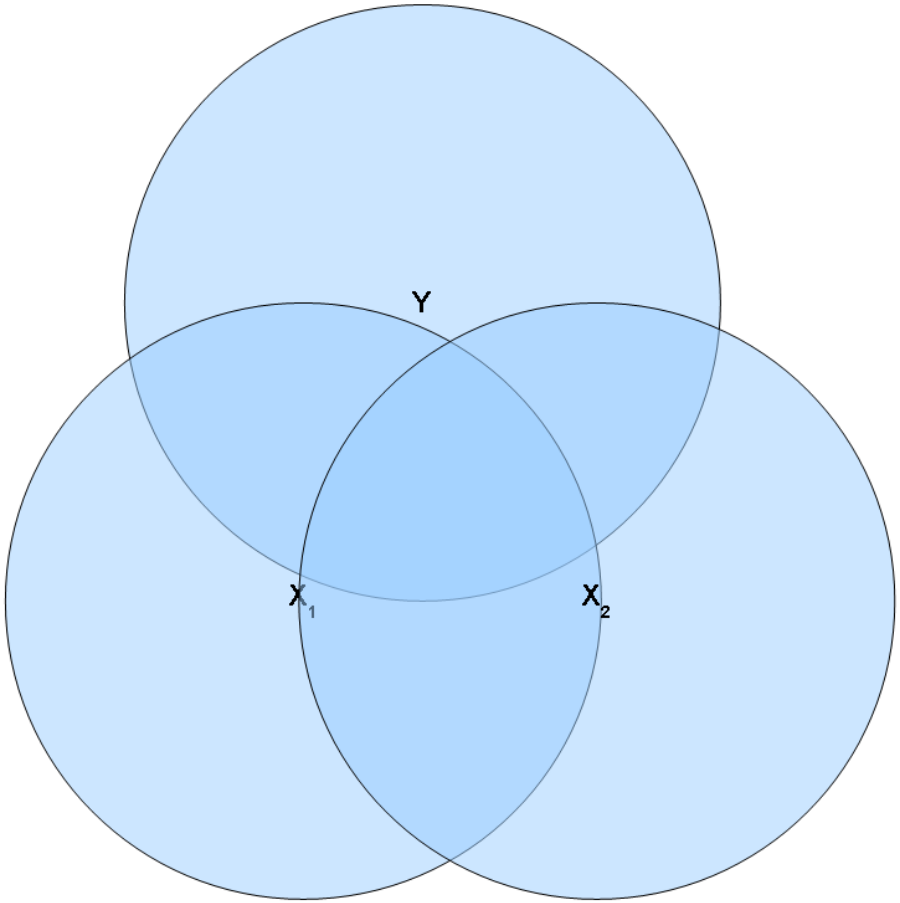
Multivariate OLS estimator
Graphically



Multivariate OLS estimator
Graphically



Multivariate OLS estimator
Graphically



Let's play with data

- We're going to create some fake data
- then we're going to run some regressions on this data
- Why make fake data (why not use real data?)
- Because we can create fake data with properties that we are sure about
- We run regressions under a bunch of different conditions (where we create the conditions) and see what happens
- Then, when we are using real data (i.e. when it counts), we have some intuition about what the results are telling us about how the real variables are related

Let's play with data
What data do we want to create?

- Let's create two x-variables and a y-variable
- The two x-variables are just going to be random data
  - That is, we're literally just going to draw random numbers out of a hat
  - BUT! The x-variables are going to be a little bit correlated
  - This is like saying that they overlap a little bit in the Ballantine drawing
- The y-variable is going to be determined by the x-variables and some other random (unobservable) data

Let's play with data
What data do we want to create??

