Proving Causality in Social Science: a potential application of ologs

By Noam Angrist

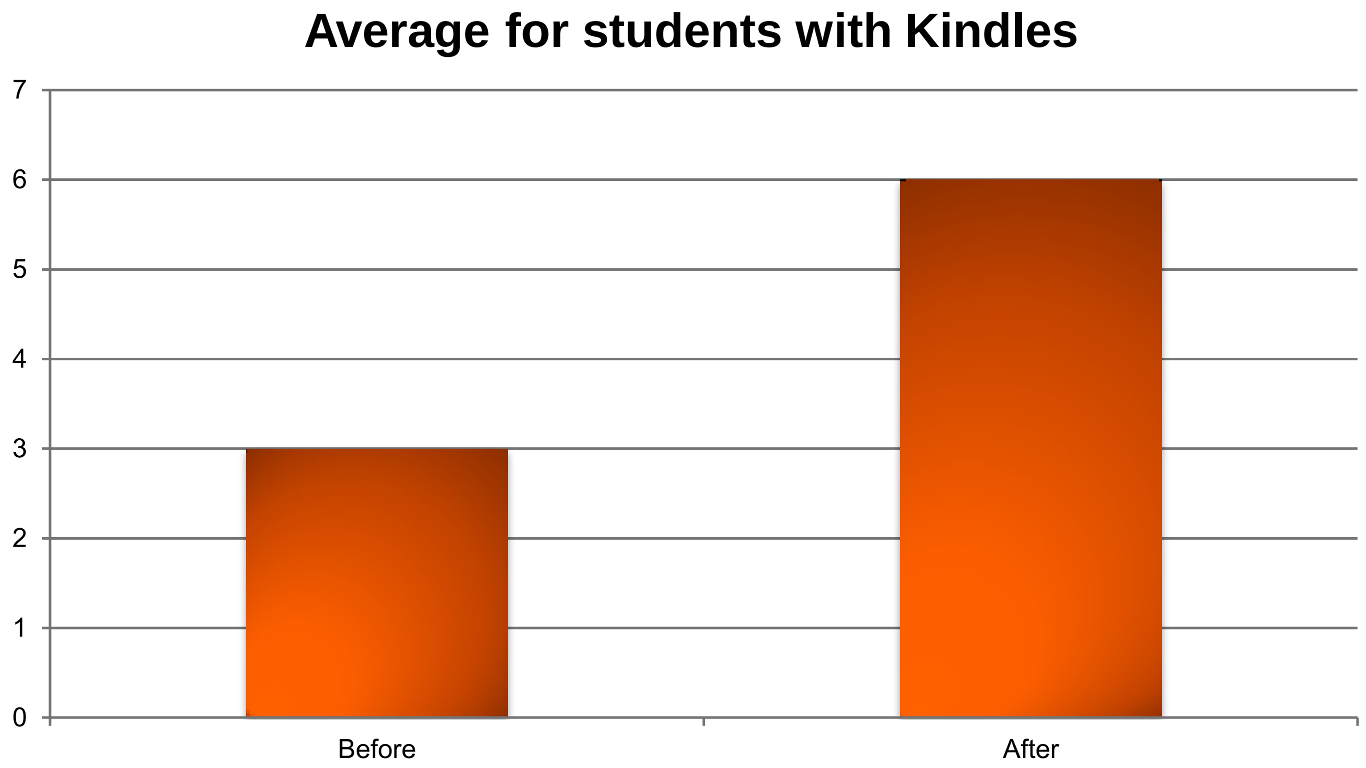
The Goals of Social Science

- Explain the world around us. What is *really* happening and why.
- **Example**: do Kindles boost test scores?

Image courtesy of Dekuwa on Flickr. Available CC BY-NC-SA.

THE GOALS OF SOCIAL SCIENCE

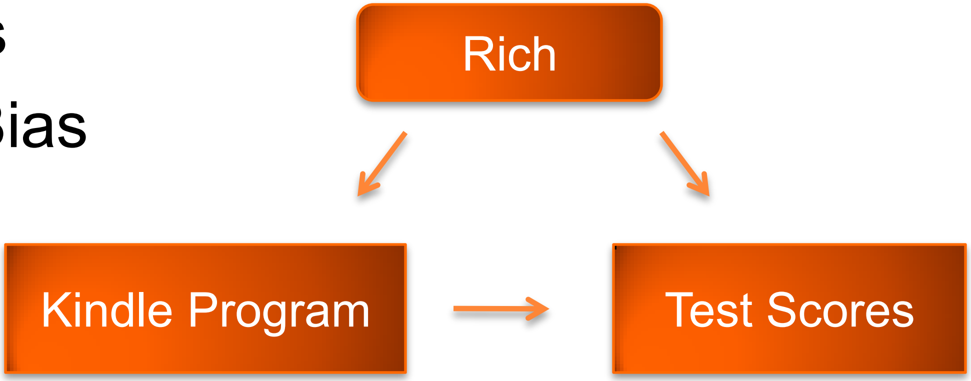
- Did the Kindle intervention work?



What's Wrong with This?

