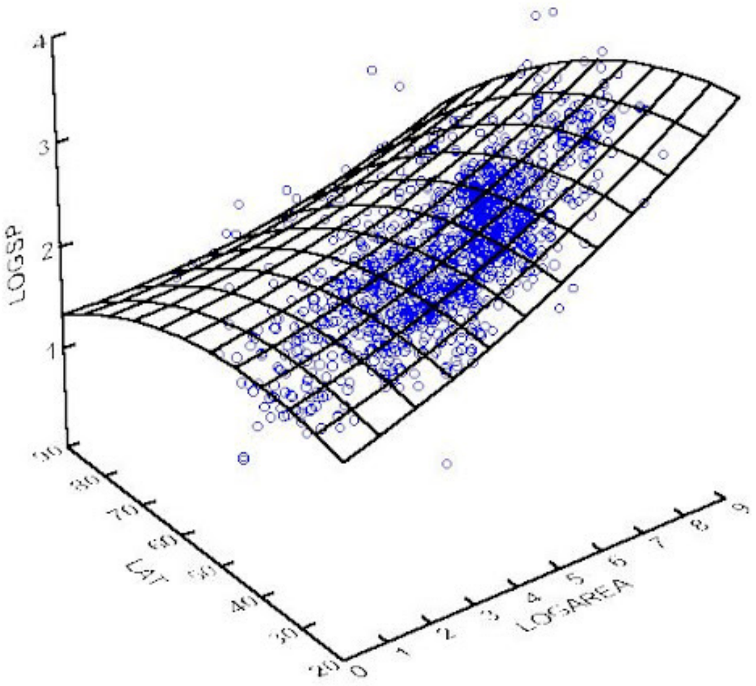
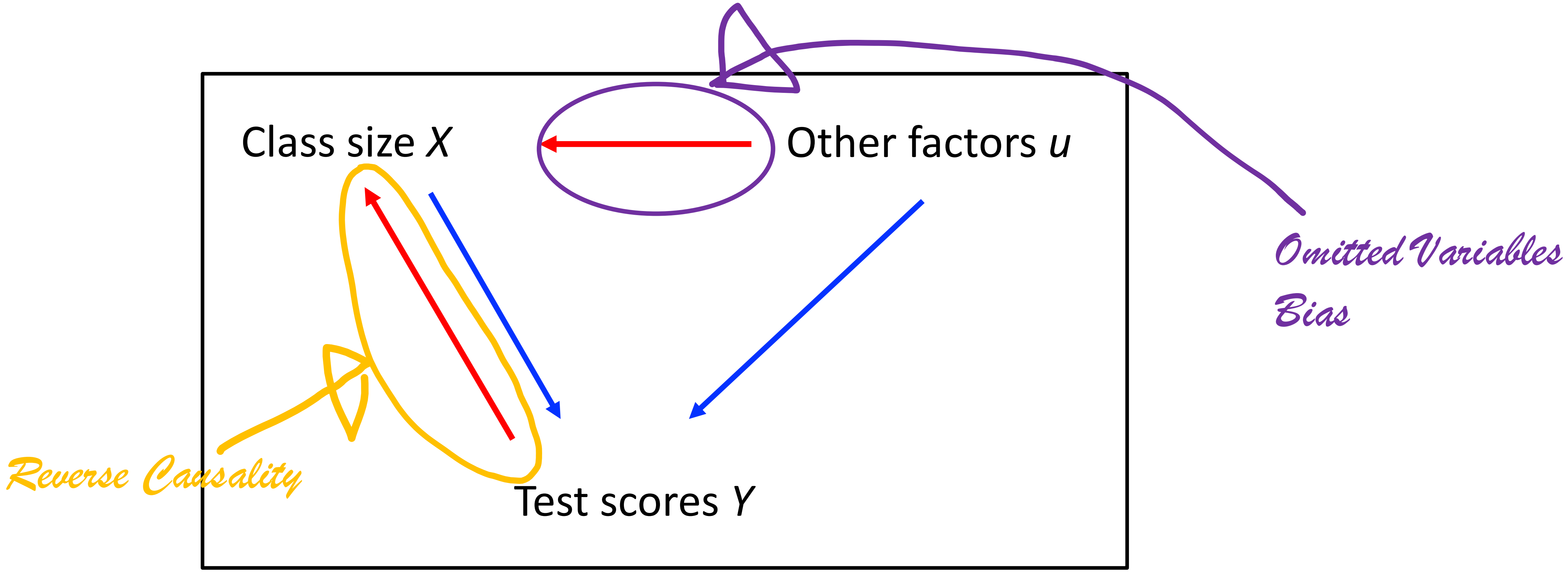
3. Linear regression with multiple regressors



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.

