Previously on Quantitative Methods…

- Linear regression model in the population:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

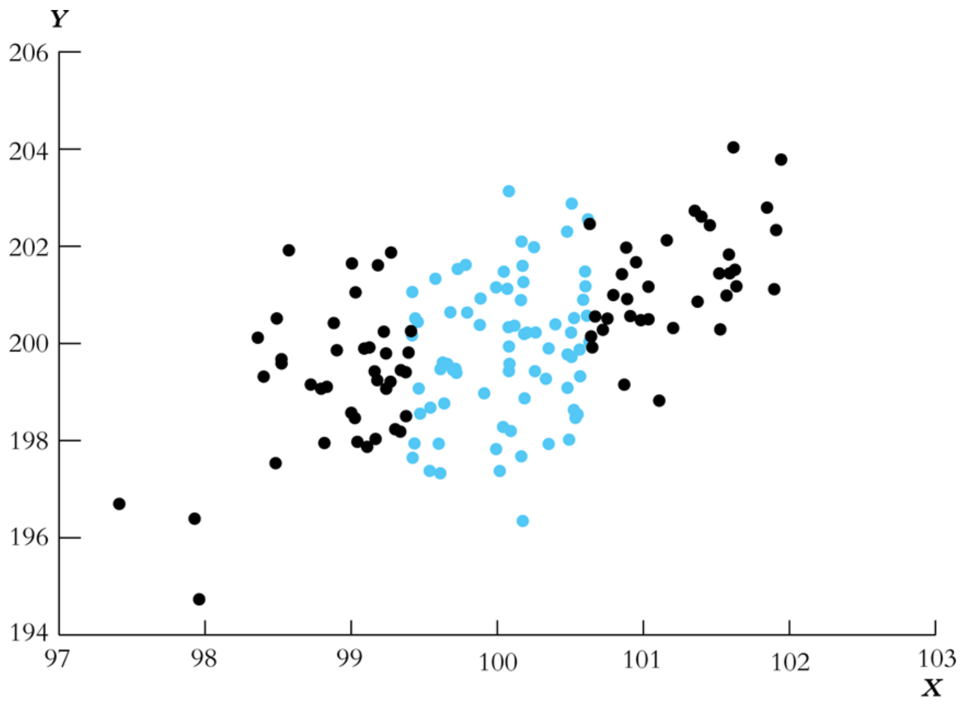
OLS estimators of β_0 & β_1

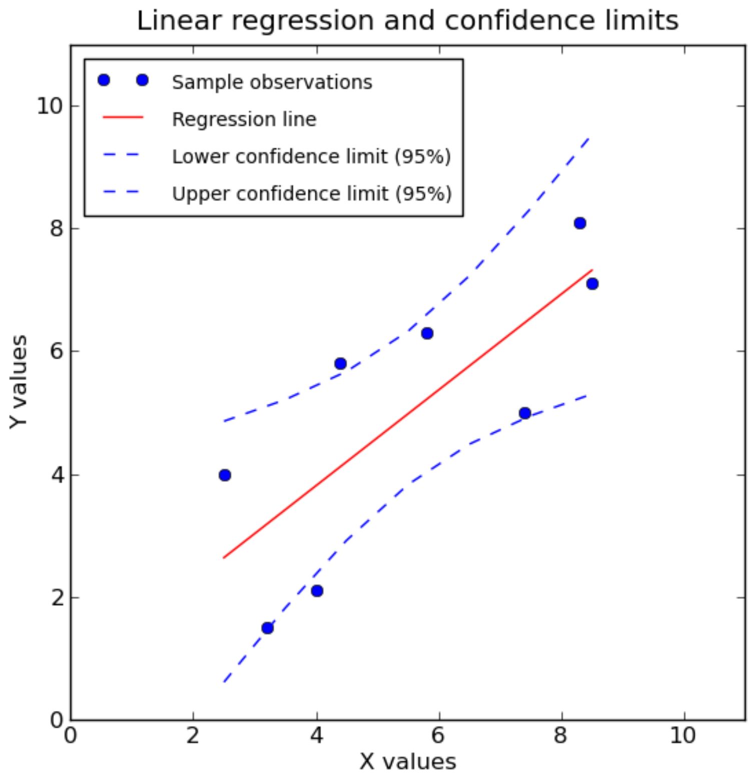
- OLS estimator from sample data:

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{u}_i$$

- $\hat{\beta}_1$ estimates a *causal effect* of X on Y *only if* $corr(X_i, u_i) = 0$.
  - No confounding factors affecting both X & Y, and no reverse causality Y->X.
  - (…and sample is random, and outliers are rare)

2. Statistical inference about linear regression





Inference about linear regression

- Test hypotheses about population coefficients β_0 & β_1.

