Econometrics (Econ 452) – Fall 2022 – Instructor: Daniele Girardi

5 – LINEAR REGRESSION II
MULTIPLE REGRESSORS

University of Massachusetts Amherst BE REVOLUTIONARY™

THE PLAN

1. Omitted Variable Bias
2. The Multiple Regression Model
3. OLS Estimation of the Multiple Regression Model
4. Measures of Fit in Multiple Regression
5. Multiple Regression and Causality: Control Variables & the CIA
6. Multicollinearity
7. Statistical Inference about a single coefficient
8. Statistical Inference about multiple coefficients at the same time
9. Model specification and presentation

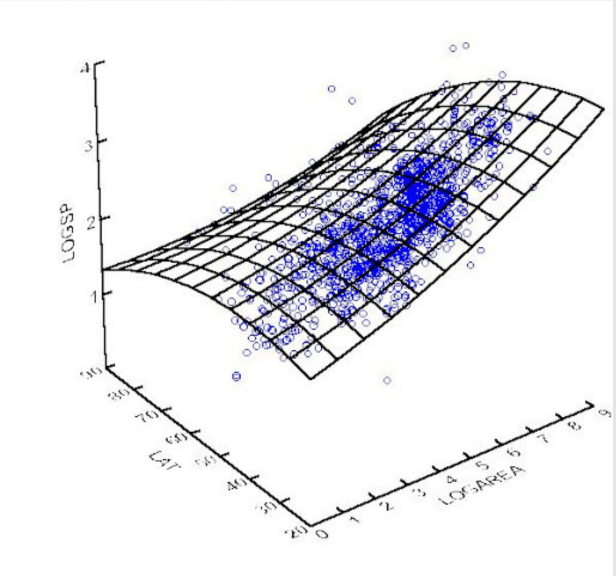
Motivation

Estimation

Causal Inference

Statistical Inference

MULTIPLE LINEAR REGRESSION: OVERVIEW



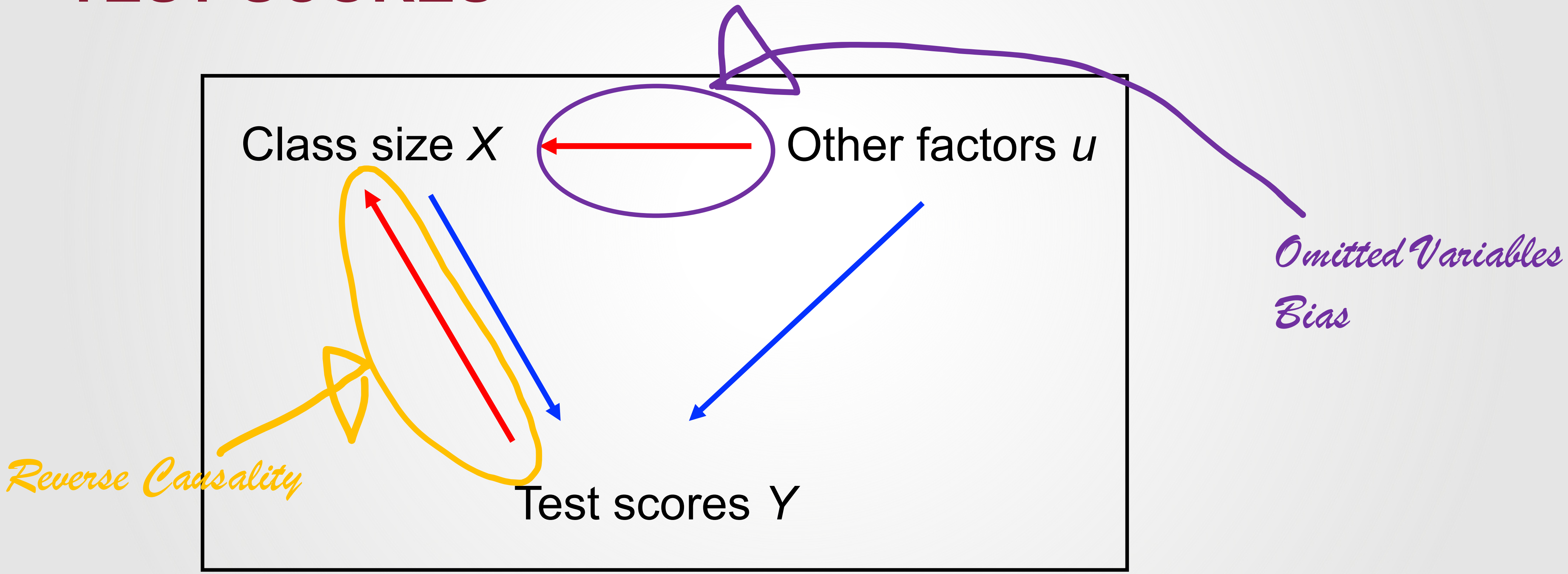
Why multiple regressors *at the same time*?

- Prediction → increase accuracy.
- Causal inference → *control for* confounding factors.

5.1 OMITTED VARIABLES BIAS (OVB)

University of Massachusetts Amherst

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.

OMITTED VARIABLES BIAS (OVB)

- Linear regression model:

$$TestScores_i = \beta_0 + \beta_1 STR + u_i$$

- Do these variables cause OVB?

1. Financial resources of the school district.
2. Outside temperature during the test.
3. Average parking lot space.
4. Percentage of English learners

OMITTED VARIABLES BIAS (OVB)

- Let $\beta_1^\star$ be the true causal effect of *X* on *Y* in the population.
- Let $\rho_{Xu} = corr(X_i, u_i)$
- OLS coefficient gives you:

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

(proof in Appendixes 4.3 & 6.1)

OMITTED VARIABLES BIAS (OVB)

- Y = dependent variable
- X = independent variable
- Z = omitted variable

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

	$Corr(Z,X) > 0$	$Corr(Z,X) < 0$
Z increases Y (& u_i)		
Z decreases Y (& u_i)		

OMITTED VARIABLES BIAS (OVB)

- Y = dependent variable
- X = independent variable
- Z = omitted variable

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

	$Corr(Z,X) > 0$	$Corr(Z,X) < 0$
Z increases Y (& u_i)	Upward bias ↑	
Z decreases Y (& u_i)		

OMITTED VARIABLES BIAS (OVB)

- Y = dependent variable
- X = independent variable
- Z = omitted variable

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

	$Corr(Z,X) > 0$	$Corr(Z,X) < 0$
Z increases Y (& u_i)	Upward bias ↑	Downward bias ↓
Z decreases Y (& u_i)		

OMITTED VARIABLES BIAS (OVB)

- Y = dependent variable
- X = independent variable
- Z = omitted variable

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

	$Corr(Z,X) > 0$	$Corr(Z,X) < 0$
Z increases Y (& u_i)	Upward bias ↑	Downward bias ↓
Z decreases Y (& u_i)	Downward bias ↓	

