5 – LINEAR REGRESSION II
MULTIPLE REGRESSORS

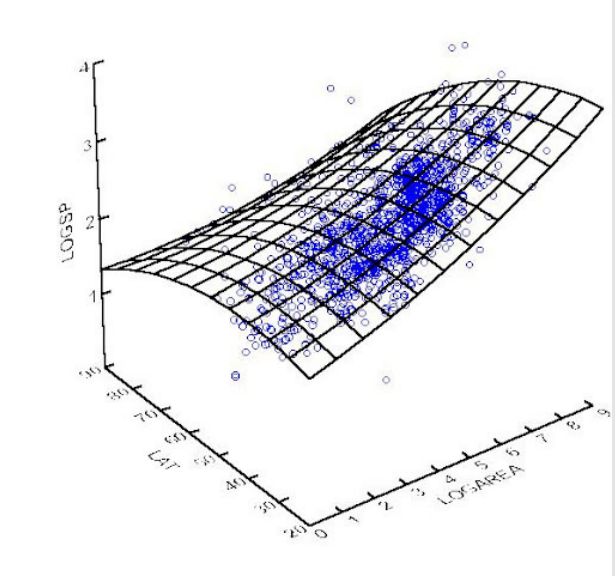
University of Massachusetts Amherst
BE REVOLUTIONARY™

SECTION 5 – LINEAR REGRESSION, PART 2

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

MULTIPLE LINEAR REGRESSION: OVERVIEW



Why multiple regressors?

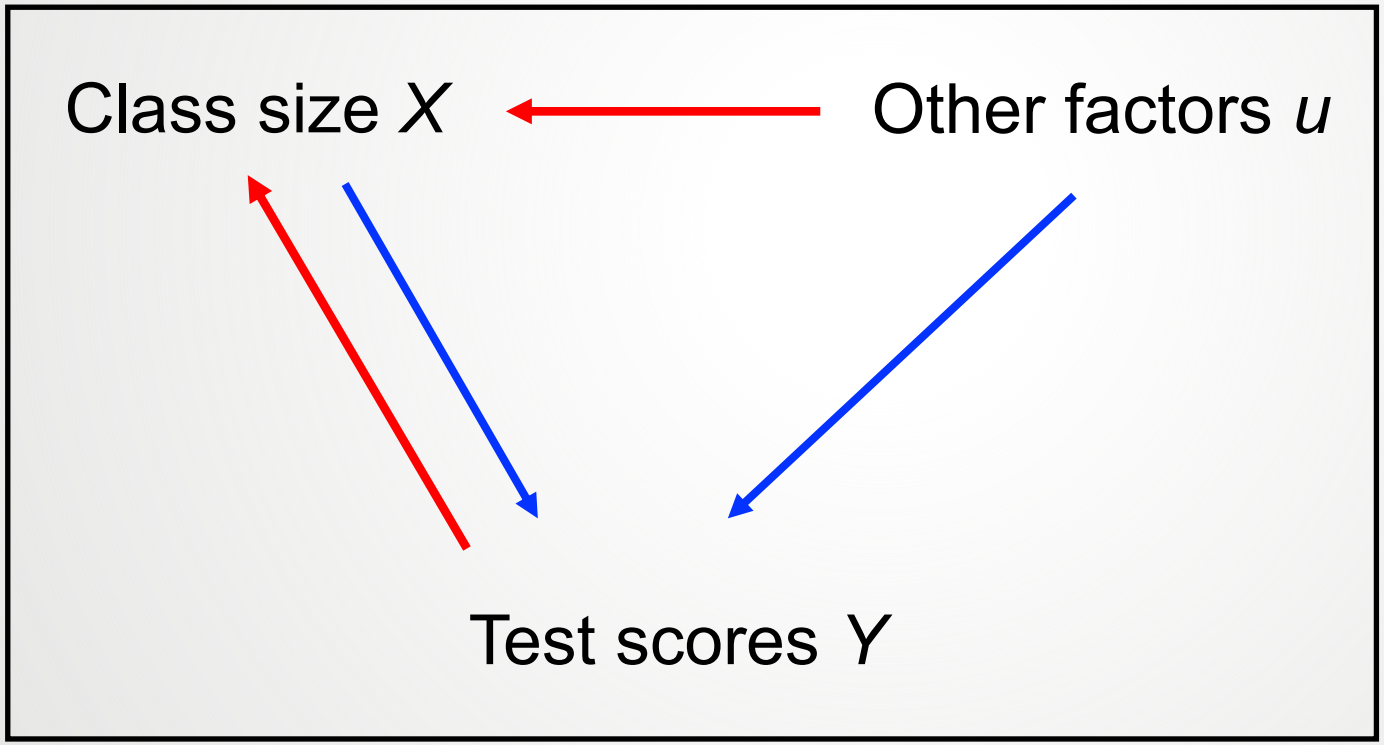
- Prediction → increase accuracy.
- Causal inference → *control for* confounding factors.