OMITTED VARIABLES BIAS (OVB)

- Linear regression model:

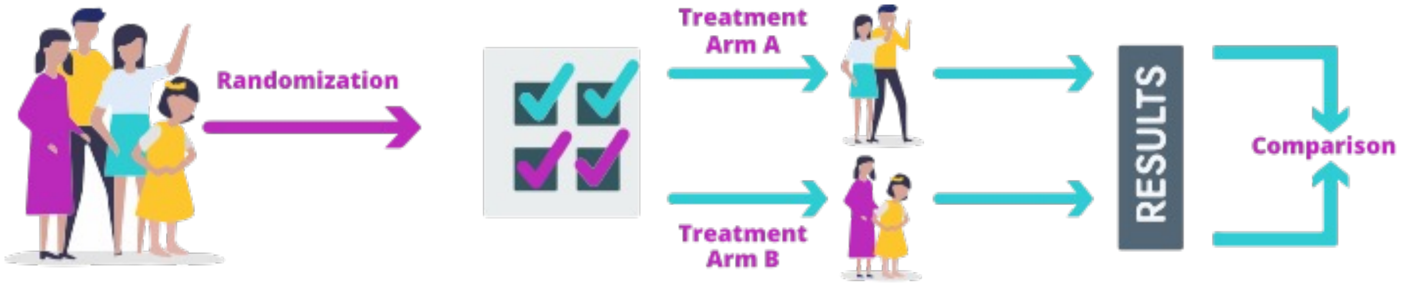
$$TestScores_i = \beta_0 + \beta_1 STR + u_i$$

- 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

RANDOMIZATION AS A SOLUTION

- Randomized Controlled Trials (RCTs).
- Random assignment of X → no OVB (& no reverse causality).
- X is purely random, so it's 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 variable Z that affects both Y and X, we can try to "control for" it.
- Include Z in the regression, so it's no longer part of u.
- Estimate the relation between X and Y, while keeping Z fixed.

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?

Multiple regression model with k regressors

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

$$\Downarrow$$

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

OLS estimation of multiple regression

- Select $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$ to *best fit* the sample 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 estimation 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)

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
------------------------------------------------------------------------------
```

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

Hypothesis tests & CIs for single coefficients in multiple regression

1. Specify H_0 & H_1.
2. Estimate $\hat{\beta}_j$ and $SE(\hat{\beta}_j)$ from the multiple regression.
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: $\{\hat{\beta}_j \pm 1.96 \times SE(\hat{\beta}_j)\}$.

Takeaway:
It's exactly the same as before, except that $\hat{\beta}_j$ and $SE(\hat{\beta}_j)$ are now estimated from a multiple regression.

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:

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

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
------------------------------------------------------------------------------
```

R^2 & adjusted R^2 in multiple regression

- $R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS}$

- Always increases if you add regressors.

- $Adjusted\ R^2 (or\ \bar{R}^2) = 1 - \frac{n-1}{n-k-1} \frac{SSR}{TSS}$

R^2 & adjusted R^2 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% 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