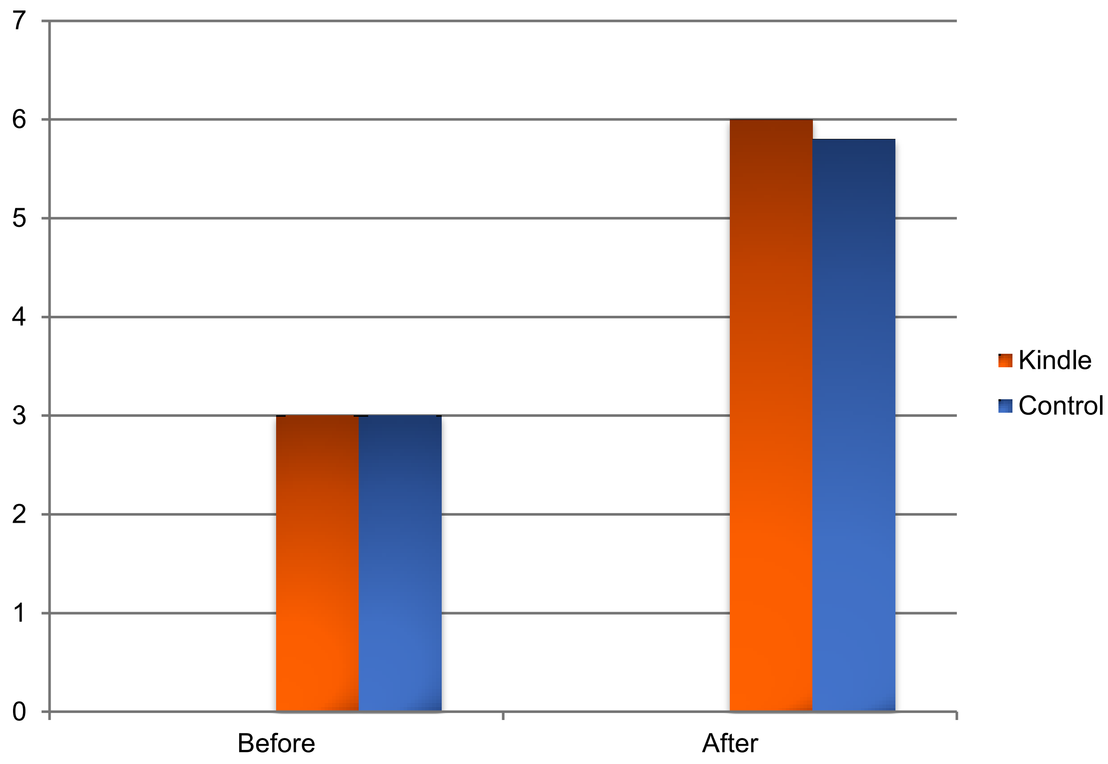
<u>Some (not all) Key Biases:</u>

- (1) Self Selection Bias
- (2) Omitted Variable Bias
- (3) Attrition Bias
- (4) Counterfactual