OMITTED VARIABLES BIAS (OVB)

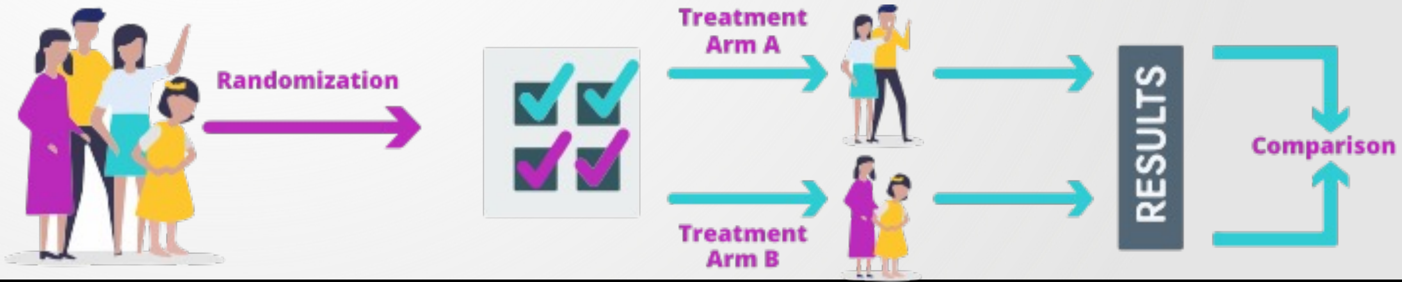
- Y = dependent variable
- X = independent variable
- Z = omitted variable

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

	$Corr(Z,X) > 0$	$Corr(Z,X) < 0$
Z increases Y (& u_i)	Upward bias ↑	Downward bias ↓
Z decreases Y (& u_i)	Downward bias ↓	Upward bias ↑

RANDOMIZATION AS A SOLUTION

- Randomized Controlled Trials (RCTs).
- Random assignment of X → no OVB (& no reverse causality).
- Same $E(X)$ for all units, independent of other factors affecting Y.
- $E(u)$ does not vary with X → $corr(X, u) = 0$.



“CONTROLLING FOR” OMITTED VARIABLES

- Observational data → no guarantee that $corr(X, u) = 0$.
- But if we can observe the omitted variables that affect both Y and X, we can try to “control for” them.
- Compare Y between units with similar levels of Z but different levels of X.

"CONTROLLING FOR" OMITTED VARIABLES

TABLE 6.1 Differences in Test Scores for California School Districts with Low and High Student–Teacher Ratios, by the Percentage of English Learners in the District

	Student–Teacher Ratio < 20		Student–Teacher Ratio ≥ 20		Difference in Test Scores, Low vs. High Student–Teacher Ratio	
	Average Test Score	*n*	Average Test Score	*n*	Difference	*t*-statistic
All districts	657.4	238	650.0	182	7.4	4.04
Percentage of English learners						
< 1.9%	664.5	76	665.4	27	−0.9	−0.30
1.9–8.8%	665.2	64	661.8	44	3.3	1.13
8.8–23.0%	654.9	54	649.7	50	5.2	1.72
> 23.0%	636.7	44	634.8	61	1.9	0.68

5.2 THE MULTIPLE REGRESSION MODEL

University of Massachusetts Amherst

MULTIPLE REGRESSION MODEL WITH 2 REGRESSORS

$$E(Y_i|X_1, X_2) = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i}$$

⬇

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + u_i$$

- How do you interpret β_1?

MULTIPLE REGRESSION MODEL WITH 2 REGRESSORS

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + u_i$$

- $\beta_1 = \frac{\Delta Y}{\Delta X_i}, holding\ X_2\ constant.$
- *Partial effect* of X_1
- How do you interpret β_2? and β_0? and u_i?

"CONTROLLING FOR" OMITTED VARIABLES

- Multiple regression model with k regressors:

$$E(Y_i|X_1, X_2, \ldots, X_n) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki}$$

$$\Downarrow$$

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \cdots + \beta_k X_{k,i} + u_i$$

5.3 OLS ESTIMATION OF THE MULTIPLE REGRESSION MODEL

University of Massachusetts Amherst

OLS ESTIMATION OF MULTIPLE REGRESSION

- OLS strategy: Select $\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k$ to *best fit* the data.
- Best fit the data = minimize (squared) prediction errors:

$$\min_{b_0, b_1, \dots, b_k} \sum_{i=1}^{n} \left(Y_i - [b_0 + b_1 X_{i,1} + b_2 X_{i,2} + \cdots + b_k X_{k,1}]\right)^2$$

- OLS estimators $(\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k)$ = the values of $b_0, b_1, \dots, b_k$ that minimize this expression

OLS ESTIMATOR OF MULTIPLE REGRESSION

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_k X_{ki} + \hat{u}_i$$

- Linear multiple regression model…
- …but with sample OLS coefficients $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ as estimators of population coefficients $\beta_0, \beta_1, \ldots, \beta_k$.
- $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_k X_{ki}$ = predicted value
- $\hat{u}_i = Y_i - \hat{Y}_i$ = regression residual (estimator of error term u_i)

THE FRISCH-WAUGH-LOVELL THEOREM

- With one regressor ($Y_i = \beta_0 + \beta_1 X_i + u_i$):

$$\hat{\beta}_1 = \frac{cov(X,Y)}{var(X)}$$

- With multiple regressors ($Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + u_i$):

$$\hat{\beta}_1 = \frac{cov(\tilde{X}_1, \tilde{Y})}{var(\tilde{X}_1)}$$

- $\tilde{X}_1 =$ residual from regression of X_1 on all other regressors ($X_2, \dots, X_k$).
- $\tilde{Y} =$ residual from regression of Y on all other regressors ($X_2, \dots, X_k$).

THE FRISCH-WAUGH-LOVELL THEOREM

- FWL theorem means that you can compute $\hat{\beta}_1$ in 3 steps:

1. Regress X_1 on $X_2, X_3, \ldots, X_k$ and obtain residuals $\tilde{X}_1$.
2. Regress Y_1 on $X_2, X_3, \ldots, X_k$ and obtain residuals $\tilde{Y}_1$.
3. Regress $\tilde{Y}_1$ on $\tilde{X}_1$
   - $\tilde{Y}_i = \beta_0 + \beta_1 \tilde{X}_i + u_i$

EXAMPLE: CLASS SIZE & TEST SCORES

- Back to our dataset of 420 California school districts in 1999.
- We estimated:

$$\widehat{TestScore} = 698.9 - 2.28 \times STR$$

- Now include percent English Learners in the district (*PctEL*):

$$\widehat{TestScore} = 686.0 - 1.10 \times STR - 0.65 \times PctEL$$

- What happened to the coefficient on STR? Why?