- Build confidence intervals about β_0 & β_1.
  - A range of values with (say) 95% probability of including true coefficients.

Hypothesis tests

- Null and alternative hypotheses:

$$H_0: \beta_1 = \beta_{1,0} \text{ vs. } H_1: \beta_1 \neq \beta_{1,0}$$

- Three steps for testing H_0 :

1. Compute $\hat{\beta}_1$ and $SE(\hat{\beta}_1)$ using sample data.
2. Compute the t-statistics
3. Compute the p-value

Hypothesis tests

- Null and alternative hypotheses:

$$H_0\!: \beta_1 = \beta_{1,0} \text{ vs. } H_1\!: \beta_1 \neq \beta_{1,0}$$

- Three steps for testing H_0 :

1. Compute $\hat{\beta}_1$ and $SE(\hat{\beta}_1)$ using sample data.
2. Compute the t-statistics
3. Compute the p-value

The standard error of $\hat{\beta}_1$

- $SE(\hat{\beta}_1)$ is an estimator of $\sigma_{\hat{\beta}_1}$.
- $SE(\hat{\beta}_1) = \sqrt{\hat{\sigma}^2_{\hat{\beta}_1}} = \frac{1}{n} \times \frac{\frac{1}{n-2}\sum_{i=1}^{n}(X_i - \bar{X})^2 \hat{u}_i^2}{\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right]^2}$
- (complicated, but STATA will do it for you)
- Also called *robust* standard error.
- To obtain this type of SE in STATA, you use the 'robust' option.

```
regress testscr str, robust

Regression with robust standard errors               Number of obs =      420
                                                     F(  1,   418) =    19.26
Prob > F        =  0.0000
                                                     R-squared     =   0.0512
                                                     Root MSE      =   18.581

------------------------------------------------------------------------
        |               Robust
testscr |      Coef.   Std. Err.       t    P>|t|     [95% Conf. Interval]
--------+---------------------------------------------------------------
    str |  -2.279808   .5194892    -4.39   0.000    -3.300945   -1.258671
  _cons |    698.933   10.36436    67.44   0.000     678.5602    719.3057
------------------------------------------------------------------------
```

Standard error for slope $SE(\hat{\beta}_1)$

Standard error for intercept $SE(\hat{\beta}_0)$

- $\hat{\beta}_1 = -2.28$ and $SE(\hat{\beta}_1) = 0.52$
- $\hat{\beta}_0 = 698.9$ and $SE(\hat{\beta}_1) = 10.36$

HYPOTHESIS TESTS

- Null and alternative hypotheses:

$$H_0: \beta_1 = \beta_{1,0} \text{ vs. } H_1: \beta_1 \neq \beta_{1,0}$$

- Three steps for testing H_0 :

1. Compute $\hat{\beta}_1$ and $SE(\hat{\beta}_1)$ using sample data.
2. Compute the t-statistics
3. Compute the p-value

t-statistics for OLS estimated coefficients

$$t = \frac{estimated\ coeff. - hypothesized\ value}{standard\ error\ of\ estimator}$$

$$t = \frac{\hat{\beta}_1 - \hat{\beta}_{1,0}}{SE(\hat{\beta}_1)}$$

- t has a *standard normal distribution* in large samples
- $t \sim N(0,1)$

```
regress testscr str, robust

Regression with robust standard errors               Number of obs =     420
                                                      F(  1,   418) =   19.26
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.0512
                                                      Root MSE      =  18.581

-------------------------------------------------------------------------
         |               Robust
 testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------+---------------------------------------------------------------
     str |  -2.279808   .5194892    -4.39   0.000    -3.300945   -1.258671
   _cons |    698.933   10.36436    67.44   0.000     678.5602    719.3057
-------------------------------------------------------------------------
```