Conventional Methods of Addressing Biases

- Add a control group – addresses counterfactual bias



Mathematical-Based Methods of Addressing Biases

- Econometrics

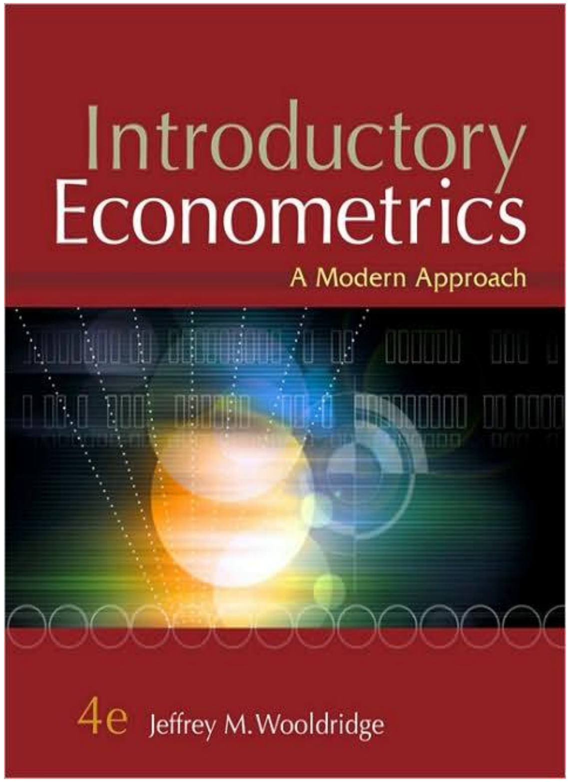
  A) Regressions

  B) Controlling

  C) Instrumental variables

  D) Randomized trials

  E) Other methods



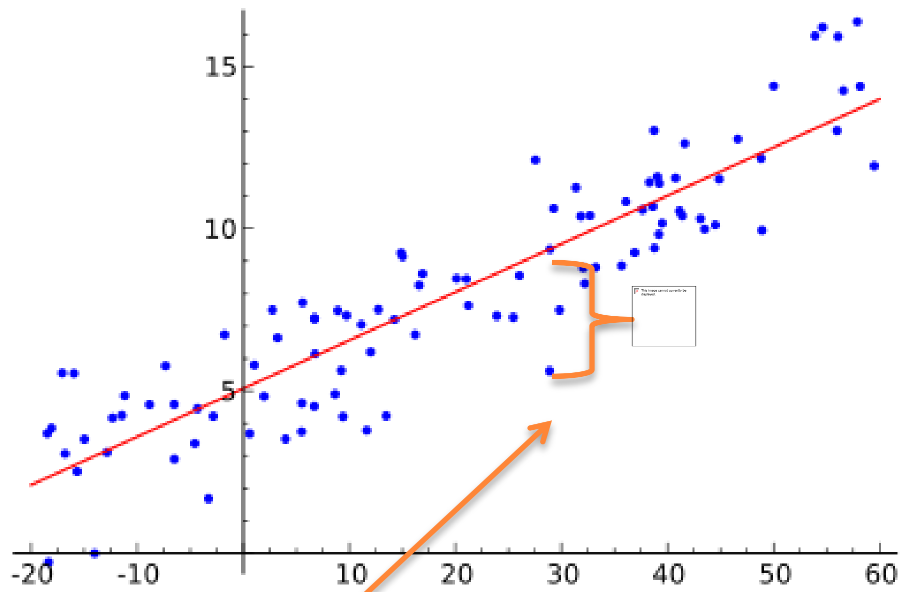
Simple Linear Regression: overview

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

where

$$\beta = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$\alpha = \bar{y} - \beta \bar{x}$$



Slope

Residual

Intercept

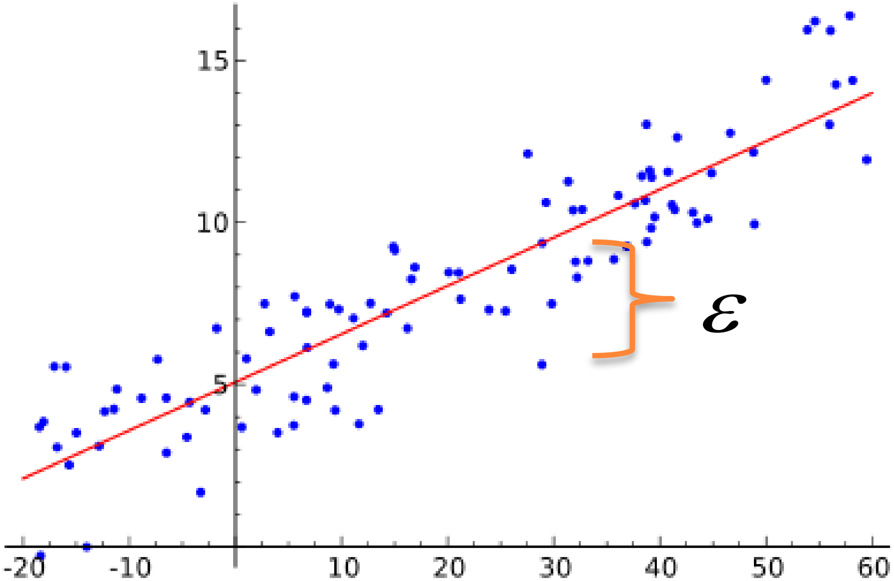
Simple Linear Regression: derivation

The goal is to minimize the sum of square residuals in order to find the line of best fit:

$$\min_{\alpha,\beta} Q(\alpha,\beta) \quad \textbf{where} \quad Q(\alpha,\beta) = \sum_{i=1}^{n} \varepsilon_i^{\,2} = \sum_{i=1}^{n} (y - \alpha - \beta x_i)^2$$

1) $$\frac{\partial}{\partial a} \sum_{i=1}^{n} (y_i - \boxed{\alpha} - \beta x_i)^2$$

2) $$\frac{\partial}{\partial b} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$



SIMPLE LINEAR REGRESSION: DERIVATION

1) $$\frac{\partial}{\partial \alpha}\sum_{i=1}^{n}(y_i-\alpha-\beta x_i)^2$$

$$=\sum_{i=1}^{n}2(y_i-\alpha-\beta x_i)(-1)=0$$

$$=-2\sum_{i=1}^{n}(y_i-\alpha-\beta x_i)=0$$

$$=-\sum_{i=1}^{n}y_i+n\alpha+\beta\sum_{i=1}^{n}x_i=0$$

This image cannot currently be displayed.

$$\alpha=\bar{y}-\beta\bar{x}$$

2) This image cannot currently be displayed.

$$=\sum_{i=1}^{n}2(y_i-\alpha-\beta x_i)(-x_i)=0$$

$$=-2\sum_{i=1}^{n}(y_i x_i-\alpha x_i-\beta x_i^2)=0$$

$$=-\sum_{i=1}^{n}x_i y_i+\alpha\sum x_i+\beta\sum_{i=1}^{n}x_i^2=0$$

Simple Linear Regression: derivation

- Plug in alpha from equation (1) into equation (2):

1) $\alpha = \bar{y} - \beta \bar{x}$

2) $= -\sum_{i=1}^{n} x_i y_i + \alpha \sum x_i + \beta \sum_{i=1}^{n} x_i^2 = 0$

$$= -\sum_{i=1}^{n} x_i y_i + (\bar{y} - \beta \bar{x}) \sum x_i + \beta \sum_{i=1}^{n} x_i^2 = 0$$

This image cannot currently be displayed.

$$\beta = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{yx}}{\sum x_i^2 - n\overline{x}^2}$$