MULTIPLE REGRESSION IN STATA

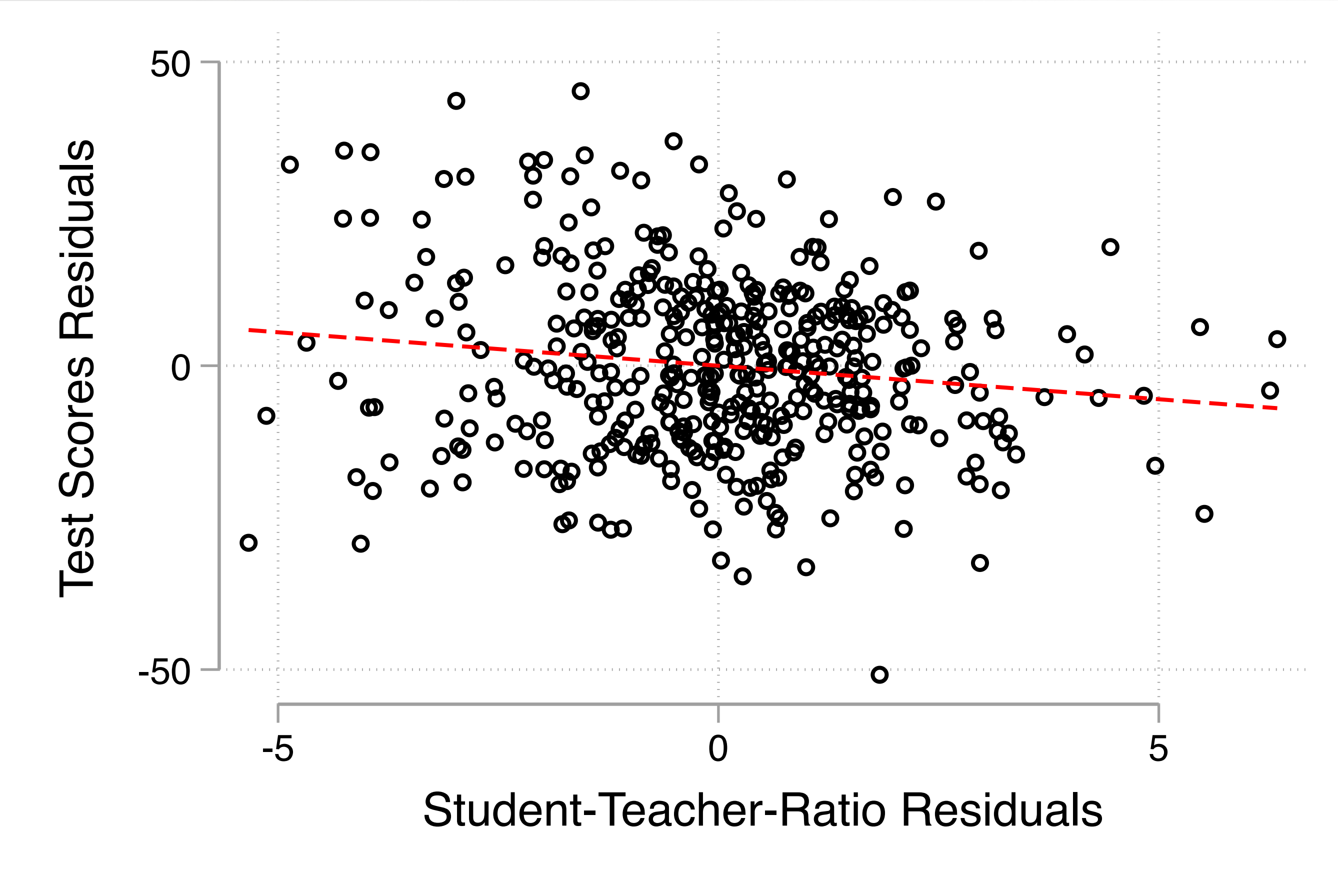
```
reg testscr str pctel, robust

Regression with robust standard errors                 Number of obs =     420
                                                       F(  2,   417) =  223.82
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.4264
                                                       Root MSE      =  14.464

------------------------------------------------------------------------------
             |               Robust
     testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         str |  -1.101296   .4328472    -2.54   0.011    -1.95213   -.2504616
       pctel |  -.6497768   .0310318   -20.94   0.000    -.710775   -.5887786
       _cons |   686.0322   8.728224    78.60   0.000    668.8754     703.189
------------------------------------------------------------------------------
```

$$\widehat{TestScore} = 686.0 - 1.10 \times STR - 0.65 \times PctEL$$

PICTURING MULTIPLE REGRESSION COEFFICIENTS: A "RESIDUALIZED" SCATTERPLOT



- What is the slope of this regression line equal to?
- Application of Frisch-Waugh-Lovell!

5.4 MEASURES OF FIT IN MULTIPLE REGRESSION

University of Massachusetts Amherst

MEASURES OF FIT IN MULTIPLE REGRESSION

1. Standard Error of the Regression (SER)
2. R^2
3. Adjusted R^2

SER

- Measures the spread of Y_i around the regression line.
- How far from the regression line is the "typical" unit?

$$SER = \sqrt{\frac{1}{n-k-1}\sum_{i=1}^{n}\hat{u}_i^2}$$

- Note: "Root MSE" in STATA regression output is *basically* the SER.

R² & ADJUSTED R²

- $R^2 = \frac{ESS}{TSS} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$
- Equivalently, $R^2 = 1 - \frac{SSR}{TSS} = 1 - \frac{\sum_{i=1}^{n} \hat{u}_i^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$
- Always increases if you add regressors.
- *Adjusted* $R^2 (or\ \bar{R}^2) = 1 - \frac{n-1}{n-k-1} \frac{SSR}{TSS}$

MEASURES OF FIT IN MULTIPLE REGRESSION

```
reg testscr str pctel, robust

Regression with robust standard errors               Number of obs =     420
                                                 F(  2,   417) =  223.82
                                                 Prob > F      =  0.0000
                                                 R-squared     =  0.4264
                                                 Root MSE      =  14.464

------------------------------------------------------------------------------
             |               Robust
     testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         str |  -1.101296   .4328472    -2.54   0.011     -1.95213   -.2504616
       pctel |  -.6497768   .0310318   -20.94   0.000     -.710775   -.5887786
       _cons |   686.0322   8.728224    78.60   0.000     668.8754     703.189
------------------------------------------------------------------------------
```

MEASURES OF FIT IN MULTIPLE REGRESSION

```
reg testscr str pctel, robust

Regression with robust standard errors                 Number of obs =     420
                                                       F(  2,   417) =  223.82
                                                       Prob > F      =  0.0000
                                                       R-squared     =
                                                       Root MSE      =

------------------------------------------------------------------------------
             |               Robust
     testscr |      Coef.   Std. Err.      t    P>|t|      [95% C
-------------+----------------------------------------------------------------
         str |  -1.101296   .4328472    -2.54   0.011     -1.952
       pctel |  -.6497768   .0310318   -20.94   0.000     -.7107
       _cons |   686.0322   8.728224    78.60   0.000     668.87
------------------------------------------------------------------------------
```

```
. est tab, stats(r2 r2_a)
```

Variable	Active
str	-1.1012959
el_pct	-.64977678
_cons	686.03225
r2	.42643136
r2_a	.42368043

5.5 MULTIPLE REGRESSION AND CAUSALITY

University *of* Massachusetts Amherst

ASSUMPTIONS FOR CAUSAL INFERENCE IN MULTIPLE REGRESSION

1. The regressors X_s are independent of the error term u_i

$$E(u_i|X_{1i}, X_{2i}, \dots X_{ki}) = 0$$

2. $(Y_i, X_{1i}, X_{2i}, \dots X_{ki})$, i = 1,…, n, are i.i.d.