Parameter	MODEL1	MODEL2	MODEL3	MODEL4	MODEL5
FEMALE	4.87	5.49	5.44	5.94	5.49
	(1.30)	(1.01)	(0.93)	(0.91)	(0.88)
Intercept	50.12	20.23	11.90	8.58	6.14
	(0.96)	(2.71)	(2.86)	(2.87)	(2.81)
MATH			0.40	0.29	0.24
			(0.07)	(0.07)	(0.07)
READ		0.57	0.33	0.23	0.13
		(0.05)	(0.06)	(0.06)	(0.06)
SCIENCE				0.26	0.24
				(0.06)	(0.06)
SOCST					0.23
					(0.05)

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.

   AND

2. The omitted variable is a determinant of the dependent variable.

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 β_1 be the true causal effect of *X* on *Y* in the population.
- Let $\rho_{Xu} = corr(X_i, u_i)$
- OLS coefficient gives you:

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

(proof in Appendixes 4.3 & 6.1)

OMITTED VARIABLES BIAS (OVB)

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

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

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

OMITTED VARIABLES BIAS (OVB)

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

	$Corr(Z, X) > 0$	$Corr(Z, X) < 0$
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

	$Corr(Z, X) > 0$	$Corr(Z, X) < 0$
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

	$Corr(Z, X) > 0$	$Corr(Z, X) < 0$
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

	$Corr(Z, X) > 0$	$Corr(Z, X) < 0$
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) = a way to address OVB (& also reverse causality).
- Imagine randomly assigning class size X to schools.
- Same $E(X)$ for all units, independent of other factors affecting Y.
- → $E(u)$ does not vary with X.
- → Randomization ensures $corr(X, u) = 0$.

"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

"CONTROLLING FOR" OMITTED VARIABLES

- 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?

"CONTROLLING FOR" OMITTED VARIABLES

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

- $\beta_1 = \frac{\Delta Y}{\Delta X_i}, holding\ X_2\ constant.$
- *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:

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

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

5.3 OLS ESTIMATION OF THE MULTIPLE REGRESSION MODEL

University of Massachusetts Amherst

OLS ESTIMATION OF MULTIPLE REGRESSION

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

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

- OLS estimators $(\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k)$ = the values of $b_0, b_1, \ldots, b_k$ 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 $(X_2, \ldots, X_k)$.
- $\tilde{Y} =$ residual from regression of Y on all other regressors $(X_2, \ldots, X_k)$.

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 $\tilde{X}_1$.