- `x1` and `x2` are a bit correlated
- Represent this with a correlation matrix

$$\mathbf{C} = \begin{pmatrix} 1 & 0.2 \\ 0.2 & 1 \end{pmatrix}$$

- If `x1` and `x2` were not correlated at all, `C` would look like this instead

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

- If `x1` and `x2` were really correlated, `C` would look like this

$$\begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix}$$

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 mu = (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(mu) ///
sds(sd) corr(C)
* Draw some "unobservable" stuff
gen u = rnormal()
```

Stata code

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

Stata code

```
* Create a dependent variable y
gen y = 5 + 2*x1 - 3*x2 + u
* 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 + u
* 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
```

Save your data now: `save lecture-05-dset-01.dta`

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
------------------------------------------------------------------------------
```

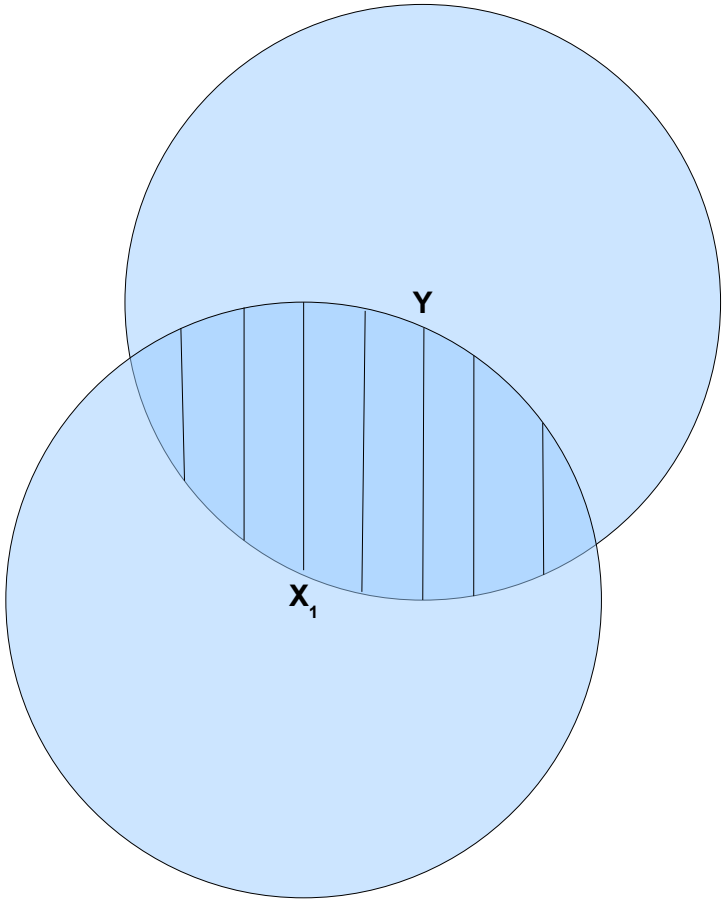
- When we regress y on x_2 alone, what happens?

```
      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
------------------------------------------------------------------------------
```

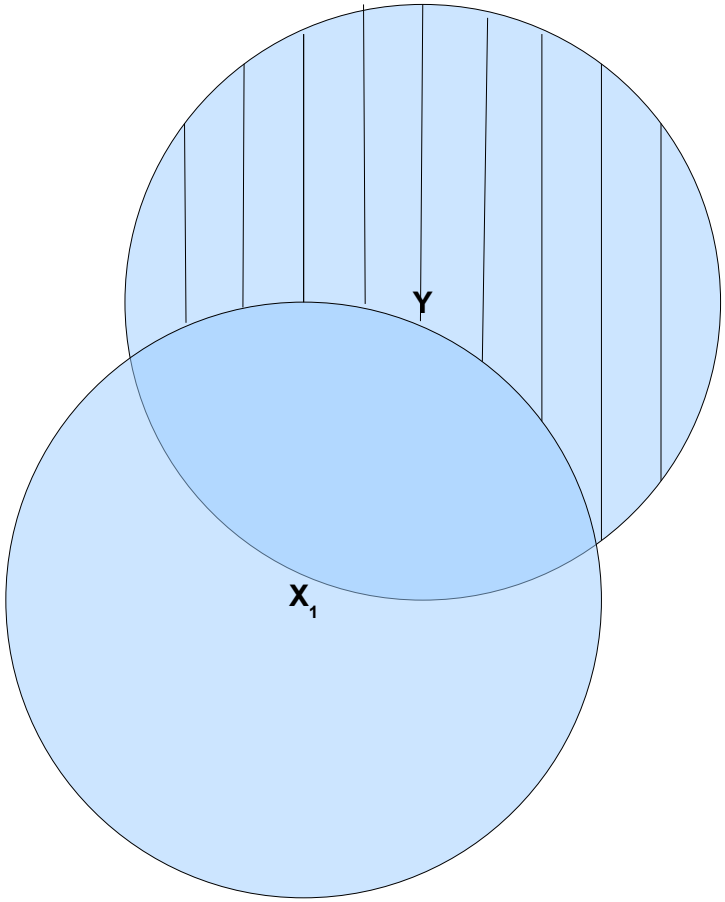
- When we regress y on x_1 and x_2 together, what happens?

```
      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
------------------------------------------------------------------------------
```