3. Large outliers are rare.

4. No perfect multicollinearity.

ASSUMPTIONS FOR CAUSAL INFERENCE IN MULTIPLE REGRESSION

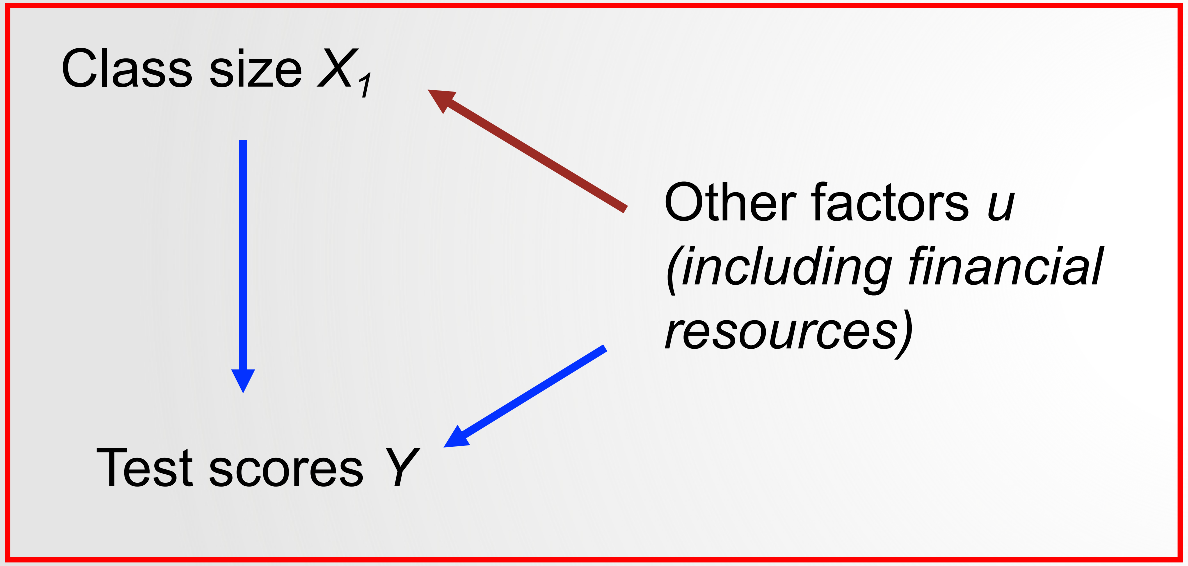
1. The regressors X_s are independent of the error term u_i

$$E(u_i|X_{1i}, X_{2i}, \ldots X_{ki}) = 0$$

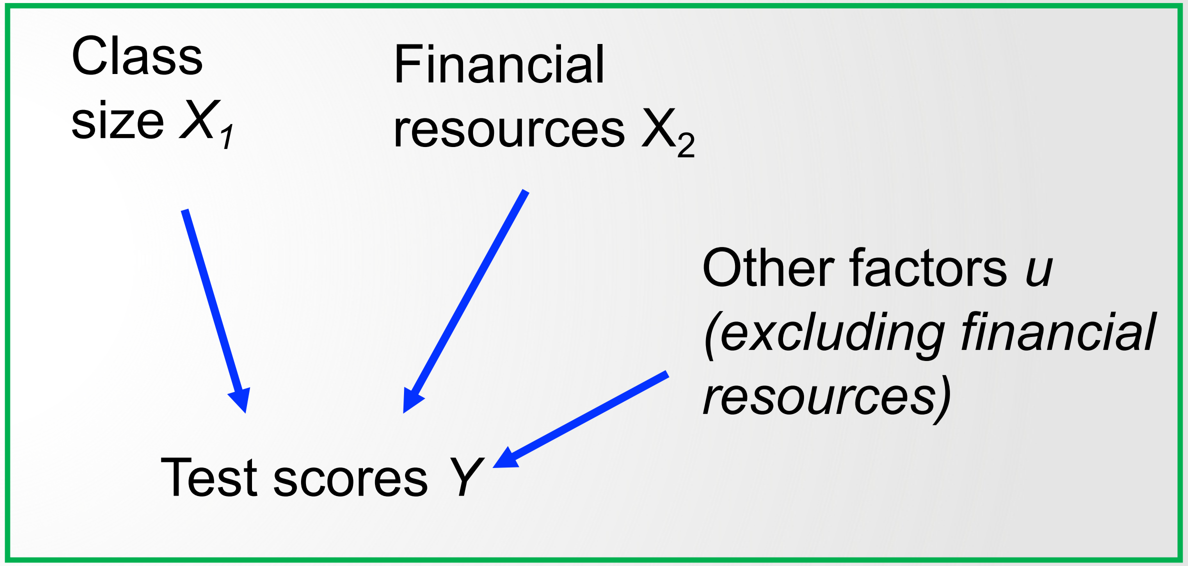
2. $(Y_i, X_{1i}, X_{2i}, \ldots X_{ki})$, i = 1,…, n, are i.i.d.
3. Large outliers are rare.
4. No perfect multicollinearity.

HYPOTHETICAL EXAMPLE

$$Y_i = \beta_0 + \beta_1 X_{1i} + u_i$$



$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$



- Hypothetical example: Class size X_1 uncorrelated with the error term *only after controlling for financial resources X_2.*

THE CIA

- X = regressor (or "treatment") of interest.
- $W_1, W_2, \ldots, W_k$ = control variables.
- Conditional Independence Assumption (CIA):

