LECTURE 1

Statistics Review II
Linear Regression Review I

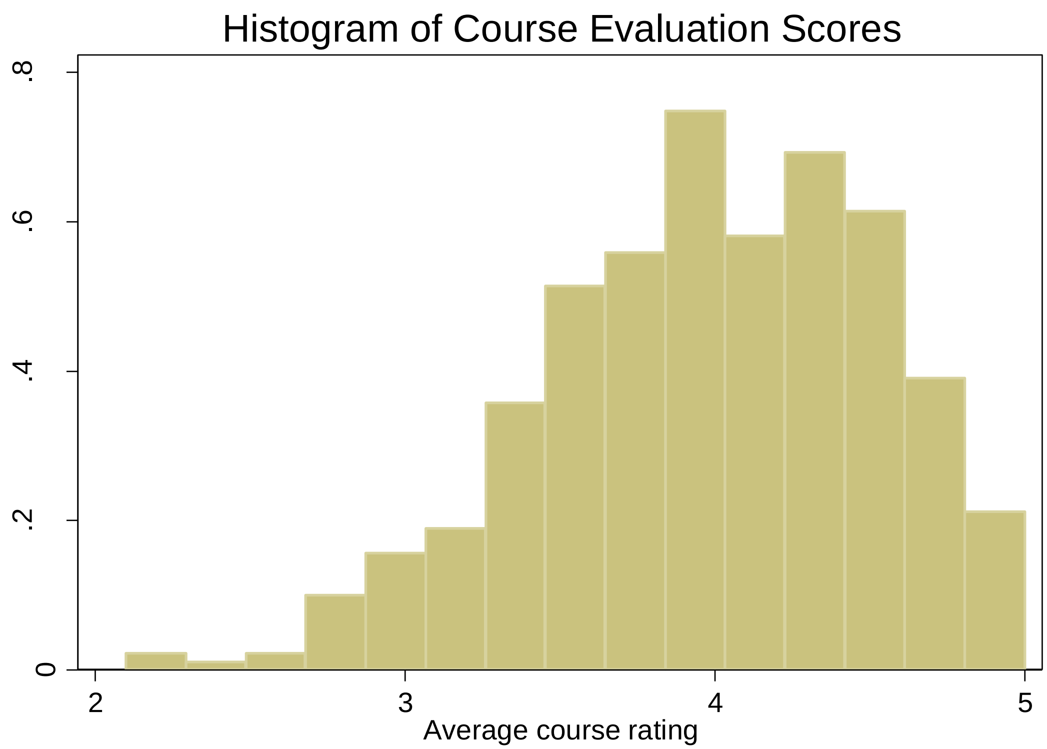
Outline

1. Statistics review (empirical example & finish)
2. Regression with one regressor
   a. Estimation: continuous regressor, discrete regressor
   b. Hypothesis tests and confidence intervals
   c. Heteroskedasticity, homoskedasticity, and HR standard errors
3. Omitted variable bias

Statistics Review: Empirical Example using STATA

Data set: U.T. Teaching evaluations

n = 463 courses at U.T. Austin, academic years 2000-2002 (Source: Hamermesh and Parker (2005))



Empirical questions

Are course evaluation scores the same on average for male and female instructors?

Let Δ = the population difference in mean scores, men – women

$= E(Y_m) - E(Y_w)$.

We are interested in:

1. Estimating Δ by the sample difference, $\hat{\Delta} = \bar{Y}_m - \bar{Y}_w$
2. Can we reject the hypothesis that male and female instructors have the same scores on average, i.e. that $\Delta = 0$?
3. Finding a 95% confidence interval for Δ

STATA output – *courseevaluation* by sex of instructor

```
Blue means you type this in

. summarize courseevaluation if(female==0)

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
courseeval~n |       268     4.06903    .5566518        2.1          5


. summarize courseevaluation if(female==1)

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
courseeval~n |       195    3.901026    .5388026        2.3        4.9
```

Question 1: Who has better evaluations – male or female instructors? What is the estimated difference ($\hat{\Delta}$) in evaluations?