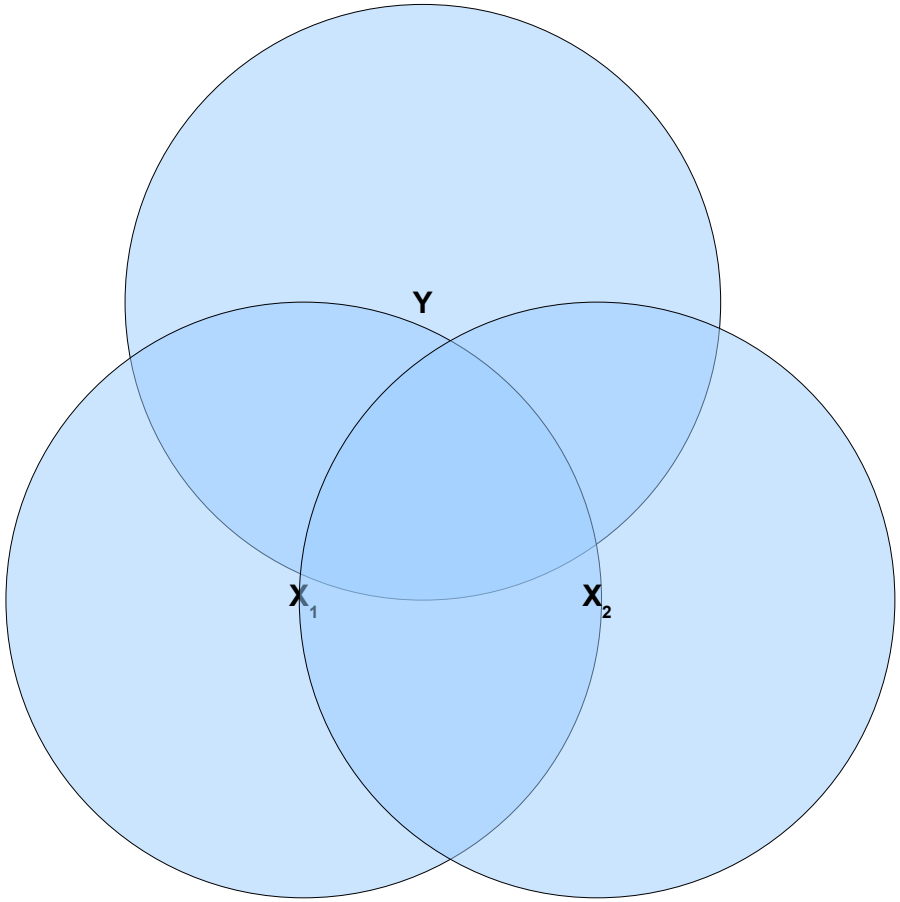
What the OLS estimator does



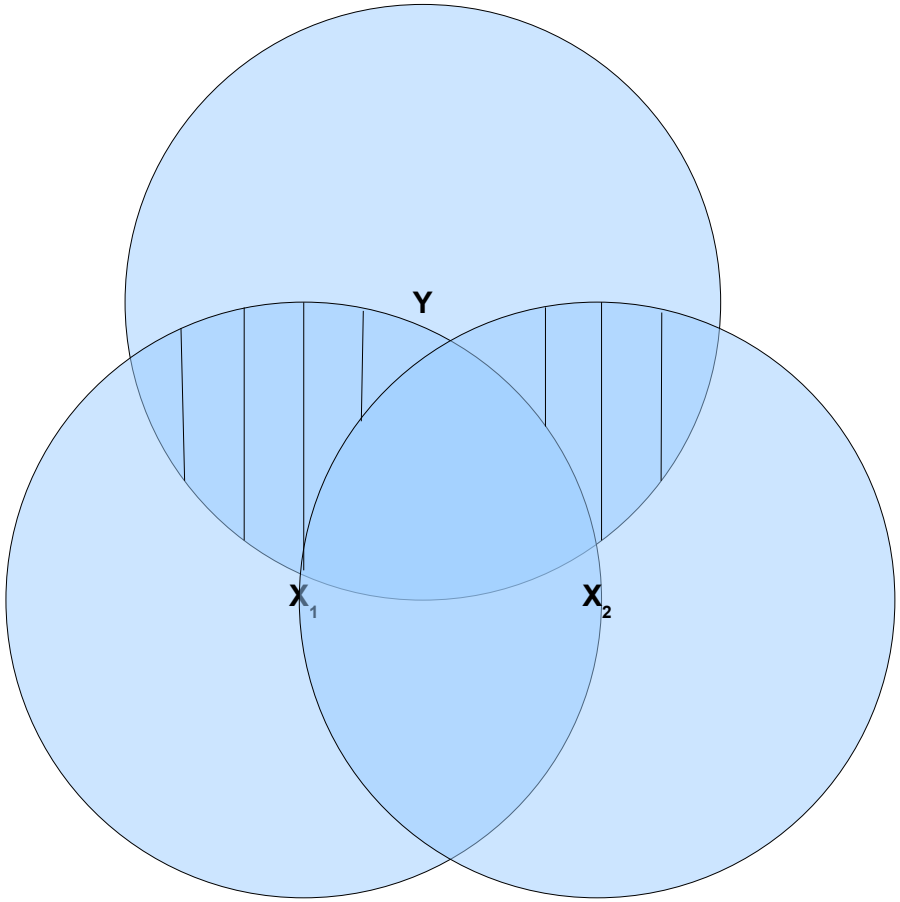
What the OLS estimator does



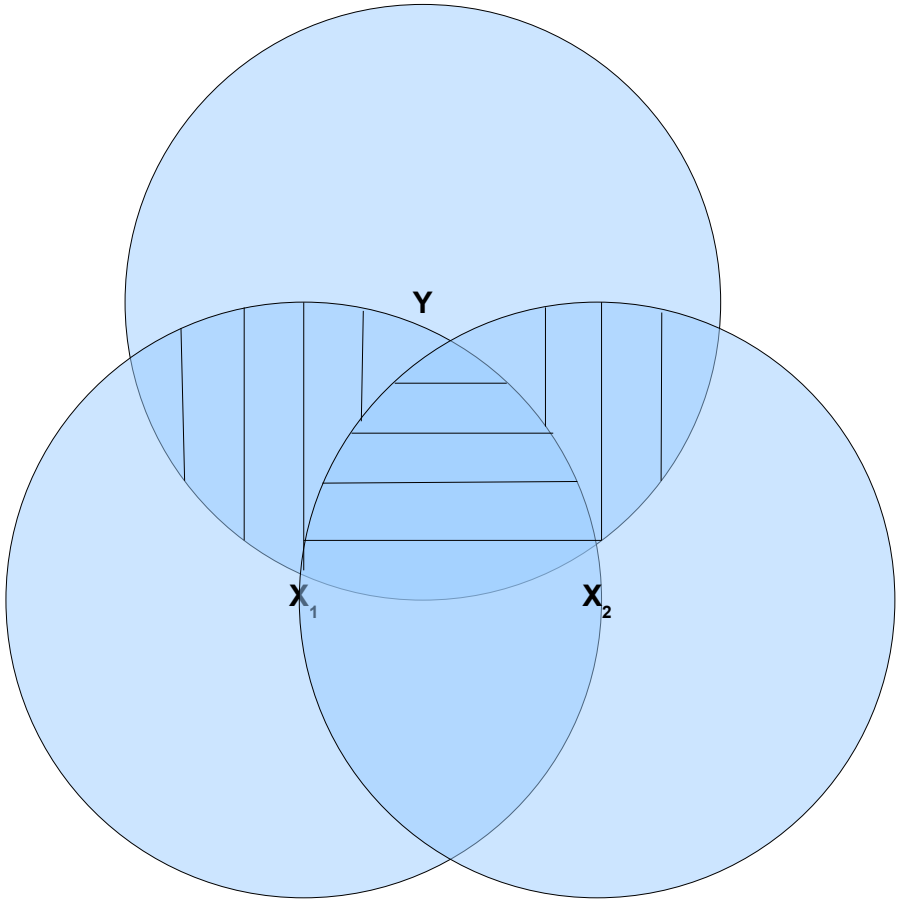
What the OLS estimator does



What the OLS estimator does



What the OLS estimator does



What the OLS estimator does

- We see that the OLS estimator changes when we regress `y` on `x1` by itself / `y` on `x2` by itself;
- The estimator works best when `y` is simultaneously regressed on both `x1` and `x2`. Why?
- The performance of the OLS estimator is worse when we regress `y` on just `x1` (as opposed to `y` on just `x2`). Why?

What the OLS estimator does

- We see that the OLS estimator changes when we regress `y` on `x1` by itself / `y` on `x2` by itself;
- The estimator works best when `y` is simultaneously regressed on both `x1` and `x2`. Why?
- The performance of the OLS estimator is worse when we regress `y` on just `x1` (as opposed to `y` on just `x2`). Why?
- Now let's see what happens when we do a similar exercise, but this time change the correlation matrix `C` so that `x1` and `x2` are independent (unrelated)

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 mu = (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(mu) ///
sds(sd) corr(C)
* Draw some "unobservable" stuff
gen u = rnormal()
```

Stata code

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

Stata code

```
* Create a dependent variable y
gen y = 5 + 2*x1 - 3*x2 + u
* 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 + u
* 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
```

Save your data now: `save lecture-05-dset-02.dta`

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
------------------------------------------------------------------------------
```