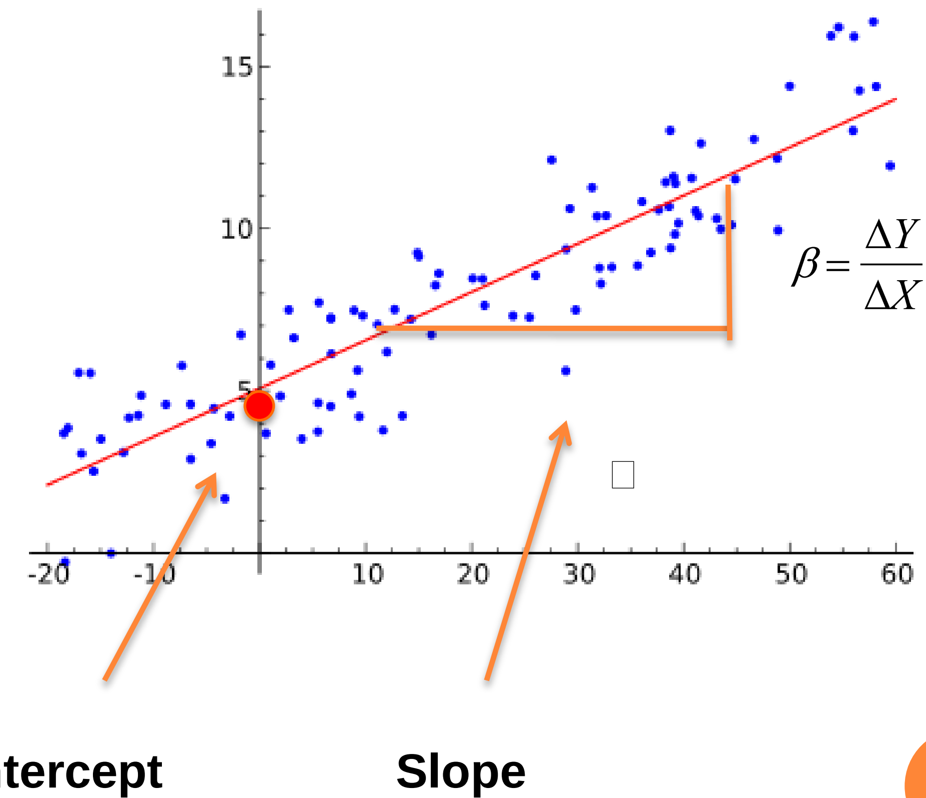
This image cannot currently be displayed.

Simple Linear Regression

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

$$\beta = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$\alpha = \bar{y} - \beta \bar{x}$$



Simple Linear Regression: An Example

- Does a kindle club (as described before) boost test scores?

- Let's find out!

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

where

$$Y_2 = \Delta testscores$$

$$X_i = participatinginKC$$

Dependent Variable

Independent Variable

READING PROGRAM: TEST SCORES

```
. regress diff kc

      Source |       SS       df       MS              Number of obs =     223
-------------+------------------------------           F(  1,   221) =    2.43
       Model |  .829634148     1  .829634148           Prob > F      =  0.1204
    Residual |  75.4243508   221  .341286655           R-squared     =  0.0109
-------------+------------------------------           Adj R-squared =  0.0064
       Total |   76.253985   222  .343486419           Root MSE      =   .5842

------------------------------------------------------------------------------
        diff |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          kc |    .213468   .1369144     1.56   0.120    -.0563569    .4832928
       _cons |    .244532   .0410026     5.96   0.000     .1637258    .3253382
------------------------------------------------------------------------------
```

Where:

diff is the difference in test score over 8 weeks

and

Kc is a dummy variable that equals 1 id a student participated in the club and 0 if they didn't

Thus, we have:

$$\beta = .213468$$

$$\alpha = .244532$$

Result:

Participation in the Kindle Club results in an increase of .2134 on standardized test scores relative to all students in the school (everyone else increased around .22 naturally)

ECONOMETRICS: CONTROLLING

I expand the simple linear regression to include more independent (or predictor) variables:

$$Y_i = \alpha + \beta_{i,1} X_{i,1} + \beta_{i,2} X_{i,2} + \beta_{i,3} X_{i,3} + \beta_{i,n} X_{i,n} + \varepsilon_i$$

Multiple regression allows me to control for certain characteristics (i.e. I can determine relationships holding/given certain variables constant).

This takes into account **covariance** among variables.

Intuition: conditional probabilities

Reading Program: controlling

I control for (1) income status and (2) grade level:

```
. regress diff kc maincomestatus0noneornotdefined1 grade

      Source |       SS       df       MS              Number of obs =     222
-------------+------------------------------           F(  3,   218) =    3.46
       Model |  3.46703073     3  1.15567691           Prob > F      =  0.0171
    Residual |  72.7171155   218  .33356475            R-squared     =  0.0455
-------------+------------------------------           Adj R-squared =  0.0324
       Total |  76.1841462   221  .344724643           Root MSE      =  .57755

------------------------------------------------------------------------------
        diff |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          kc |   .2540788   .1362198     1.87   0.063    -.0143975    .5225552
maincomest~1 |   .1029078   .0672073     1.53   0.127    -.0295514     .235367
       grade |  -.1076672   .0475764    -2.26   0.025    -.2014359   -.0138986
       _cons |   .9019443   .3431936     2.63   0.009     .2255422    1.578346
------------------------------------------------------------------------------
```

Result: our estimate for impact of kindle club participation on test score increase relative to the whole school goes up by .04

Explanation 1: free lunch, harder to improve, so had more kids with free lunch such that when we control, we have a higher impact

Explanation 2: higher grade level, less room for improvement since higher baseline so we had more kids at a higher grade level in group

DID IT WORK!?