Assumptions for causal inference in multiple regression

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

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

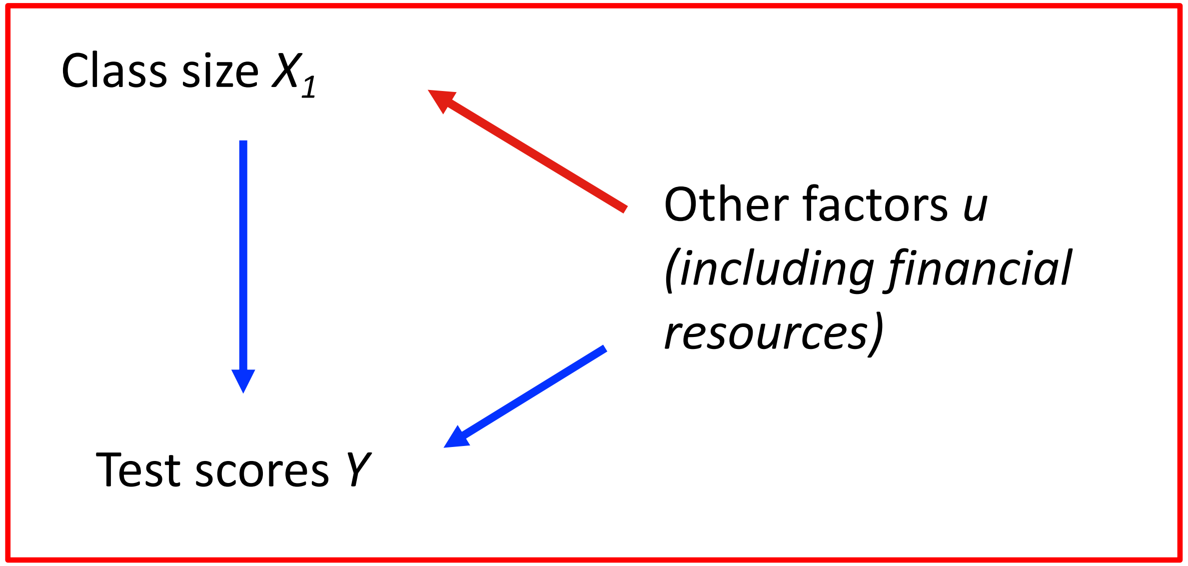
2. $(Y_i, X_{1i}, X_{2i}, \dots X_{ki})$, *i* = 1,…, *n*, are i.i.d.

3. Large outliers are rare.

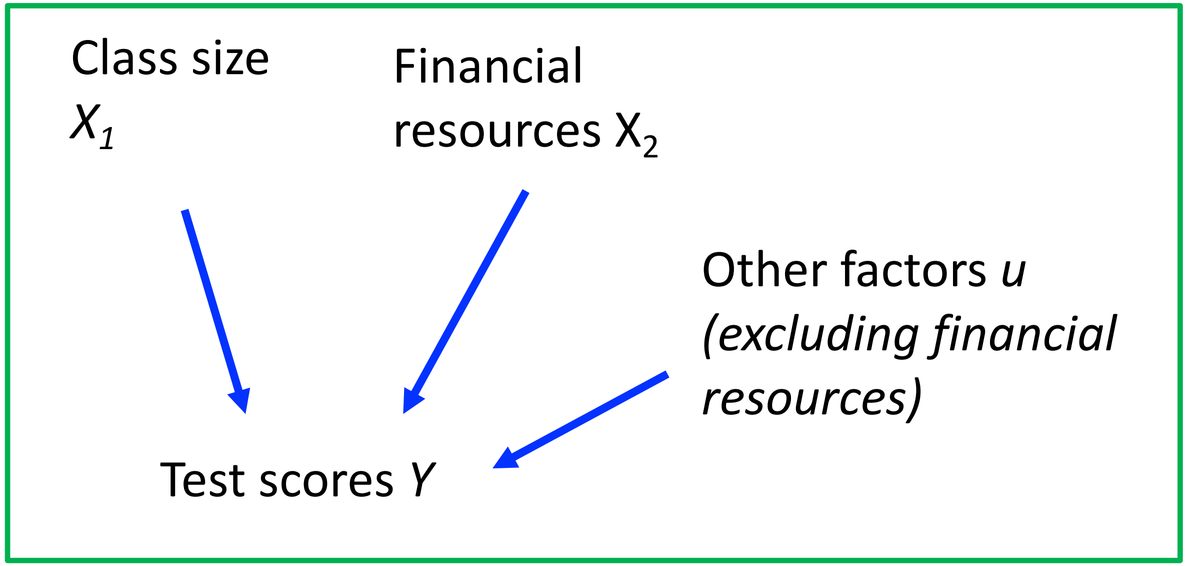
4. No perfect multicollinearity (no regressor is an exact linear function of other regressors).