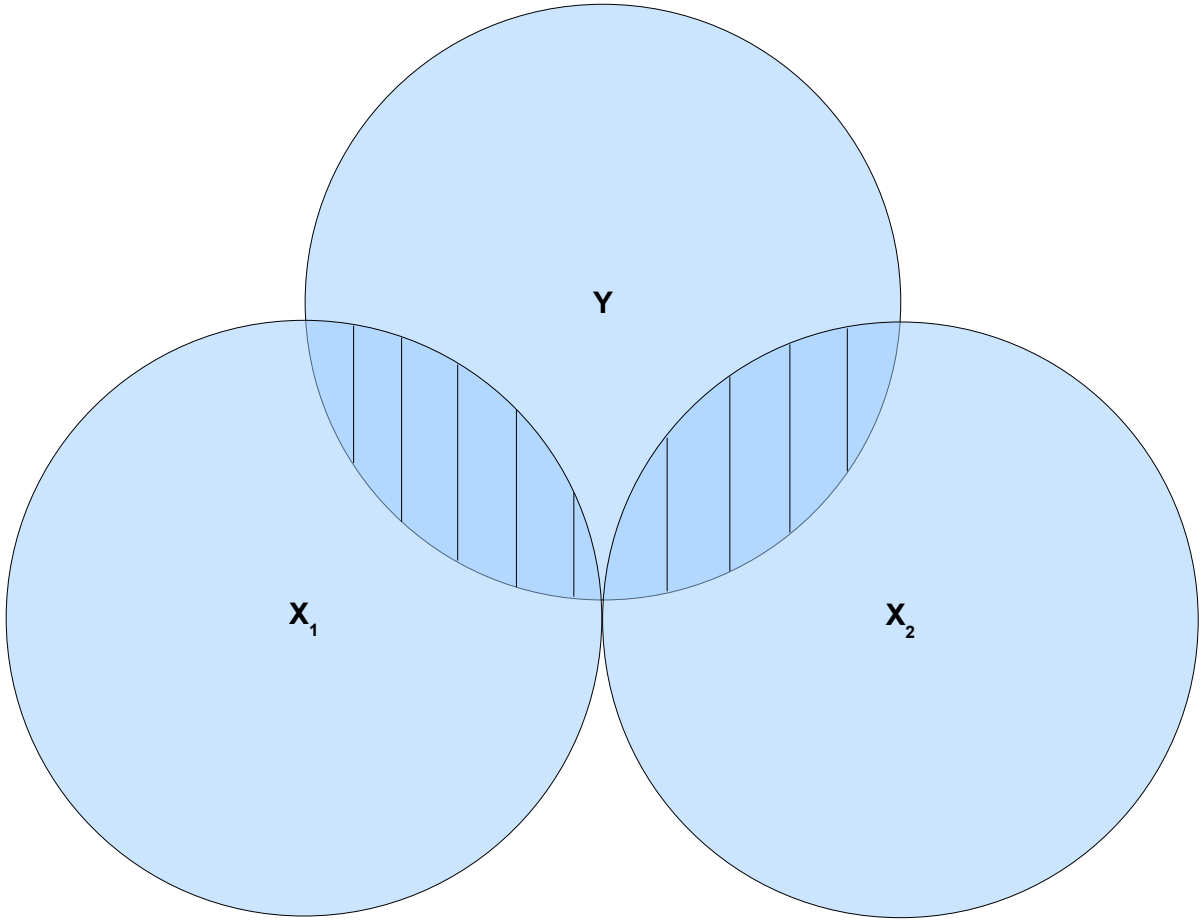
- When we regress y on x_2 alone, what happens?

```
      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?
- Lessons:
  1. It's variation that drives the bus: The more variation there is in an x-variable, the easier it is to estimate the coefficient associated with that variable (the easier it is to *detect* the x-variable's effect on y)
  2. When x-variables are related (when they co-move) it is harder to estimate their independent effects on the y-variable (this is multicollinearity)
  3. Omitting variables that *should* be in the regression is potentially harmful. That harm is minimized when the omitted variable is independent of the other variable(s).

Tricks with OLS

- See the difference?
- Lessons:
  1. It's variation that drives the bus: The more variation there is in an x-variable, the easier it is to estimate the coefficient associated with that variable (the easier it is to *detect* the x-variable's effect on y)
  2. When x-variables are related (when they co-move) it is harder to estimate their independent effects on the y-variable (this is multicollinearity)
  3. Omitting variables that *should* be in the regression is potentially harmful. That harm is minimized when the omitted variable is independent of the other variable(s).
- 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 β_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!