- More issues

  - **Omitted Variable Bias**

    (can't control for everything) – factors not included in regression which impact independent and dependent variable
  - **Selection Bias**
  - **Attrition bias**

Econometrics: Instrumental Variables

Examples of Z_i	Causal Outcome
Birth Date	Returns of an extra year of schooling
Gender	Title IX affect on labor market outcomes
Twins	Family size effect on schooling

Econometrics: Instrumental Variables

- Using a random variable to "instrument" for causality such that Z has no correlation with Y outcome variable, but is highly correlated with X such that you can attribute a causal impact of X on Y

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

$$\downarrow$$

$$Y_i = \alpha + \beta Z_i + \varepsilon_i$$

Where

$$Cov(Z_{i,} \varepsilon_i) = 0$$ **and**

$$Cov(Z_{i,} X_i) > 0$$

Econometrics: Randomized Trials

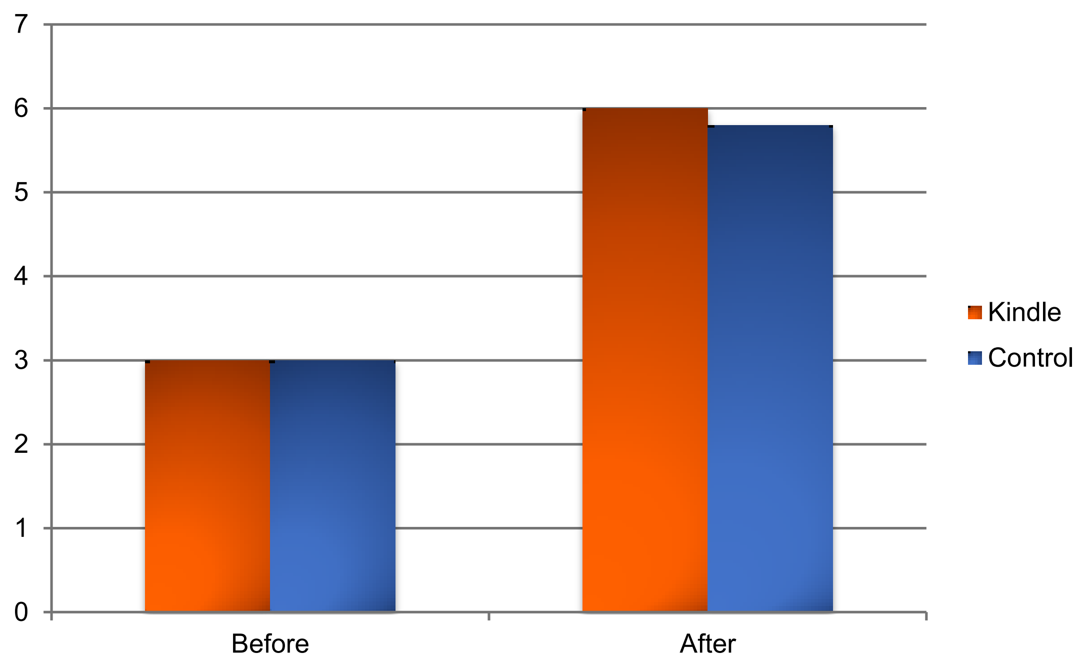
- **Program design**: randomly assign treatment and control group (like clinical trials in medicine) – eliminates motivation/demographics biases in intervention
- In this case Z_i can act as an instrumental variable since our treatment dummy variable is determined by random lottery

$$Y_i = \alpha + \beta Z_i + \varepsilon_i$$

Now, β has a casual interpretation, not just correlation

Using Random Assignment for IV approach

- Add a **random** control group – addresses counterfactual bias, omitted variable bias, self-selection



Reading Program: Randomized Trial

I regress the dummy treatment variable X (1 if randomly selected into the KC, 0 if randomly not selected) on the difference in test scores after 8 weeks

```
. regress diff treated2

      Source |       SS       df       MS              Number of obs =     223
-------------+------------------------------           F(  1,   221) =    4.57
       Model |  1.54468656     1  1.54468656           Prob > F      =  0.0336
    Residual |  74.7092984   221  .338051124           R-squared     =  0.0203
-------------+------------------------------           Adj R-squared =  0.0158
       Total |   76.253985   222  .343486419           Root MSE      =  .58142

------------------------------------------------------------------------------
        diff |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    treated2 |   .2685595   .1256352     2.14   0.034      .020963    .5161559
       _cons |   .2347739   .0412159     5.70   0.000     .1535474    .3160003
------------------------------------------------------------------------------
```

Result: our estimate for impact of kindle club on test score increase is .26 of a reading level (causal since relative to random control group)

Note 1: this is a rigorous result. Also, notice that the regression with controls yields a result closest to the controlled regression
Note 2: it is critical to check for statistical significance
Note 3: measuring intention to treat effect, so underestimate of impact

Is Category Theory Useful for Social Scientists?