Estimated difference = $\hat{\Delta} = \bar{Y}_m - \bar{Y}_w$ = 4.069 – 3.901= 0.168

Question 2: Can we reject the hypothesis that male and female instructors have the same scores on average?

To conduct this hypothesis test, compute the *t*-statistic testing the hypothesis that $\Delta = 0$:

$$t \text{ (testing } \Delta=0) = \frac{\bar{Y}_w - \bar{Y}_m}{SE(\bar{Y}_w - \bar{Y}_m)}$$

We need to compute the standard error of $\hat{\Delta}$, $SE(\hat{\Delta})$:

$$SE(\bar{Y}_m - \bar{Y}_w) = \sqrt{\frac{s_m^2}{n_m} + \frac{s_w^2}{n_w}}$$

```
. summarize courseevaluation if(female==0)

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
courseeval~n |       268     4.06903    .5566518        2.1          5

. summarize courseevaluation if(female==1)

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
courseeval~n |       195    3.901026    .5388026        2.3        4.9
```

$$SE(\bar{Y}_m - \bar{Y}_w) = \sqrt{\frac{s_m^2}{n_m} + \frac{s_w^2}{n_w}} = \sqrt{\frac{0.5567^2}{268} + \frac{0.5388^2}{195}} = 0.0514$$

$$t \text{ (testing } \Delta=0) = \frac{\bar{Y}_w - \bar{Y}_m}{SE(\bar{Y}_w - \bar{Y}_m)} = 0.168/0.0514 = 3.27$$

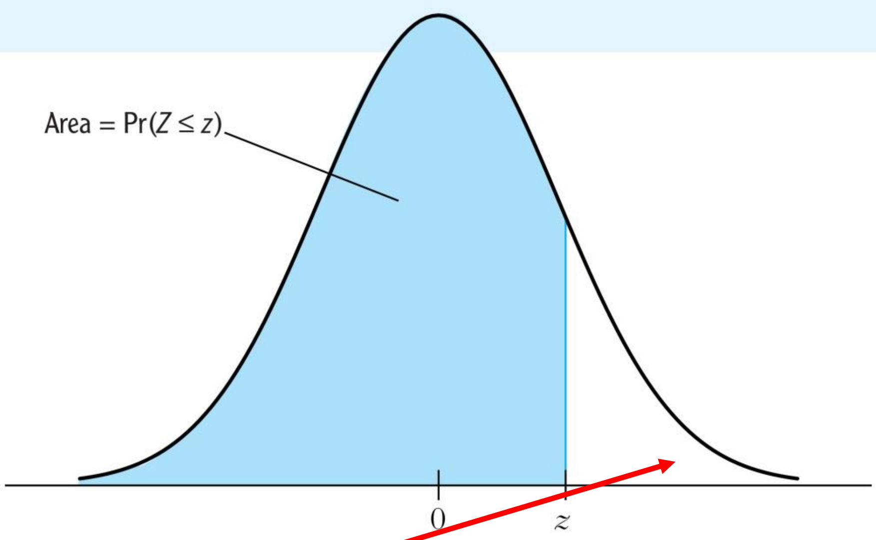
Two methods to evaluate this t-statistic:

a) compare it to 1.96

b) compute the p-value:

p-value = Pr(|z| > 3.27) = 0.0011 = 0.11%

TABLE 1 The Cumulative Standard Normal Distribution Function, $\Phi(z) = \Pr(Z \leq z)$



area for $z > 3.27$ is 0.0006 = 0.06% = $\Pr(z > 3.27)$

p-value = $\Pr(|z| > 3.27) = 2\times\Pr(z > 3.27)$

$= 2\times0.06\% = 0.12\%$

(different from 0.11% due to rounding)

Question 3: What is the 95% confidence interval for this difference?