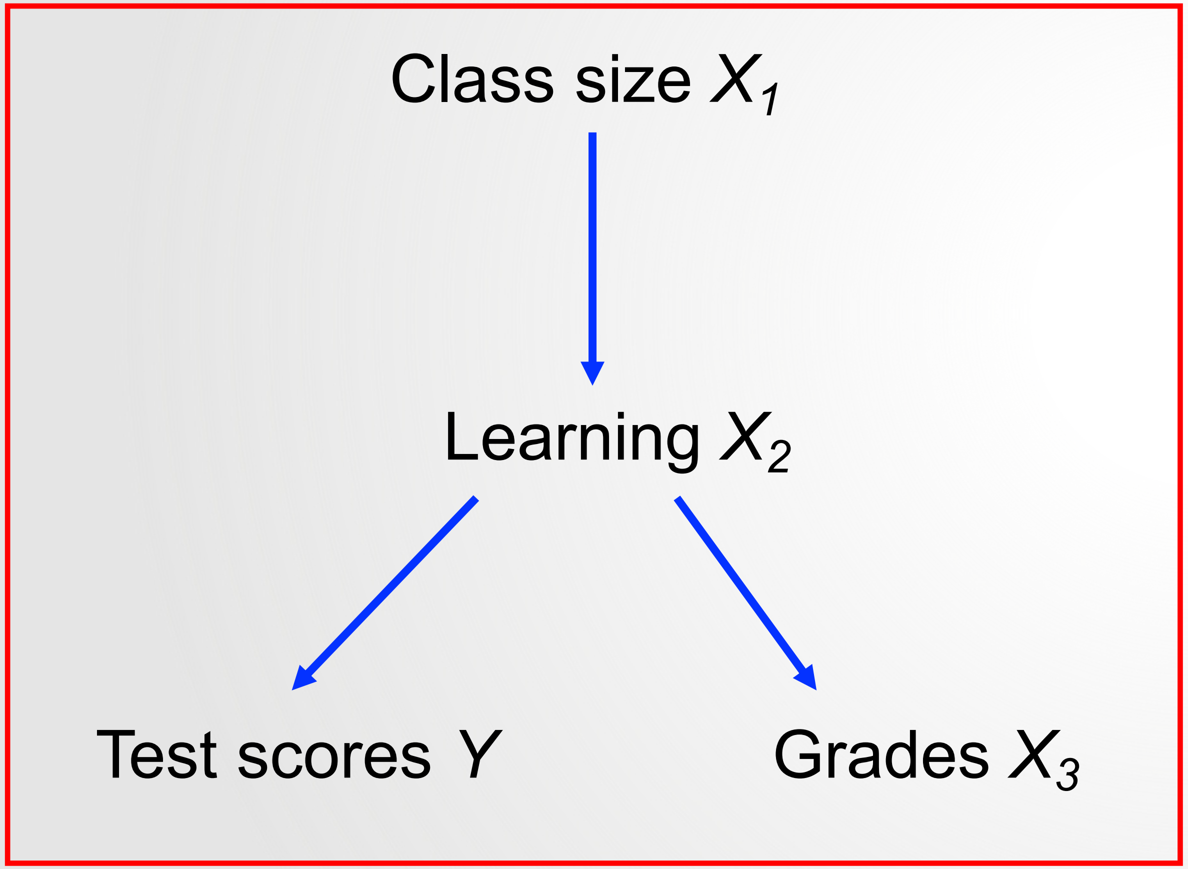
$$E(u_i|X, W_1, \ldots, W_k) = E(u_i|W_1, \ldots, W_k)$$

In words: u and *X* are uncorrelated, after controlling for the W_s

CONTROL VARIABLES: GOOD AND BAD

- Not all variables are suitable as control variables.
- *Bad controls*: variables that are affected by the X of interest.
  - By "holding them fixed", you *create* bias.
- *Good controls* are pre-determined with respect to the X of interest.
- In estimating the effect of class size on test scores, the amount of *learning* by students (if observable) would be a *bad control*.

EXAMPLE OF BAD CONTROL VARIABLES



- We are after the effect of class size on test scores.
- Don't control for *learning!* we don't want to hold learning fixed
- Similarly, don't control for grades! Doesn't make sense to hold them fixed, when class size affects them through learning.
- "Learning" and grades are *bad controls*.
- **Don't control for anything that is affected by the regressor of interest!**

ASSUMPTIONS FOR CAUSAL INFERENCE IN MULTIPLE REGRESSION

1. The regressors X_s are independent of the error term u_i

$$E(u_i|X_{1i}, X_{2i}, \ldots X_{ki}) = 0$$

2. $(Y_i, X_{1i}, X_{2i}, \ldots X_{ki})$, *i* = 1,…, *n*, are i.i.d.

3. Large outliers are rare.

4. No perfect multicollinearity.

5.6 MULTICOLLINEARITY

PERFECT MULTICOLLINEARITY: EXAMPLE

$$TestScores_i = \beta_0 + \beta_1 STR_i + \beta_2 PctEL_i + \beta_3 FracEL_i + u_i$$

- $PctEL$ = percentage of English learners (from 0 to 100).
- $FracEL$ = fraction of English learners (from 0 to 1).
- *Perfect multicollinearity*: $PctEL = 100xFracEL$
- β_2 = effect of increasing $PctEL$ by 1 while keeping $FracEL$ fixed. Nonsense!!
- STATA will drop one of the two multicollinear regressors.

THE DUMMY VARIABLE TRAP

- 2 indicator variables for sex at birth
  - *Female* =1 if woman; 0 if man.
  - *Male* =1 if man; 0 if woman
- $Y_i = \beta_0 + \beta_1 Female + \beta_2 Male + u_i$ cannot be estimated
  - Perfect multicollinearity: $Female_i + Male_i = 1 = X_{oi}$
- Can estimate one of these three:

1. $Y_i = \beta_0 + \beta_1 Female + u_i$
2. $Y_i = \beta_0 + \beta_1 Male + u_i$
3. $Y_i = \beta_1 Female + \beta_2 Male + u_i$

THE DUMMY VARIABLE TRAP

- *General rule*:
  If you have G indicator variables, and each observation falls into one (and only one) category, *you cannot estimate all G indicators plus an intercept.*
- Conventional solution: include G-1 indicators + the intercept
- Then coefficient on one included indicator = difference between that category and the “excluded category”.
- Can also exclude the intercept and include all G indicators.

IMPERFECT MULTICOLLINEARITY

- Example:

$$AHE_i = \beta_0 + \beta_1 Age + \beta_2 Experience + u_i$$

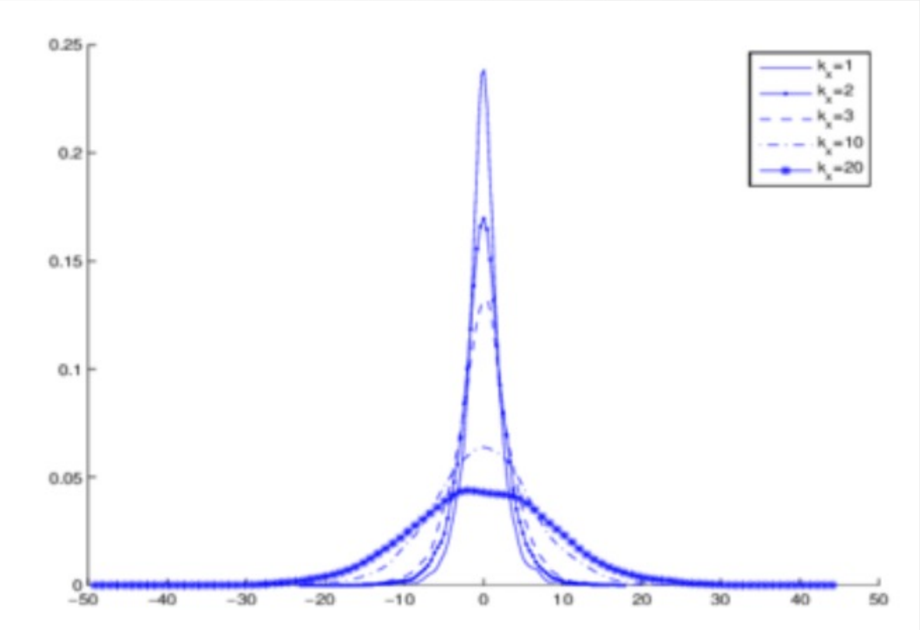
  - AHE = average hourly earnings.
  - Experience=years since entering the labor force.
- Nothing wrong with this regression.
- But β_1 & β_2 will probably be imprecisely estimated (large SE).
- There is probably little variation in experience within each given age group.