3.1.2 Monoid actions

Definition 3.1.2.1 (Monoid action). Let $(M, e, \star)$ be a monoid and let S be a set. An *action of* $(M, e, \star)$ *on* S, or simply an *action of* M *on* S or an M-*action on* S, is a function

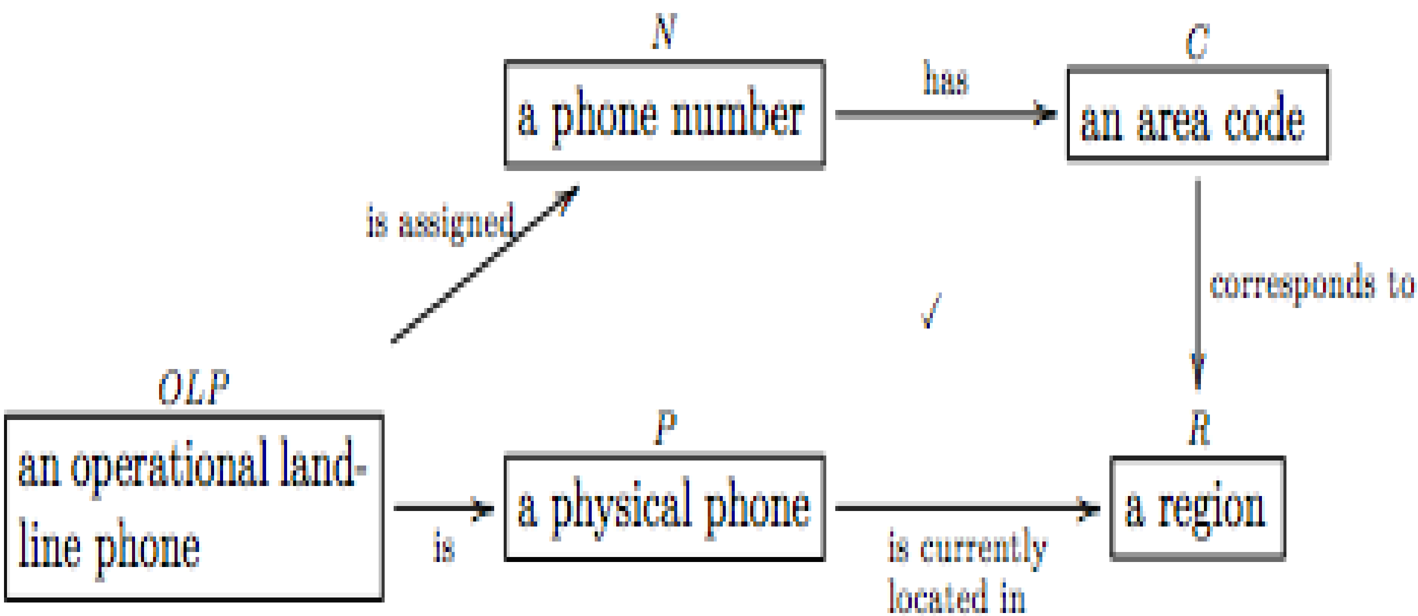
$$\circlearrowleft\ : M \times S \to S$$

such that the following conditions hold for all $m, n \in M$ and all $s \in S$:

- $e \circlearrowleft s = s$
- $m \circlearrowleft (n \circlearrowleft s) = (m \star n) \circlearrowleft s$. [4]

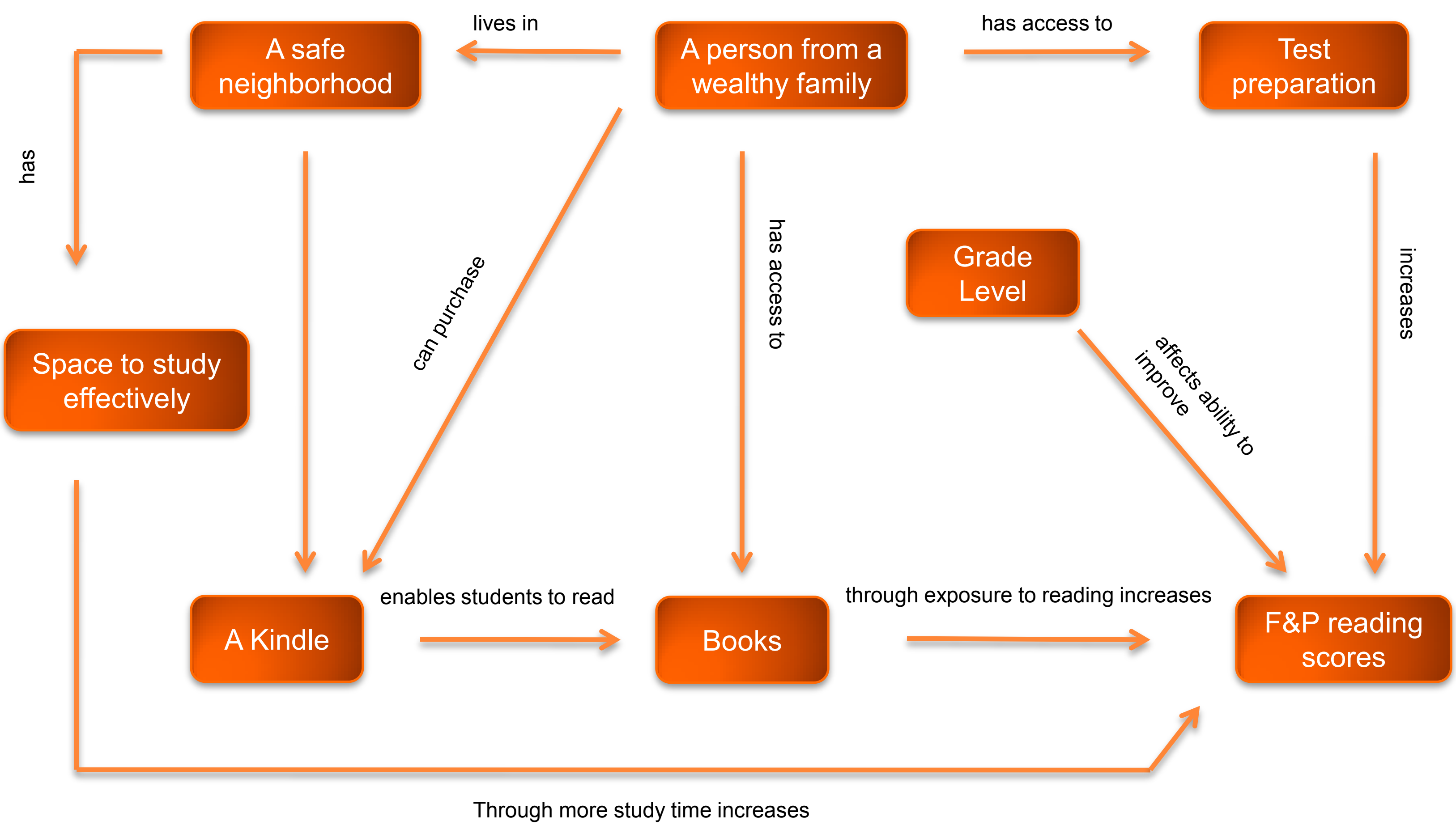
So where Does Category Theory Come In?