t-stat for $H_0: \beta_1 = 0$

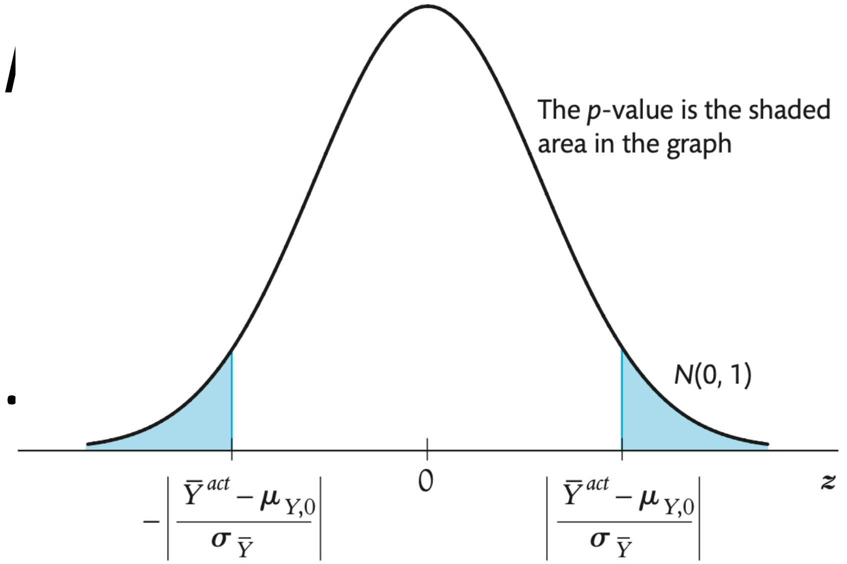
t-stat for $H_0: \beta_0 = 0$

- $\hat{\beta}_1 = -2.28$ and $SE(\hat{\beta}_1) = 0.52$ and $t = -4.39$
- $\hat{\beta}_0 = 698.9$ and $SE(\hat{\beta}_1) = 10.36$ and $t = 67.44$

Hypothesis tests

- Null and alternative hypotheses:

$$H_0: \beta_1 = \beta_{1,0} \text{ vs. } H_1: \beta_1 \neq$$



- Three steps for testing H_0 :

1. Compute $\hat{\beta}_1$ and $SE(\hat{\beta}_1)$ using sample data.
2. Compute the t-statistics
3. Compute the p-value

Computing the p-value

- p-value $= Pr_{H_0}[|\hat{\beta}_1 - \beta_{1,0}| > |\hat{\beta}_1^{act} - \hat{\beta}_{1,0}|$

"Probability under the null hypothesis…

…that the difference between the estimated coefficient and the null hypothesis…

…is at least as large as the one we obtained in our sample."

$$= 2\phi(-|t^{act}|)$$

```
regress testscr str, robust

Regression with robust standard errors               Number of obs =     420
                                                      F(  1,   418) =   19.26
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.0512
                                                      Root MSE      =  18.581

-------------------------------------------------------------------------
        |               Robust
testscr |      Coef.   Std. Err.       t     P>|t|      [95% Conf. Interval]
--------+----------------------------------------------------------------
    str |  -2.279808   .5194892     -4.39    0.000     -3.300945   -1.258671
  _cons |    698.933   10.36436     67.44    0.000      678.5602    719.3057
-------------------------------------------------------------------------
```

p-value for $H_0: \beta_1 = 0$

p-value for $H_0: \beta_0 = 0$

- $\hat{\beta}_1 = -2.28$ and $SE(\hat{\beta}_1) = 0.52$ and $t = -4.39$ and $p < 0.001$
- $\hat{\beta}_0 = 698.9$ and $SE(\hat{\beta}_1) = 10.36$ and $t = 67.44$ and $p < 0.001$

Confidence interval for β_1

- **95% confidence interval:** a range of values that is 95% likely to include the "true" population coefficient $\boldsymbol{\beta_1}$.

- The set of β_1 values that we *cannot* reject at the 5% significance level.