5.7 STATISTICAL INFERENCE ABOUT A SINGLE COEFFICIENT

University *of* Massachusetts Amherst

DISTRIBUTION OF OLS ESTIMATORS

- $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$ are random variables.
- $E(\hat{\beta}_j) = \beta_j$ for $j = 1, \ldots, k$.
- $Var(\hat{\beta}_j)$ is inversely proportional to *n*.
- $\hat{\beta}_j \to \beta_j$ (law of large numbers)
- Each $\hat{\beta}_j$ is normally distributed in large samples (CLT).



HYPOTHESIS TESTS & CI_S FOR SINGLE COEFFICIENTS

1. Specify H_0 & H_1.
2. Estimate $\hat{\beta}_j$ and $SE(\hat{\beta}_j)$.
3. Compute t-statistics: $t = \frac{\hat{\beta}_j - \beta_{j,0}}{SE(\hat{\beta}_j)}$
4. Compute p-value: $p = 2\Phi(-|t|)$.
5. Compute 95% CI: $\{\hat{\beta}_j \pm 1.96 \times SE(\hat{\beta}_j)\}$.

APPLICATION: STR & TEST SCORES

$$\widehat{TestScore} = \underset{(8.7)}{686.0} - \underset{(0.43)}{1.10} \times STR - \underset{(0.031)}{0.650} \times PctEL.$$

1. Null hypothesis: $H_0: \beta_1 = 0$
2. t-statistic: $t = \frac{-1.10-0}{0.43} = -2.54$
3. p-value: $2\Phi(-2.54) = 0.011 = 1.1\%.$
4. 95% confidence interval for β_1:

$$-1.10 \pm 1.96 \times 0.43 = (-1.95, -0.26)$$

APPLICATION: STR & TEST SCORES

$$\widehat{TestScore} = \underset{(8.7)}{686.0} - \underset{(0.43)}{1.10} \times STR - \underset{(0.031)}{0.650} \times PctEL.$$

- **YOUR TURN**: Test $H_0: \beta_2 = 0$ and compute 95% c.i. for β_2.

APPLICATION: STR & TEST SCORES

$$\widehat{TestScore} = \underset{(8.7)}{686.0} - \underset{(0.43)}{1.10} \times STR - \underset{(0.031)}{0.650} \times PctEL.$$

- **YOUR TURN**: Test $H_0: \beta_2 = 0$ and compute 95% c.i. for β_2
- t-statistic: $t = \frac{-0.650-0}{0.031} = -20.9$
- p-value: $2\Phi(-20.9) = 5.3\times10^{-97}$
- 95% confidence interval for β_1:

$$-0.65 \pm 1.96\times0.031 = (-0.71, -0.59)$$

IN STATA

```
reg testscr str pctel, robust

Regression with robust standard errors                 Number of obs =     420
                                                       F(  2,   417) =  223.82
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.4264
                                                       Root MSE      =  14.464

------------------------------------------------------------------------------
             |               Robust
     testscr |      Coef.   Std. Err.       t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         str |  -1.101296   .4328472    -2.54   0.011    -1.95213   -.2504616
       pctel |  -.6497768   .0310318   -20.94   0.000    -.710775   -.5887786
       _cons |   686.0322   8.728224    78.60   0.000    668.8754     703.189
------------------------------------------------------------------------------
```

5.8 JOINT HYPOTHESES:

STATISTICAL INFERENCE ABOUT MULTIPLE COEFFICIENTS AT THE SAME TIME

University *of* Massachusetts Amherst

TESTS OF JOINT HYPOTHESES

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \cdots + \beta_k X_{k,i} + u_i$$

- Example of joint hypotheses:

$$H_0\colon\ \beta_1 = 0 \text{ and } \beta_2 = 0$$

$$H_1\colon\ \beta_1 \neq 0 \text{ and/or } \beta_2 \neq 0$$

- Can also test more than two *restrictions*.
- In general:

H_0: $\beta_j = \beta_{j,0}$, $\beta_m = \beta_{m,0}$, ... up to q restrictions

H_1: one or more of the q restrictions doesn't hold

THE F-STATISTIC

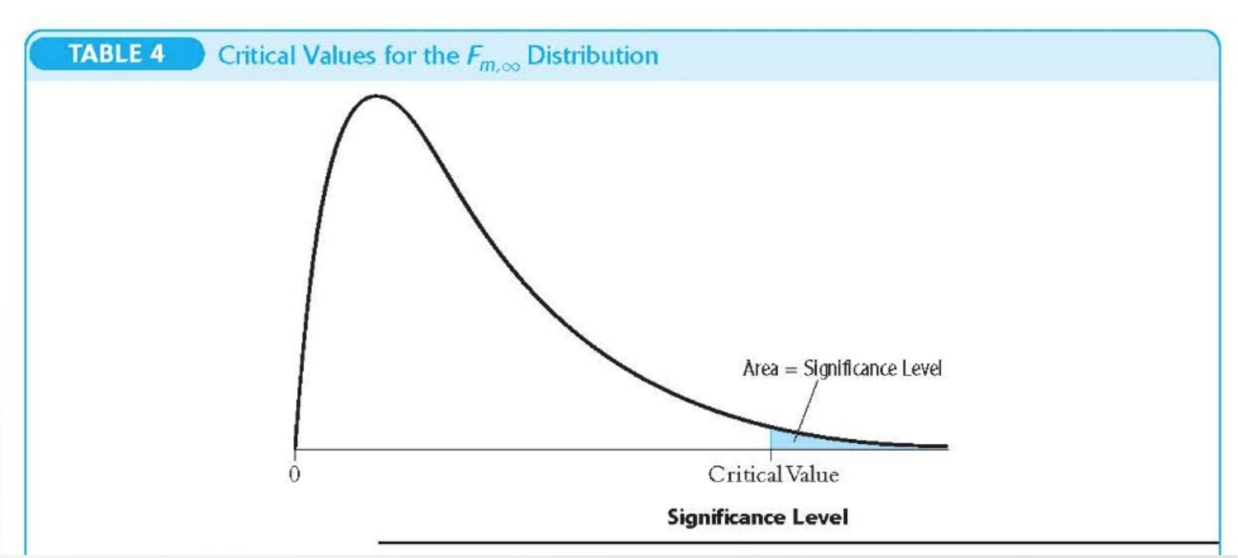
- Tests all components of the joint hypothesis at once.
- With q=2 restrictions (H_0: $\beta_1 = \beta_{1,0}$ ***and*** $\beta_2 = \beta_{2,0}$):

$$F = \frac{1}{2}\left(\frac{t_1^2 + t_2^2 - 2\hat{\rho}_{t_1,t_2}}{1 - \hat{\rho}_{t_1,t_2}^2}\right)$$