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

- Another way to see assumption 1, when you are mainly interested in the effect of one particular regressor.
- X = regressor (or "treatment") of interest.
- $W_1, W_2, \ldots, W_k$ = control variables.
- Conditional Independence Assumption (CIA):

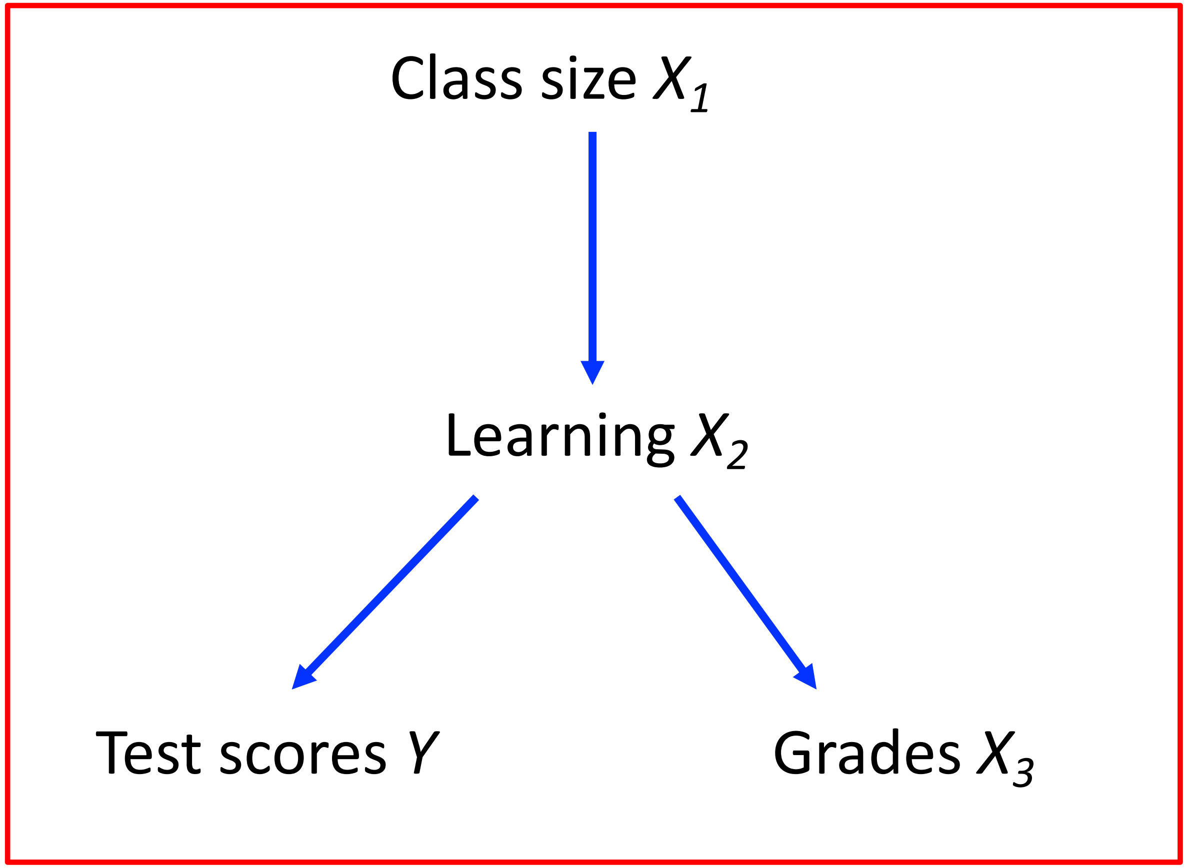
$$E(u_i|X, W_1, \ldots, W_k) = E(u_i|W_1, \ldots, W_k)$$

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 VARIABLE



- 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!**

Data Analysis Lab (5SSPP267)

Module convenors:

Dr Maia King and Professor Jonathan Portes

This module enables students to develop their **skills and confidence in data analysis in Excel,** and in presenting this analysis in **clear and accessible written reports.**

They will explore data sources and applications that are relevant for the study of **current topics in economics and social sciences.**

Students will practise their skills in **interactive weekly workshops, exploring data sourcing, analysis and visualisation** on a variety of relevant topics.

During these sessions, they will have the chance to develop and deploy their **critical thinking skills** in relation to data analysis practices in economics and social sciences, and **how data is used in public debate**.

Students will also develop important **employability skills** like communication skills, report writing, and clearly explaining complex findings.

The module **does not cover regression or other causal methods,** and therefore does not overlap with 5SSPP213 or 5SSPP241.

Teaching arrangement:

- 6 hours of lectures
- 10 weekly computer workshops of 1 hour each

Assessment:

- **1,500-word data analysis report** (40%) due in Reading Week
- **1,500-word data analysis report** (60%) due after teaching ends

Thank you for your attention