```

- We have just learned how to isolate variation in our independent variables
- Turns out this is pretty useful
- Lesson
  1. The *only* variation used to estimate the ceteris paribus relationship between an x-variable and y is the variation in x that is independent (not explained) by the other x-variables
- If we want to run the regression

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$$

- . . . then the estimate for β_1 (i.e. $\widehat{\beta_1}$) is obtained by "holding constant" the effects of x_2 and x_3 on y
- The OLS estimator does this (does the "holding constant") via a formula that mimics the following procedure

1. Regress x_1 on x_2 and x_3. That is, run the regression

$$x_1 = \alpha_0 + \alpha_1 x_2 + \alpha_2 x_3 + w$$

2. Obtain the predicted x_1 from the regression above; call it $\widehat{x_1}$. (note: don't be confused by the α's — they are just like β's. I am using α's just so we don't confuse the coefficients in this regression with the coefficients in the main regression that has y as the left-hand-side variable)
3. Use the predictions $\widehat{x_1}$ to obtain the unexplained variation in x_1 (the variation in x_1 not explained by variation in x_2 or x_3): $x_1 u = x_1 - \widehat{x_1}$ ($x_1 u$ means "x_1-unexplained")
4. Finally, regress y on $x_1 u$. This gives us the effect of x_1 on y, *independent* of the effect of x_2 and x_3 on y.