95% confidence interval = $\hat{\Delta} \pm 1.96 \times SE(\hat{\Delta})$

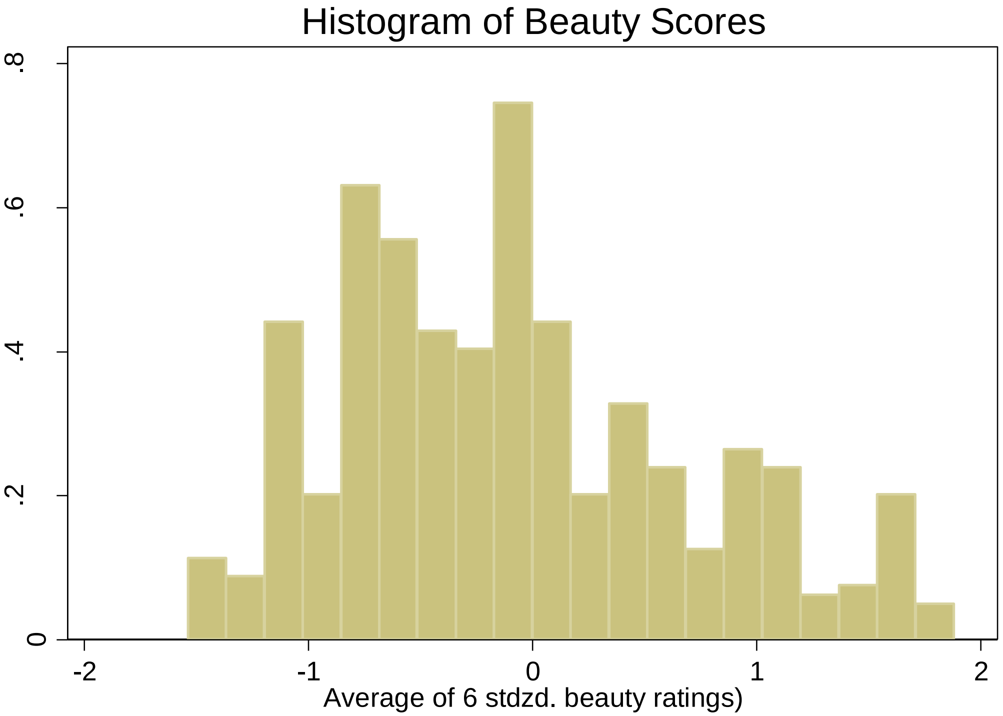
$\hat{\Delta} \pm 1.96 \times SE(\hat{\Delta}) = 0.168 \pm 1.96 \times 0.0514 = (0.067, 0.269)$

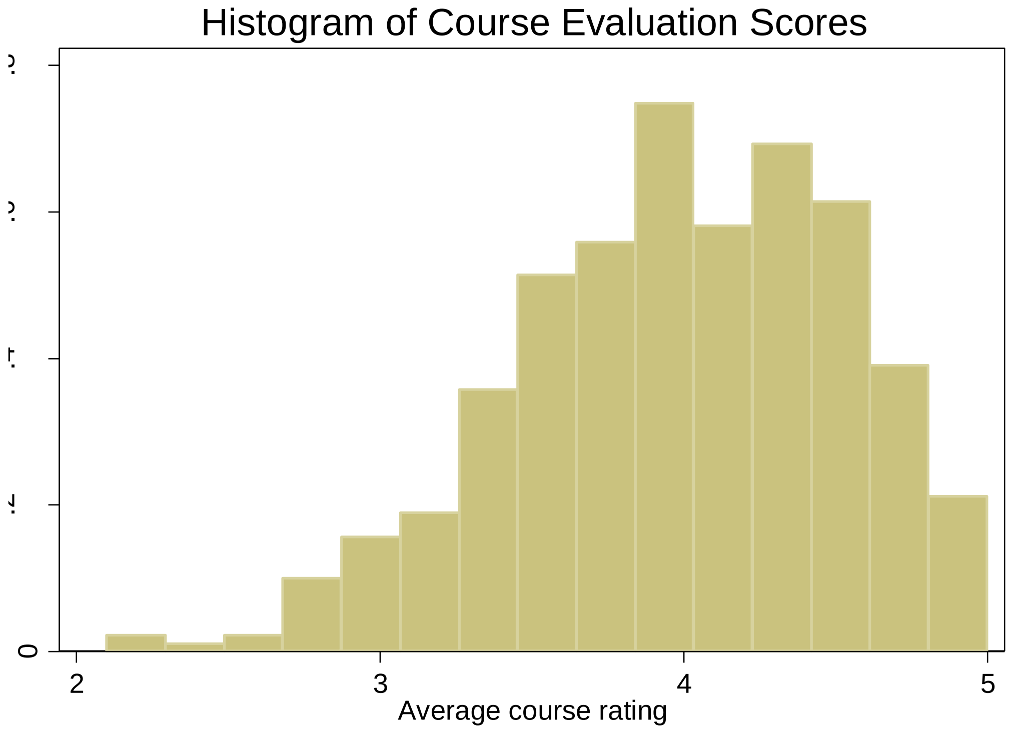
These calculations, done using **ttest** in STATA:

```
. ttest courseevaluation, by(female) unequal;

Two-sample t test with unequal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |     268     4.06903    .0340029    .5566518    4.002082    4.135978
       1 |     195    3.901026    .0385845    .5388026    3.824927    3.977125
---------+--------------------------------------------------------------------
combined |     463    3.998272    .0257868    .5548656    3.947598    4.048946
---------+--------------------------------------------------------------------
    diff |            .1680042    .0514292                .0669175    .2690909
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =   3.2667
Ho: diff = 0                             Satterthwaite's degrees of freedom =  425.756

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.9994         Pr(|T| > |t|) = 0.0012          Pr(T > t) = 0.0006
```