EXAMPLE: CLASS SIZE & TEST SCORES

- Back to our dataset of 420 California school districts
- 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

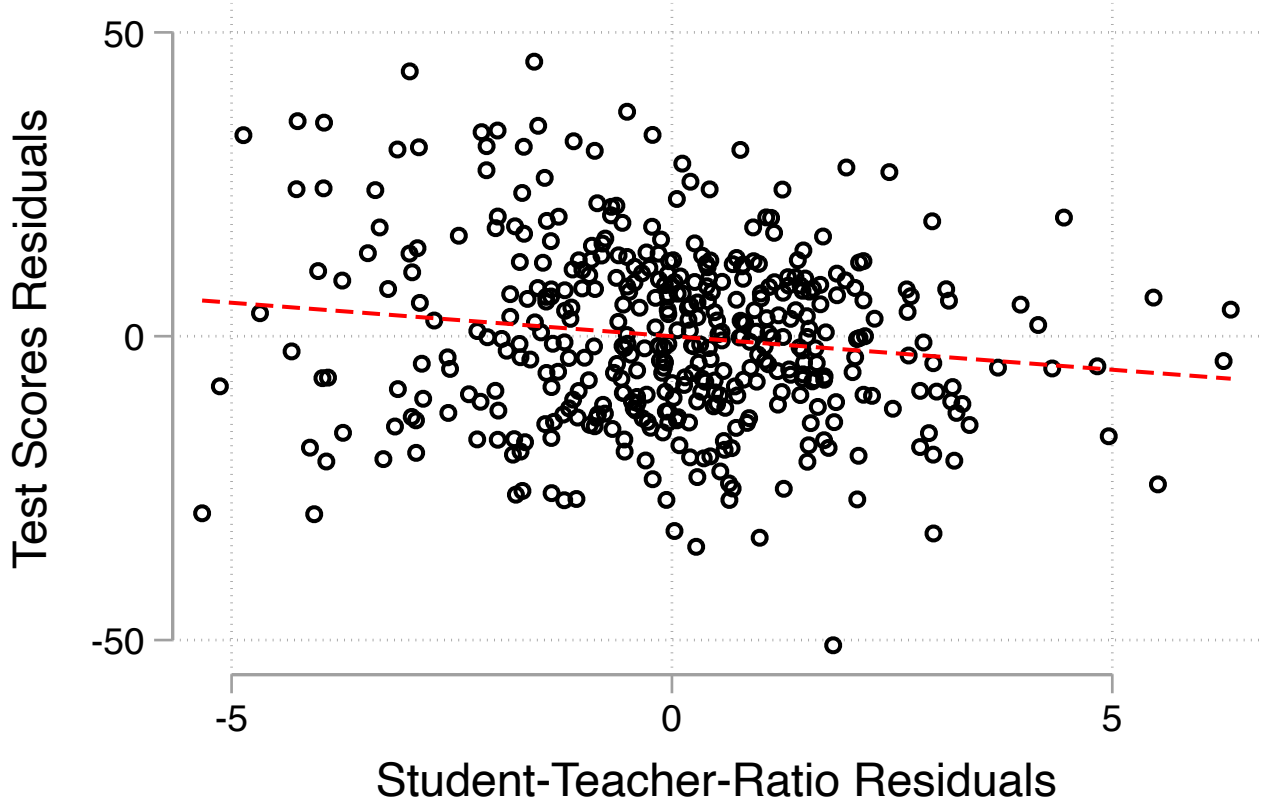
```
reg testscr str pctel, robust

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.

R^2 & ADJUSTED R^2

- $R^2 = \frac{ESS}{TSS} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$
- 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     668.8
------------------------------------------------------------------------------