  - t_1 = individual t-stat for $\beta_1 = \beta_{1,0}$
  - t_2 = individual t-stat for $\beta_2 = \beta_{2,0}$
  - $\hat{\rho}_{t_1,t_2}$=correlation between t_1 & t_2

THE F-STATISTIC

- In large samples, the F-stat is distributed $F_{q,\infty}$.
- p-value=$\Pr[F_{q,\infty} > F^{act}]$
- 'test' command in STATA
  - it's a *post-estimation* command



F-STATISTICS: APPLICATION

$$TestScore_i = \beta_0 + \beta_1 STR_i + \beta_2 Expn_i + \beta_3 PctEL_i + u_i$$

$H_0: \beta_1 = 0$ and $\beta_2 = 0$

vs.

$H_1: \beta_1 \neq 0$ and/or $\beta_2 \neq 0$.

```
reg testscr str expn_stu pctel, robust
Regression with robust standard errors                 Number of obs =     420
                                                       F(  3,   416) =  147.20
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.4366
                                                       Root MSE      =  14.353
------------------------------------------------------------------------------
             |               Robust
     testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         str |  -.2863992   .4820728    -0.59   0.553    -1.234001     .661203
    expn_stu |   .0038679   .0015807     2.45   0.015     .0007607    .0069751
       pctel |  -.6560227   .0317844   -20.64   0.000    -.7185008   -.5935446
       _cons |   649.5779   15.45834    42.02   0.000     619.1917    679.9641
------------------------------------------------------------------------------

test str expn_stu

 ( 1)  str = 0.0
 ( 2)  expn_stu = 0.0
       F(  2,   416) =    5.43
       Prob > F =    0.0047
```

THE "OVERALL" REGRESSION F-STAT

- $H_0: \beta_1 = 0, \beta_2 = 0, \ldots, \beta_k = 0$
- H_1: $\beta_j \neq 0$ for at least one j
- → *does any of the included regressors help explain Y?*
- → *Does the model do better than simply computing the sample mean?*
- Part of STATA 'regress' output

```
reg testscr str expn_stu pctel, r;
Regression with robust standard errors               Number of obs =     420
                                                     F(  3,   416) =  147.20
                                                     Prob > F      =  0.0000
                                                     R-squared     =  0.4366
                                                     Root MSE      =  14.353
------------------------------------------------------------------------------
             |               Robust
     testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         str |  -.2863992   .4820728    -0.59   0.553    -1.234001    .661203
    expn_stu |   .0038679   .0015807     2.45   0.015     .0007607   .0069751
      el_pct |  -.6560227   .0317844   -20.64   0.000    -.7185008  -.5935446
       _cons |   649.5779   15.45834    42.02   0.000     619.1917   679.9641
------------------------------------------------------------------------------
```

TESTING SINGLE RESTRICTIONS ON MULTIPLE COEFFICIENTS

- Example:

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + u_i$$

- Hypothesis:

$$H_0: \beta_1 = \beta_2 \text{ vs } H_1: \beta_1 \neq \beta_2$$

- One single hypothesis…
- …but about multiple coefficients.

TESTING SINGLE RESTRICTIONS ON MULTIPLE COEFFICIENTS

- Example:

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + u_i$$

- Hypothesis:

$H_0: \beta_1 = \beta_2$ vs $H_1: \beta_1 \neq \beta_2$

- Two ways to test this:

1. Rearrange ("transform") the regression
2. Perform the test directly

METHOD 1: REARRANGE THE REGRESSION

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + u_i$$

- We want to test $H_0: \beta_1 = \beta_2$ vs $H_1: \beta_1 \neq \beta_2$
- Add and subtract $\beta_2 X_{1,i}$:

$$Y_i = \beta_0 + \beta_1 X_{1,i} - \beta_2 X_{1,i} + \beta_2 X_{2,i} + \beta_2 X_{1,i} + u_i$$

$$Y_i = \beta_0 + X_{1,i}(\beta_1 - \beta_2) + \beta_2 (X_{1,i} + X_{2,i}) + u_i$$

$$Y_i = \beta_0 + \gamma_1 X_{1,i} + \beta_2 (X_{1,i} + X_{2,i}) + u_i$$

With $\gamma_1 = \beta_1 - \beta_2$

METHOD 1: REARRANGE THE REGRESSION

$$(a)\ Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + u_i$$

$$H_0: \beta_1 = \beta_2 \text{ vs } H_1: \beta_1 \neq \beta_2$$

$$(b)\ Y_i = \beta_0 + \gamma_1 X_{1,i} + \beta_2 (X_{1,i} + X_{2,i}) + u_i$$

- (b) and (b) are equivalent
    - same R^2, predicted values, and residuals.
- But now the test boils down to whether γ_1=0 in regression (b)!

METHOD 2: PERFORM THE TEST DIRECTLY USING SOFTWARE

- Regression: $Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + u_i$
- Hypothesis: $H_0: \beta_1 = \beta_2$ vs $H_1: \beta_1 \neq \beta_2$
- Example:

$$TestScore_i = \beta_0 + \beta_1 STR_i + \beta_2 Expn_i + \beta_3 PctEL_i + u_i$$

- 'test' command in STATA after running the regression:

1. *regress testscr str expn_stu el_pct, r*
2. *test str=expn*

```
. regress testscr str expn_stu el_pct, r

Linear regression                               Number of obs     =        420
                                                F(3, 416)         =     147.20
                                                Prob > F          =     0.0000
                                                R-squared         =     0.4366
                                                Root MSE          =     14.353
```

testscr	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
str	-.2863992	.4820728	-0.59	0.553	-1.234002	.661203
expn_stu	.0038679	.0015807	2.45	0.015	.0007607	.0069751
el_pct	-.6560227	.0317844	-20.64	0.000	-.7185008	-.5935446
_cons	649.5779	15.45834	42.02	0.000	619.1917	679.9641

```
. test str=expn

 ( 1)  str - expn_stu = 0

       F(  1,   416) =    0.36
            Prob > F =    0.5467
```