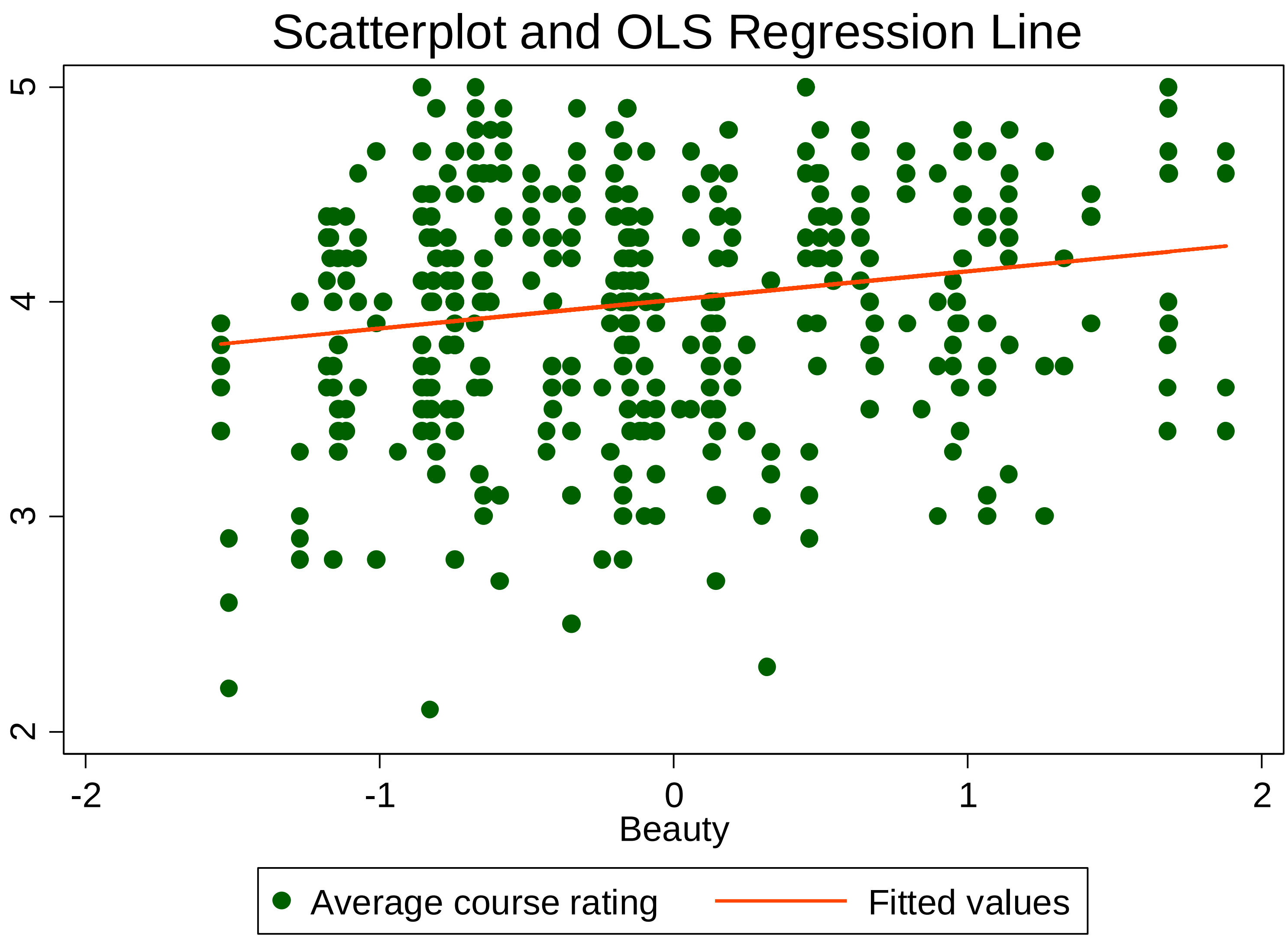
Regression with a single regressor

Is instructor attractiveness related to course evaluations?





Course evaluations and beauty scores:

