LECTURE 2

Multiple Regression

Outline

1. Multiple regression:
   a. Omitted variable bias and control variables
   b. Mechanics
   c. Perfect multicolinearity
   d. Tests of q restrictions & F-statistics (finish)
   e. Critical values: t_{n-k-1} v. z; $F_{q,n-k-1}$ v. χ^2_q / q

Omitted Variable Bias

- Let β_1 be the causal effect of a change in X_1
- If Y depends on another variable, not in the regression, and that variable co-varies with X_1, then that omitted variable is a confounding factor and the OLS estimator of β_1 is biased.
- The bias in the OLS estimator that occurs because of an omitted factor is called ***omitted variable*** bias.

For OVB to occur, the omitted factor "Z" must satisfy both:

1. Z is a determinant of Y (i.e. Z is part of u); and
2. Z is correlated with X (*i.e.* corr(Z,X) $\neq 0$)

- The best solution to OVB is including Z if it is available.
- Or, it might be possible to include a "control" variable that controls for the effect of Z, if Z is not available (much more on this later)

OVB example: *Beauty* and *onecredit*

```
. reg courseevaluation btystdave, r
Linear regression                               Number of obs =     463
                                                F(  1,   461) =   16.94
                                                Prob > F      =  0.0000
                                                R-squared     =  0.0357
                                                Root MSE      =  .54545
------------------------------------------------------------------------------
             |               Robust
courseeval~n |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   btystdave |   .1330014   .0323189     4.12   0.000     .0694908    .1965121
       _cons |   4.010023   .0253299   158.31   0.000     3.960246    4.059799
------------------------------------------------------------------------------

. * (i) Is onecredit a determinant of courseevaluation?
. reg courseevaluation btystdave onecredit, r
Linear regression                               Number of obs =     463
                                                F(  2,   460) =   28.47
                                                Prob > F      =  0.0000
                                                R-squared     =  0.0993
                                                Root MSE      =  .52773
------------------------------------------------------------------------------
             |               Robust
courseeval~n |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   btystdave |   .1480829   .0318059     4.66   0.000       .08558    .2105857
   onecredit |   .5985639   .0909013     6.58   0.000     .4199307    .7771971
       _cons |    3.97645   .0255452   155.66   0.000      3.92625    4.026649
------------------------------------------------------------------------------
```

```
. * (ii) are beauty and onecredit correlated?

. corr btystdave onecredit
(obs=463)

             | btyst~ve onecre~t
-------------+------------------
   btystdave |   1.0000
   onecredit |  -0.0847   1.0000

. reg btystdave onecredit, r
Linear regression                               Number of obs =     463
                                                F(  1,   461) =    6.08
                                                Prob > F      =  0.0140
                                                R-squared     =  0.0072
                                                Root MSE      =  .78667
------------------------------------------------------------------------------
             |               Robust
   btystdave |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   onecredit |  -.2847549   .1154962    -2.47   0.014    -.5117193   -.0577906
       _cons |  -.0717435   .0382323    -1.88   0.061    -.1468747    .0033877
------------------------------------------------------------------------------
```

Omitted variable bias formula

Suppose there is a single omitted variable Z:

$$E\hat{\beta}_1 = \beta_1 + \left(\frac{\sigma_u}{\sigma_X}\right)\rho_{Xu}$$

where ρ_{Xu} = corr(X,u).

- If an omitted factor Z is ***both***:

  (1) a determinant of Y (that is, it is contained in u); ***and***

  (2) correlated with X,

  then $\rho_{Xu} \neq 0$ and the OLS estimator $\hat{\beta}_1$ is biased.

- If the data are from an ideal randomized controlled experiment, then $E(u|X) = 0$, $\rho_{Xu} = 0$, and there is no omitted variable bias.
- If $\rho_{Xu} \neq 0$, then $E(u|X) \neq 0$, so OLS is biased, that is, $E(\hat{\beta}_1) \neq \beta_1$.

Digression: derivation of the OV bias formula

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$ (formula for OLS estimator).

Now $Y_i - \bar{Y} = \beta_1(X_i - \bar{X}) + (u_i - \bar{u})$, so

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})[\beta_1(X_i - \bar{X}) + (u_i - \bar{u})]}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

or

$$\hat{\beta}_1 - \beta_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(u_i - \bar{u})}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

When n is large,

$$\hat{\beta}_1 - \beta_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(u_i - \bar{u})}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$\xrightarrow{p} \frac{\text{cov}(X_i, u_i)}{\text{var}(X_i)} = \frac{\sigma_{Xu}}{\sigma_X^2} \quad (\text{“}\xrightarrow{p}\text{” means large-}n\text{ limit})$$

$$= \left(\frac{\sigma_u}{\sigma_X}\right) \times \left(\frac{\sigma_{Xu}}{\sigma_X \sigma_u}\right) = \left(\frac{\sigma_u}{\sigma_X}\right) \rho_{Xu},$$

where $\rho_{Xu} = \text{corr}(X,u)$. Rearranging the final expression yields,

$$\hat{\beta}_1 = \beta_1 + \left(\frac{\sigma_u}{\sigma_X}\right) \rho_{Xu}.$$

Technical note: this is a limit for large n.