Endogeneity

- To reiterate: we have just learned how to isolate variation in our independent variables — this is what we are doing when we run a multiple regression
- 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?
- We encountered one kind of endogeneity when we ran a regression of y on only one of x_1 or x_2, when really we should have been running a regression of y on x_1 *and* x_2
- We call this type of endogeneity "omitted variable bias"

Endogeneity
What is it?

- In general, endogeneity is whenever the one or more of the included regressors (the x-variables) is related to the error term — remember, ideally we want the error term to be conditionally independent of the x-variables: $E(u|x) = E(u) = 0$
  - A regressor is *endogenous* when it is related to the unobservables
  - That is: $E(u|x) \neq 0$

- In general, endogeneity is whenever the one or more of the included regressors (the x-variables) is related to the error term — remember, ideally we want the error term to be conditionally independent of the x-variables: $E(u|x) = E(u) = 0$
  - A regressor is *endogenous* when it is related to the unobservables
  - That is: $E(u|x) \neq 0$
- You may recognize that Wooldridge writes this as

$$E(u|x_1, x_2, \ldots, x_k) \neq 0$$

- We mean the same thing

Endogeneity
Common flavors

1. Omitted variables (the only flavor we have considered so far)
2. Reverse causality (simultaneity)
3. Measurement error
4. Sample selection
5. These other flavors are just previews — you don't need to know anything about them yet

Omitted variables

- Omitted variables is a straightforward case of relation between a regressor and the unobservables
- Suppose that y is caused by $x1$, $x2$, and $x3$

$$y = \beta_0 + x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + u \tag{1}$$

Omitted variables

- Suppose we are specifically interested in the coefficient β_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 u
- If x_3 is uncorrelated with x_1 and x_2, we've got no problem
- If, instead, x_3 is correlated with x_1 *or* x_2, then when we regress

$$y = \beta_0 + x_1\beta_1 + x_2\beta_2 + \epsilon$$

we again have a situation where $E(u|x1, x2) \neq 0$

- Intuition: sometimes u is moving around (u is *varying* and this is causing y to move around). When we regress y on x_1 and x_2 (omitting x_3), we are mis-assigning to x_1 effects that are really due to movements in u.

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? Usually we are missing lots of stuff from the "true" equation
- If you're missing several variables, 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_3. Then include x_2. See?

Omitted variables
What is the lesson?

- Omitting variables from a regression can be costly
- Throwing in the "kitchen sink" might work . . . if you really have the kitchen sink
- If you throw in one more variable, the direction of the "correction" is unknown
- THEORY is important! You usually don't have all the data — so you have to think through the likely effect of omitted variables on your model

Omitted variables
Variables that *should* be omitted

- Load `lecture-05-dset-01.dta`
- Regress `y` on `x1,x2,x3`
- Does `x3` belong in the regression? Does it matter?

Summing-up

- You should be developing stronger intuition for what it means to "control" for all other factors when estimating a coefficient like β_1
  - Remove potential confounds
  - Look only at what we are sure is caused by x_1
  - Look at effect of x_1 on y, holding $x_2, \ldots, x_k$ fixed

Summing-up

- You should be developing stronger intuition for what it means to "control" for all other factors when estimating a coefficient like β_1
  - Remove potential confounds
  - Look only at what we are sure is caused by x_1
  - Look at effect of x_1 on y, holding $x_2, \ldots, x_k$ fixed
- Why add more variables to a regression?
  - To make the estimates of the betas correct (Ballantine?)
  - To predict y better (build better models of y) (Ballantine?)
  - To get a functional relationship correct

Exercise

- Read the article posted on my website under lecture 5 (handout)
- Q: if you ran a regression predicting poverty level (*y* variable) using the x-variables `poverty` and `black`, what would happen?
- Use the Ballantine to think through the problem

Next time

- How the inclusion of multiple variables influences the variance of the OLS estimates
- What can we infer from the data when we specify a multivariate model?
- Reading: Focus on 3.4 and 3.5, 4.1, 4.2, 4.5 (note that reading 4.3 and 4.4 might make 4.5 easier — I only direct your focus so that you will know what I think is most important to retain)

Homework

- C3.1 – C3.8 (starting on page 110)