- 95% confidence interval for β_1:

$$\hat{\beta}_1 - 1.96 * SE\left(\hat{\beta}_1\right) \leq \beta_1 \leq \hat{\beta}_1 + 1.96 * SE\left(\hat{\beta}_1\right)$$

```
regress testscr str, robust

Regression with robust standard errors               Number of obs =     420
                                                     F(  1,   418) =   19.26
                                                     Prob > F      =  0.0000
                                                     R-squared     =  0.0512
                                                     Root MSE      =  18.581

-------------------------------------------------------------------------
        |               Robust
testscr |      Coef.   Std. Err.       t    P>|t|     [95% Conf. Interval]
--------+----------------------------------------------------------------
    str |  -2.279808   .5194892    -4.39   0.000    -3.300945   -1.258671
  _cons |    698.933   10.36436    67.44   0.000     678.5602    719.3057
-------------------------------------------------------------------------
```

- Confidence interval for β_1: [-3.30 $\leq \beta_1 \leq$ -1.26]

Confidence interval for predicted effects

- Confidence interval for the effect of a Δx change in X:

$$\left[\,(\hat{\beta}_1\ lower\ bound) \times \Delta x\ \ ;\ (\ \hat{\beta}_1\ upper\ bound) \times \Delta x\,\right]$$

$$\left[\,(\hat{\beta}_1 - 1.96 * SE(\hat{\beta}_1)) \times \Delta x\ \ ;\ \hat{\beta}_1 + 1.96 * SE(\hat{\beta}_1) \times \Delta x\,\right]$$

Example: Confidence interval for the average effect of a 3.5 increase in STR

- Confidence interval for β_1(coefficient of STR): [-3.30 $\leq \beta_1 \leq$ -1.26]
- Confidence interval for 3.5 increase in STR:
  - Lower bound: -3.30 * 3.5 = -11.55
  - Upper bound: -1.26 * 3.5 = -4.41
- An increase in STR by 3.5 students is associated with a decrease in test scores between 4.41 and 11.55 points. [-11.55 $\leq \beta_1 \leq$ -4.41]

Regression with binary regressor

- Binary (or *indicator* or *dummy*) variables
  - Sex at birth (1 = female; 0 = male)
  - Urban or rural (1 = urban; 0 = rural)
  - Treatment or placebo
    (1 = treatment; 0 = placebo)
  - ….