$\gamma_1 = \beta_1 - \beta_2$

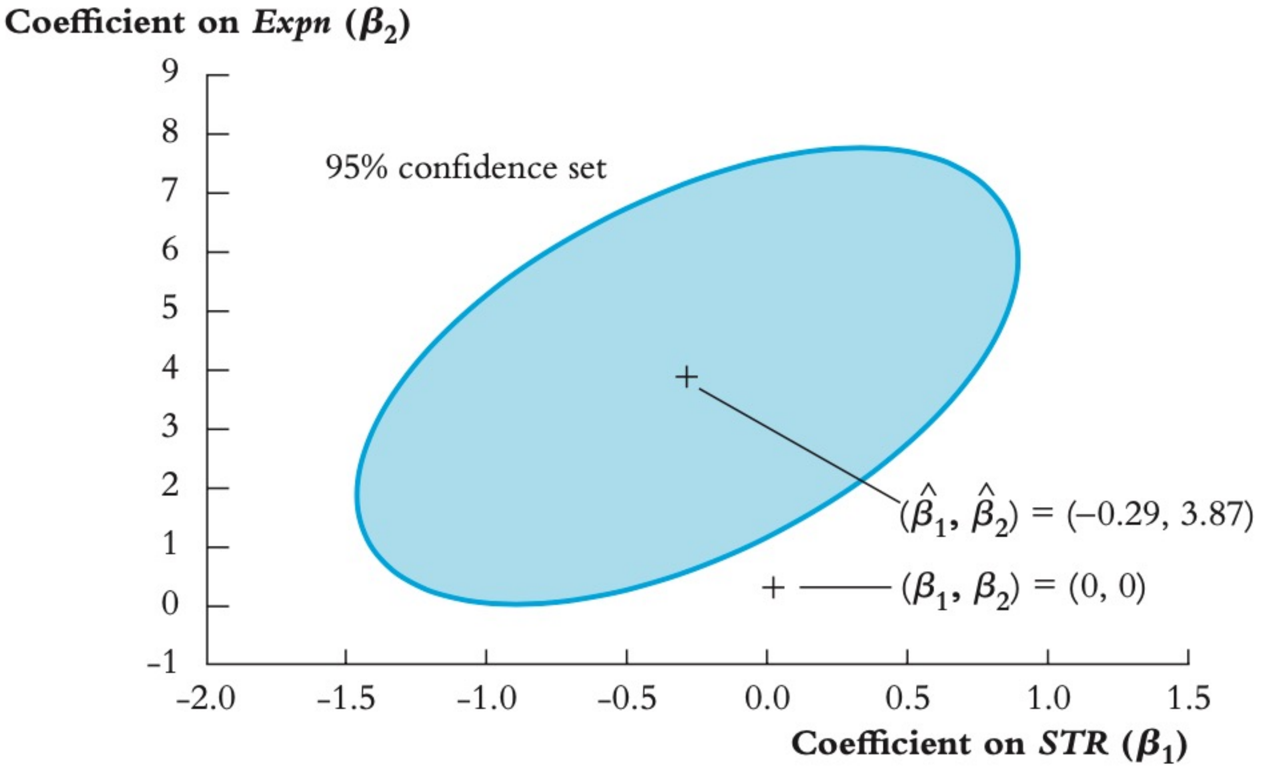
CONFIDENCE SETS FOR MULTIPLE COEFFICIENTS

- Regression: $Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + u_i$
- A *joint* confidence set for β_1 and β_2:
  - A set of pairs of values (β_1, β_2) such that it is 95% likely to contain the true pair.
  - The set of pairs of values (β_1, β_2) that cannot be rejected at the 5% significance level (using F-stat)

CONFIDENCE SETS FOR MULTIPLE COEFFICIENTS

FIGURE 7.1 95% Confidence Set for Coefficients on *STR* and *Expn* from Equation (7.6)

The 95% confidence set for the coefficients on *STR* (β_1) and *Expn* (β_2) is an ellipse. The ellipse contains the pairs of values of β_1 and β_2 that cannot be rejected using the *F*-statistic at the 5% significance level. The point $(\beta_1, \beta_2) = (0, 0)$ is not contained in the confidence set, so the null hypothesis H_0: $\beta_1 = 0$ and $\beta_2 = 0$ is rejected at the 5% significance level.



5.9 MODEL SPECIFICATION & PRESENTATION

University *of* Massachusetts Amherst

HOW TO CHOOSE REGRESSORS

- You want to estimate the effect of X_1 on Y.
- Include control variables W_i that are correlated with X_1 and affect Y.
  - Objective: $E(u_i|\, X_1, W_i) = E(u_i|W_i)$
  - X should be *as if randomly assigned* among units w/ same value of W_i.
  - Some things are hard to measure, so we use *proxies*.
- Can also include variables that affect Y but are not expected to correlate with X, to increase precision.
- Baseline specification & alternative specifications (*robustness*).

PRESENTING REGRESSION RESULTS

- We usually run several regressions (baseline + alternative specification).
- Use tables to present results from multiple specifications.
- Each specification is a column.
- Table should include, for each specification:
    1. Estimated Coefficients.
    2. Standard Errors.
    3. Number of observations.
    4. Measures of fit.
    5. Relevant F-stats, if any.

TABLE 7.1 Results of Regressions of Test Scores on the Student–Teacher Ratio and Student Characteristic Control Variables Using California Elementary School Districts

Dependent variable: average test score in the district.

Regressor	(1)	(2)	(3)	(4)	(5)
Student–teacher ratio (X_1)	−2.28 (0.52) [−3.30, −1.26]	−1.10 (0.43) [−1.95, −0.25]	−1.00 (0.27) [−1.53, −0.47]	−1.31 (0.34) [−1.97, −0.64]	−1.01 (0.27) [−1.54, −0.49]
Control variables					
Percentage English learners (X_2)		−0.650 (0.031)	−0.122 (0.033)	−0.488 (0.030)	−0.130 (0.036)
Percentage eligible for subsidized lunch (X_3)			−0.547 (0.024)		−0.529 (0.038)
Percentage qualifying for income assistance (X_4)				−0.790 (0.068)	0.048 (0.059)
Intercept	698.9 (10.4)	686.0 (8.7)	700.2 (5.6)	698.0 (6.9)	700.4 (5.5)
Summary Statistics					
SER	18.58	14.46	9.08	11.65	9.08
$\bar{R}^2$	0.049	0.424	0.773	0.626	0.773
n	420	420	420	420	420

These regressions were estimated using the data on K–8 school districts in California, described in Appendix 4.1. Heteroskedasticity-robust standard errors are given in parentheses under coefficients. For the variable of interest, the student–teacher ratio, the 95% confidence interval is given in brackets below the standard error.

- SE in parenthesis.
- Start from simplest specification (column 1)
- (3) & (4) use two alternative proxies for financial resources.
- Column (5) uses both.
- Coefficient on STR falls from (1) to (2) but is relatively stable across (2)-(5).