```

```
. 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

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

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.

ASSUMPTIONS FOR CAUSAL INFERENCE IN MULTIPLE REGRESSION

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

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

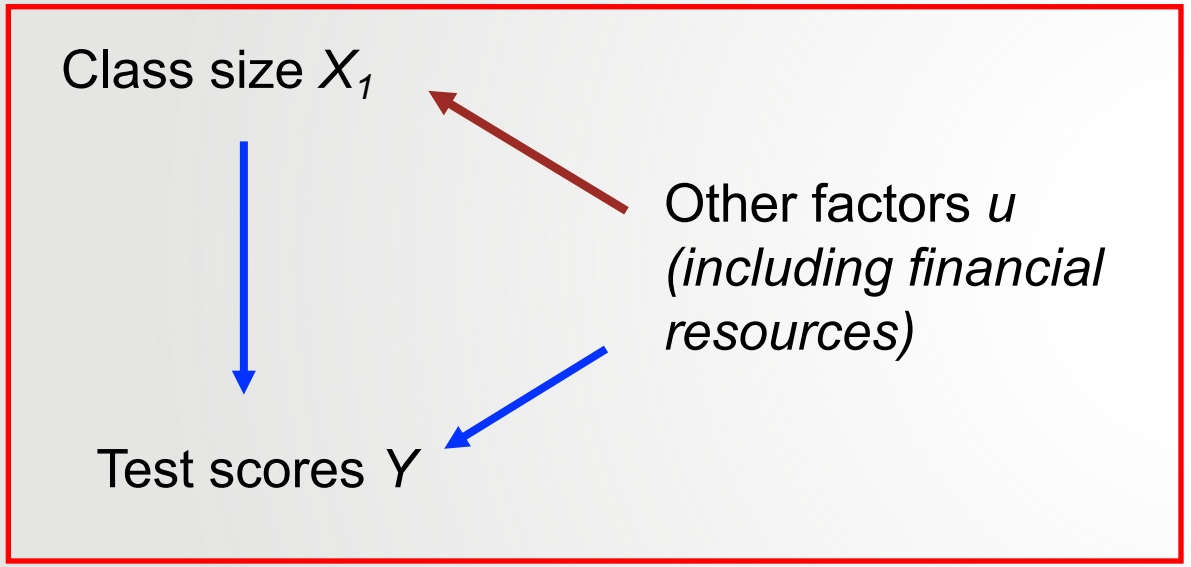
2. $(Y_i, X_{1i}, X_{2i}, \ldots X_{ki})$, i = 1,…, n, are i.i.d.