Course evaluations and beauty: multiple regression

female = 1 if instructor is female

age = age of instructor

minority = 1 if instructor is minority

nonenglish = 1 if instructor is not native English speaker

tenuretrack = 1 if instructor is Asst/Assoc/Prof

lower = 1 if lower-division course

onecredit = 1 if single-credit course (PE, Glee Club, studio art,…)

Adjusted R^2, RMSE: Course evaluations & multiple regressors

```
. reg courseevaluation btystdave female age minority
      nonenglish tenuretrack lower onecredit, r

Linear regression                               Number of obs =     463
                                                F(  8,   454) =   12.85
                                                Prob > F      =  0.0000
                                                R-squared     =  0.1577
                                                Root MSE      =  .51372

------------------------------------------------------------------------------
             |               Robust
courseeval~n |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   btystdave |    .1564742   .0301056     5.20   0.000     .0973106    .2156379
      female |   -.1910459   .0529734    -3.61   0.000    -.2951493   -.0869425
         age |   -.0024049   .0026586    -0.90   0.366    -.0076296    .0028198
    minority |   -.1595549   .0682725    -2.34   0.020    -.2937242   -.0253856
  nonenglish |   -.2344735   .0975351    -2.40   0.017    -.4261498   -.0427971
 tenuretrack |   -.0650419   .0579802    -1.12   0.263    -.1789848    .0489009
       lower |    .0046318   .0563316     0.08   0.935    -.1060712    .1153348
   onecredit |    .5964602   .1095069     5.45   0.000     .3812569    .8116634
       _cons |    4.259466   .1541533    27.63   0.000     3.956524    4.562408
------------------------------------------------------------------------------

.  dis "Adjusted R-squared = " e(r2_a);
Adjusted R-squared = .14282055
```

Perfect multicolinearity example:

Include both *female* and *male* = 1 – *female*:

```
. generate male = 1-female;
. reg courseevaluation btystdave female male, r;

Linear regression                               Number of obs =     463
                                                F(  2,   460) =   18.22
                                                Prob > F      =  0.0000
                                                R-squared     =  0.0663
                                                Root MSE      =  .53732
------------------------------------------------------------------------------
             |               Robust
courseeval~n |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   btystdave |   .1485876   .0321911     4.62   0.000     .0853278    .2118475
      female |  -.1978096   .0502136    -3.94   0.000    -.2964862    -.099133
        male |  (dropped)
       _cons |    4.09471   .0324991   125.99   0.000     4.030845    4.158576
------------------------------------------------------------------------------
```

F-test of joint hypotheses

Do instructor characteristics (other than *Beauty*) matter?

```
. reg courseevaluation btystdave female age minority
      nonenglish tenuretrack lower onecredit, r

Linear regression                               Number of obs =     463
                                                F(  8,   454) =   12.85
                                                Prob > F      =  0.0000
                                                R-squared     =  0.1577
                                                Root MSE      =  .51372

------------------------------------------------------------------------------
             |               Robust
courseeval~n |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   btystdave |   .1564742   .0301056     5.20   0.000     .0973106    .2156379
      female |  -.1910459   .0529734    -3.61   0.000    -.2951493   -.0869425
         age |  -.0024049   .0026586    -0.90   0.366    -.0076296    .0028198
    minority |  -.1595549   .0682725    -2.34   0.020    -.2937242   -.0253856
  nonenglish |  -.2344735   .0975351    -2.40   0.017    -.4261498   -.0427971
 tenuretrack |  -.0650419   .0579802    -1.12   0.263    -.1789848    .0489009
       lower |   .0046318   .0563316     0.08   0.935    -.1060712    .1153348
   onecredit |   .5964602   .1095069     5.45   0.000     .3812569    .8116634
       _cons |   4.259466   .1541533    27.63   0.000     3.956524    4.562408
------------------------------------------------------------------------------
```

F-test of joint hypotheses

(a) do instructor characteristics (other than *Beauty*) matter?

```
. test female age minority nonenglish tenuretrack

 ( 1)  female = 0
 ( 2)  age = 0
 ( 3)  minority = 0
 ( 4)  nonenglish = 0
 ( 5)  tenuretrack = 0

       F(  5,   454) =    7.18
            Prob > F =    0.0000
```

NOTES

The test command follows the relevant regression

There are q=5 restrictions being tested

The 5% critical value for q=5 is 2.21

Stata computes the p-value for you

(b) do course characteristics matter?

```
. test lower onecredit

 ( 1)  lower = 0
 ( 2)  onecredit = 0

       F(  2,   454) =   18.97
            Prob > F =    0.0000
```

Second F-test, for the same regression

The 5% critical value for q=2 is 3.00

The homoskedasticity-only *F*-statistic

Example: do instructor characteristics (other than *Beauty*) matter?