- Defining and determining Omitted Variable Bias through some comprehensive olog

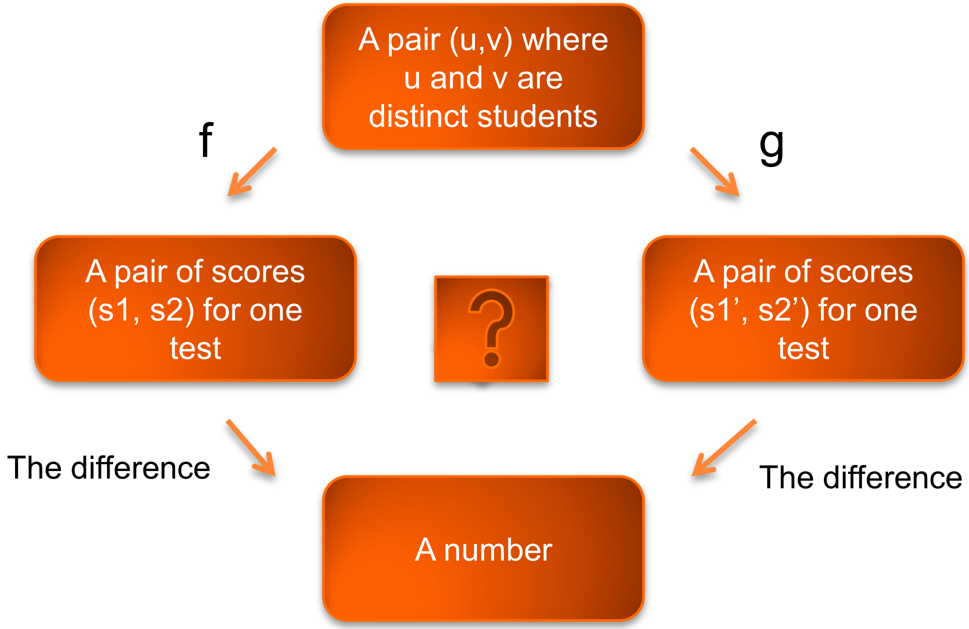


(2.23)

ONE POTENTIAL OLOG: KINDLES AND TEST SCORES…WHICH DOESN'T WORK BUT IS USEFUL NONETHELESS (NOT FUNCTIONS FROM SETS TO SETS)



A REAL OLOG



f: sends v to a bad school without a Kindle; send u to a good school without a Kindle.
g: send v to a bad school with a Kindle; send u to a good school with a Kindle.

Note: choose a global variable which captures effects from other variables

What can I conclude from this olog?

- Creating an olog helps the social scientist think through the various processes and factors which might affect our outcomes of interest

- There are multiple sources of omitted variable bias
  - The process of creating an olog helps a scientist determine a comprehensive system which can includes as many factors as the social scientist deems relevant

- “A wealthy family” captures many of the omitted variables, seen by the connecting arrows
  - thus controlling for having a wealthy family should yield estimates close to those causal estimated using a randomized controlled trial