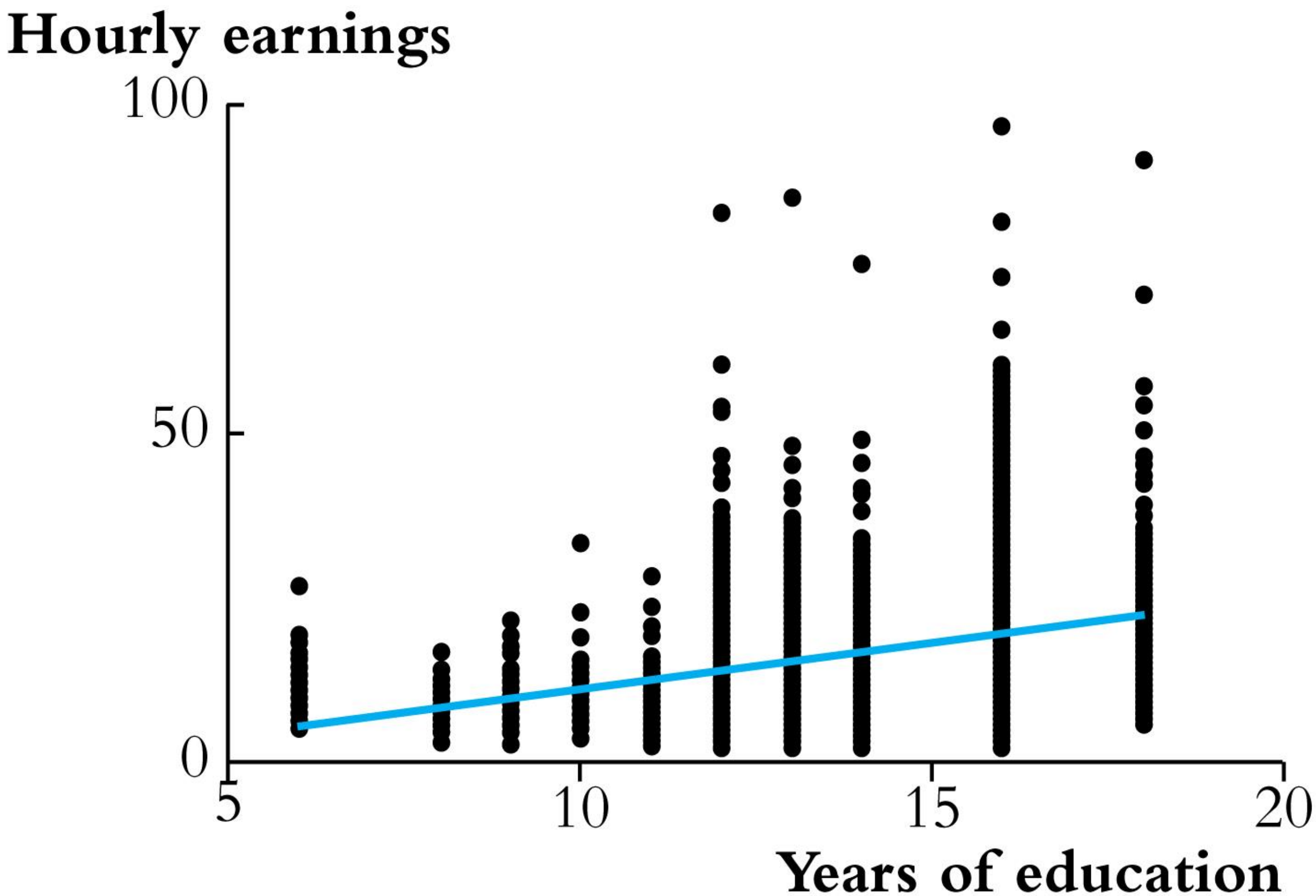


Homoskedastic or heteroskedastic?

```
. reg courseevaluation btystdave

      Source |       SS       df       MS              Number of obs =     463
-------------+------------------------------           F(  1,   461) =   17.08
       Model |  5.08300724     1  5.08300724           Prob > F      =  0.0000
    Residual |  137.155613   461  .297517599           R-squared     =  0.0357
-------------+------------------------------           Adj R-squared =  0.0336
       Total |   142.23862   462  .307875801           Root MSE      =  .54545

------------------------------------------------------------------------------
courseeval~n |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   btystdave |   .1330014   .0321775     4.13   0.000     .0697687    .1962342
       _cons |   4.010023   .0255082   157.21   0.000     3.959896    4.060149
------------------------------------------------------------------------------

. reg courseevaluation btystdave, robust

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
```

```
. reg courseevaluation female

      Source |       SS       df       MS              Number of obs =     463
-------------+------------------------------           F(  1,   461) =   10.56
       Model |  3.18587533     1  3.18587533           Prob > F      =  0.0012
    Residual |  139.052745   461  .301632852           R-squared     =  0.0224
-------------+------------------------------           Adj R-squared =  0.0203
       Total |   142.23862   462  .307875801           Root MSE      =  .54921

------------------------------------------------------------------------------
courseeval~n |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -.1680042   .0516946    -3.25   0.001    -.2695905    -.066418
       _cons |    4.06903   .0335484   121.29   0.000     4.003103    4.134957
------------------------------------------------------------------------------

. reg courseevaluation female, r

Linear regression                                      Number of obs =     463
                                                       F(  1,   461) =   10.67
                                                       Prob > F      =  0.0012
                                                       R-squared     =  0.0224
                                                       Root MSE      =  .54921

------------------------------------------------------------------------------
             |               Robust
courseeval~n |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -.1680042   .0514241    -3.27   0.001    -.2690588   -.0669496
       _cons |    4.06903    .034013   119.63   0.000      4.00219     4.13587
------------------------------------------------------------------------------
```

Average hourly earnings vs. years of education (data source: Current Population Survey):



Homoskedastic or heteroskedastic?

Omitted Variable Bias

- In multiple regression, β_1 is the effect of X_1 holding other X's constant.
- The main reason to include additional X's is if they co-vary with X_1 – in which case they would be confounding factors if they are omitted
- The bias in the OLS estimator that occurs as a result 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.* $\text{corr}(Z,X) \neq 0$)

- 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

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

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,

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

$$\xrightarrow{p} \frac{\text{cov}(X_i, u_i)}{\text{var}(X_i)} = \frac{\sigma_{Xu}}{\sigma_X^2}$$ ("$\xrightarrow{p}$" means large-n 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 ρ_{Xu} = 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.