3. Large outliers are rare.

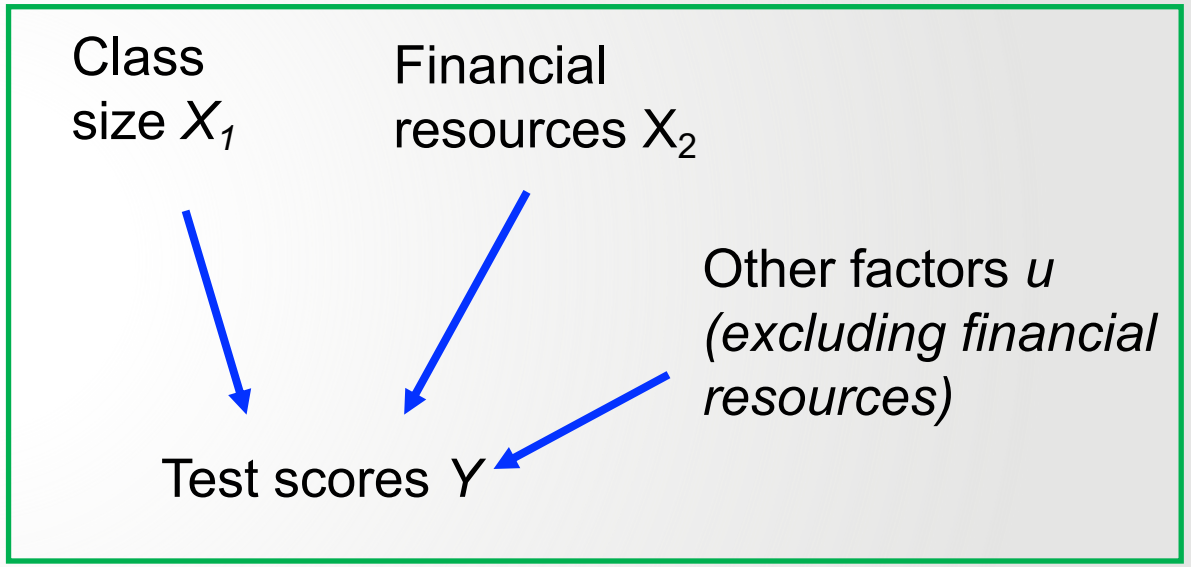
4. No perfect multicollinearity.

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):

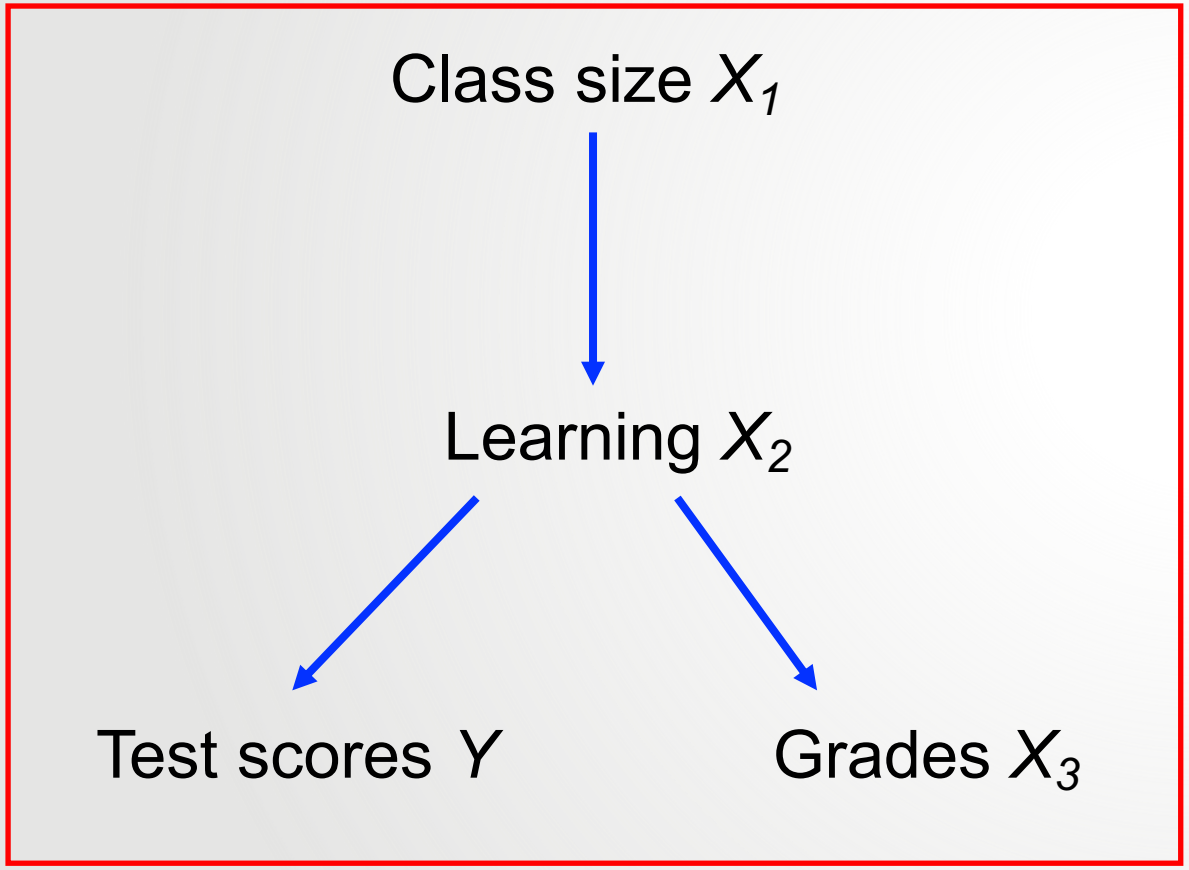
$$E(u_i|X, W_1, \ldots, W_k) = E(u_i|W_1, \ldots, W_k)$$

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

5.5 MULTICOLLINEARITY

University of Massachusetts Amherst

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*: $PctEL = 100 x FracEL$
- 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.