Reading Program: controlling

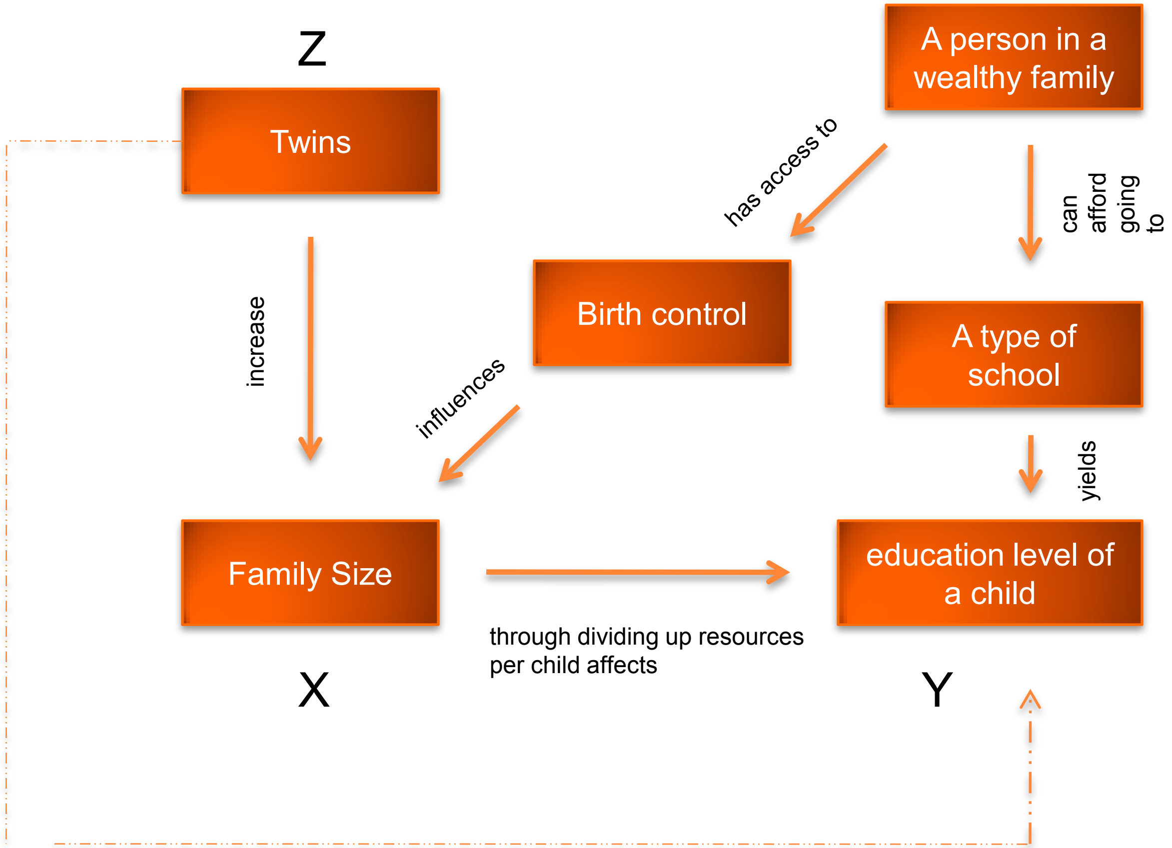
I control for (1) income status and (2) grade level:

```
. regress diff kc maincomestatus0noneornotdefined1 grade

      Source |       SS       df       MS              Number of obs =     222
-------------+------------------------------           F(  3,   218) =    3.46
       Model |  3.46703073     3  1.15567691           Prob > F      =  0.0171
    Residual |  72.7171155   218   .33356475           R-squared     =  0.0455
-------------+------------------------------           Adj R-squared =  0.0324
       Total |  76.1841462   221  .344724643           Root MSE      =  .57755

------------------------------------------------------------------------------
        diff |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          kc |   .2540788   .1362198     1.87   0.063    -.0143975    .5225552
maincomest~1 |   .1029078   .0672073     1.53   0.127    -.0295514     .235367
       grade |  -.1076672   .0475764    -2.26   0.025    -.2014359   -.0138986
       _cons |   .9019443   .3431936     2.63   0.009     .2255422    1.578346
------------------------------------------------------------------------------
```

Instrumental variables application using Ologgy-like stuff (another broken but useful olog)



Conclusion

- If we design ologs before our analysis phase we can make sure that:
  - (1) we come up with credible instrumental variables
  - (2) when we control for all relevant variables that might have otherwise been omitted and determine which variables can proxy for others
- This is important because:
  - Randomized trials are expensive and we often resort to controlling as an alternative option to determine causal relationships
  - In the absence of randomized trials, we also need good instruments to determine causal relationships

Questions/Comments

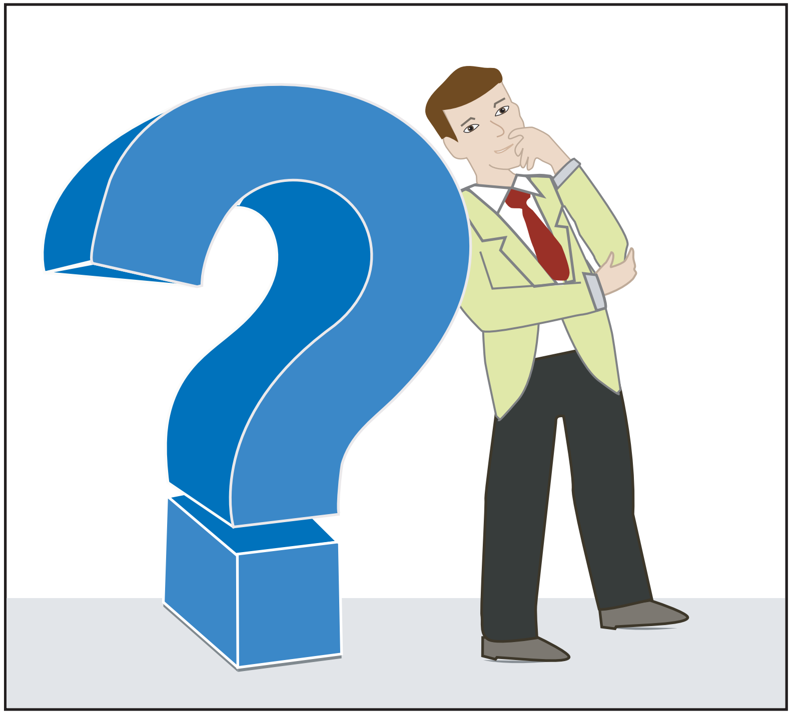


Image by MIT OpenCourseWare.