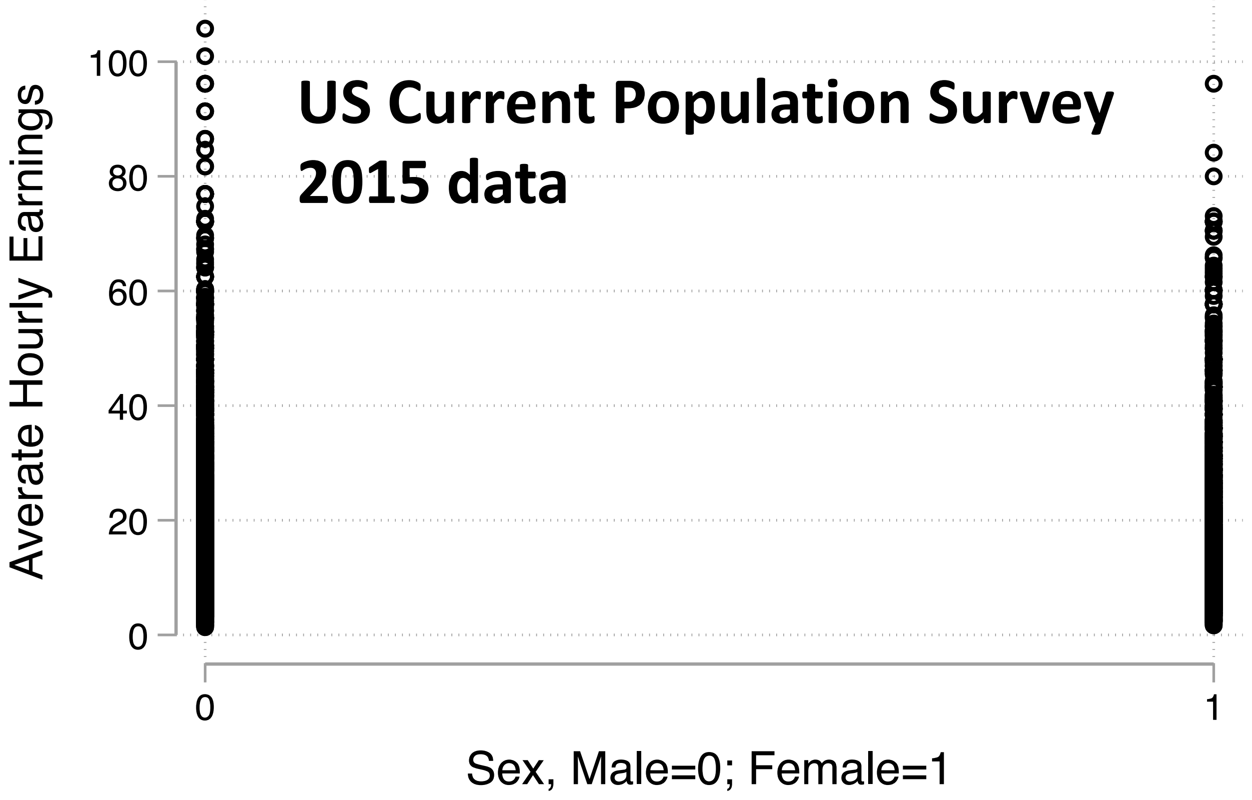
Example: the gender pay gap

Y = Average Hourly Earnings (*AHE*)

D = Sex at birth (*Female*)

$$Y_i = \beta_0 + \beta_1 D_i + u_i$$

How do we interpret β_1?



in STATA: *scatter ahe female*

Regression with binary regressor

$$E(Y|D) = \beta_0 + \beta_1 D$$

$$Y_i = \beta_0 + \beta_1 D_i + u_i$$

For a male worker ($D_i = 0$):

$$E(Y|D = 0) = \beta_0 + \beta_1 \times 0 = \beta_0$$

For a female worker ($D_i = 1$):

$$E(Y|D = 1) = \beta_0 + \beta_1 \times 1 = \beta_0 + \beta_1$$

$$\Downarrow$$

$$\beta_1 = E(Y|D = 1) - E(Y|D = 0)$$

Regression with binary regressor

$$Y_i = \beta_0 + \beta_1 D_i + u_i$$

- $\hat{\beta}_0$ = sample mean of Y when D=0
- $\hat{\beta}_0 + \hat{\beta}_1$ = sample mean of Y when D=1
- $\hat{\beta}_1$= difference in group means
- T-stats, p-value, confidence intervals calculated as usual.
- Will give the same result as a t-test for difference in means.

Example: the gender pay gap

```
. reg ahe female, robust

Linear regression                               Number of obs     =     13,201
                                                F(1, 13199)       =     184.93
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0131
                                                Root MSE          =     10.695

------------------------------------------------------------------------------
             |               Robust
         ahe | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      female |  -2.495648   .1835205   -13.60   0.000    -2.855375   -2.135922
       _cons |   18.32845   .1300679   140.91   0.000      18.0735     18.5834
------------------------------------------------------------------------------
```

- AHE for men (D=0):

$$\hat{\beta}_0 = 18.33$$

- Difference between women and men:

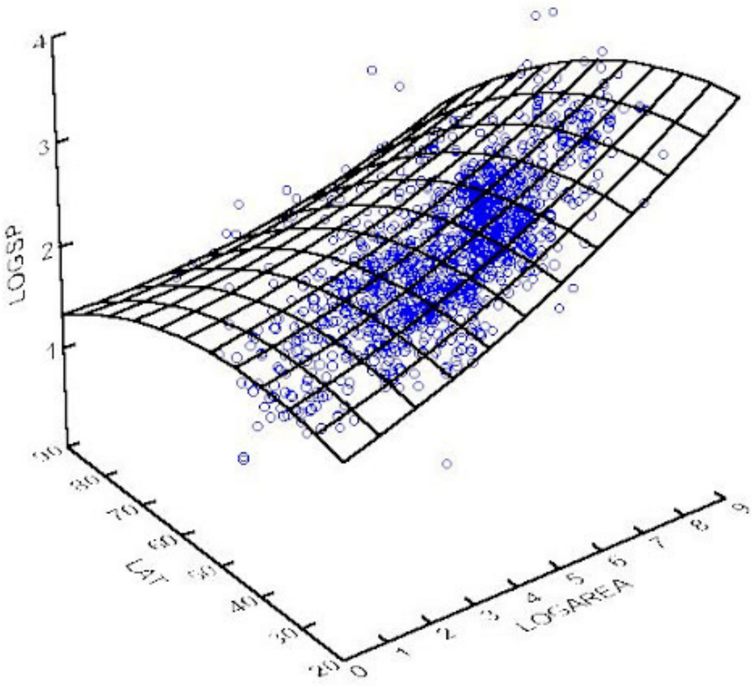
$$\hat{\beta}_1 = -2.50$$

- AHE for women (D=1):

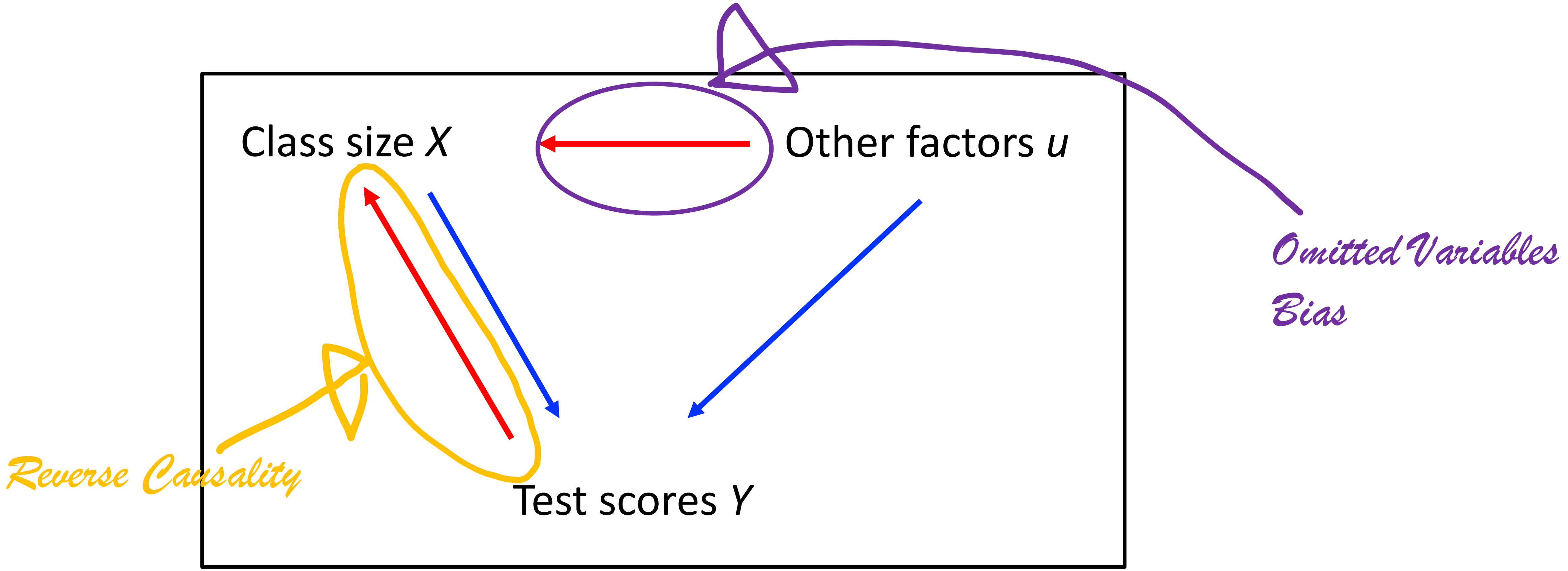
$$\hat{\beta}_0 + \hat{\beta}_1 =$$
$$= 18.32 - 2.50 = 15.83$$

(US Current Population Survey 2015 data)

3. Linear regression with multiple regressors



CAUSAL RELATIONS BETWEEN CLASS SIZE & TEST SCORES



Omitted variables bias

Omitted Variables Bias (OVB) occurs if:

1. The omitted variable is correlated with the included regressor X.

   AND

2. The omitted variable affects the dependent variable Y.

Thank you for your attention