Unrestricted population regression (under H_1):
CourseEvaluations are a function of beauty, lower division, one-credit, and personal attributes (female, age, minority, non-English, tenuretrack)

Restricted population regression (that is, under H_0):
CourseEvaluations are a function of beauty, lower division, one-credit.

What is q?

Unrestricted regression:

```
. reg courseevaluation btystdave female age minority
      nonenglish tenuretrack lower onecredit, r

Linear regression                               Number of obs =     463
                                                F(  8,   454) =   12.85
                                                Prob > F      =  0.0000
                                                R-squared     =  0.1577
                                                Root MSE      =  .51372

------------------------------------------------------------------------------
             |               Robust
courseeval~n |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   btystdave |   .1564742   .0301056     5.20   0.000     .0973106    .2156379
      female |  -.1910459   .0529734    -3.61   0.000    -.2951493   -.0869425
         age |  -.0024049   .0026586    -0.90   0.366    -.0076296    .0028198
    minority |  -.1595549   .0682725    -2.34   0.020    -.2937242   -.0253856
  nonenglish |  -.2344735   .0975351    -2.40   0.017    -.4261498   -.0427971
 tenuretrack |  -.0650419   .0579802    -1.12   0.263    -.1789848    .0489009
       lower |   .0046318   .0563316     0.08   0.935    -.1060712    .1153348
   onecredit |   .5964602   .1095069     5.45   0.000     .3812569    .8116634
       _cons |   4.259466   .1541533    27.63   0.000     3.956524    4.562408
------------------------------------------------------------------------------
```

Restricted regression:

```
. reg courseevaluation btystdave lower onecredit, r

Linear regression                               Number of obs =     463
                                                F(  3,   459) =   19.75
                                                Prob > F      =  0.0000
                                                R-squared     =  0.1000
                                                Root MSE      =   .5281

------------------------------------------------------------------------------
             |               Robust
courseeval~n |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   btystdave |   .1468724   .0317983     4.62   0.000     .0843841    .2093608
       lower |    .033135   .0580285     0.57   0.568    -.0808995    .1471695
   onecredit |   .5762671   .1038297     5.55   0.000     .3722266    .7803076
       _cons |   3.966407   .0303433   130.72   0.000     3.906778    4.026036
------------------------------------------------------------------------------
```

$$F = \frac{(R^2_{unrestricted} - R^2_{restricted})/q}{(1 - R^2_{unrestricted})/(n - k_{unrestricted} - 1)}$$

$$= \frac{(.1577 - .1000)/5}{(1 - .1577)/(463 - 8 - 1)} = \mathbf{6.21}$$

Note: Heteroskedasticity-robust F = **7.18**

STATA .do file (linear specifications – course evaluation data)

```
clear
cd "\courses\e1123\stata\beauty"
cap log close
set more off
set scheme s1color
log using beauty_1_f18_lect2_3.log, replace
*****************************************************
*  beauty_1_f18_lect2_3.do
*  Ec1123 - fall 18
*  Course evaluation data description + basic regressions
*****************************************************
use hamermesh_beauty
describe
*****************************************************
*    Histograms
*****************************************************
histogram courseevaluation, bin(15) ///
   title("Histogram of Course Evaluation Scores")
graph export courseval_histo_all.emf, replace
histogram btystdave, bin(20) title("Histogram of Beauty Scores")
graph export beauty_histo_all.emf, replace
*****************************************************
*    differences in means
*****************************************************
summarize courseevaluation female
summarize courseevaluation if(female==0)
```

```
summarize courseevaluation if(female==1)
ttest courseevaluation, by(female) unequal
*************************************************************
*    Scatterplot and regressions
*************************************************************
reg courseevaluation btystdave, robust
 predict peval
twoway scatter courseevaluation peval btystdave, ms(O i) ///
   connect(. l) title("Scatterplot and OLS Regression Line") ///
   xtitle("Beauty") ytitle("Course Overall")
graph export beauty_1b.emf, replace
ttest courseevaluation, by(female) unequal
reg courseevaluation female, r
reg courseevaluation female
reg courseevaluation btystdave
reg courseevaluation btystdave, r
reg courseevaluation btystdave female age minority ///
        nonenglish tenuretrack lower onecredit, r
generate male = 1-female
reg courseevaluation btystdave female male, r
reg courseevaluation btystdave female age minority ///
        nonenglish tenuretrack lower onecredit, r
test female age minority nonenglish tenuretrack
test lower onecredit
* restricted regression for homoskedastic-only F-test
reg courseevaluation btystdave lower onecredit, r
test lower onecredit
*************************************************************
log close
```