Econometrics 1

Chapter 4: linear regressions and causality

Xavier D'Haultfœuille and Elia Lapenta

Introduction

- So far, we have considered regressions as a tool for prediction in a stable environment.
- But another key motivation is measuring causal effects.
- For instance, a regression of wages on education shows that one additional year of education predicts a wage that is $\simeq$8% higher.
- Does it mean that each year of schooling causes this increase ?
- Or is it that more educated people differ from less educated people in other dimensions (for instance, motivation) ?
- Often crucial to answer policy questions. Here, for instance : should we subsidize education ? Or subsidize it more ?
- In this chapter, we study under which conditions we can give a causal interpretation to regressions.

Outline

The Neyman-Rubin causal framework

- We are interested in the causal effect of a binary variable (often called a « treatment ») $D \in \{0, 1\}$ on a variable Y.
- Examples :
  - Effect on a health measure (Y) of going ($D = 1$) or not ($D = 0$) to the hospital, for elderly people.
  - Effect of a job training program on unemployment duration and future earnings of unemployed people.
- To model causality, we introduce the *potential outcomes* $Y(0)$ and $Y(1)$.
- $Y(1)$=outcome the individual would have if she received the treatment.
- $Y(0)$=outcome the individual would have without the treatment.
- Key issue : we never observe together $Y(0)$ and $Y(1)$ for the same individual !
- We only observe $Y = Y(D) = DY(1) + (1 - D)Y(0)$. We also observe D.

Causal effects

- The causal effect of the treatment is then equal to

$$\Delta = Y(1) - Y(0)\,.$$

  In general, Δ is random : it varies from one individual to another.

- We will focus on :
  - the average treatment effect, $\delta = E[\Delta]$;
  - the average treatment effect on the treated, $\delta^T = E[\Delta|D = 1]$.
- δ measures the average effect of the treatment if everybody switched from being « untreated » to being « treated » .
- δ^T measures the average effect only among those that are treated.
- In general, $\delta \neq \delta^T$. We have $\delta^T > \delta$ if the treatment is given to those who benefit the most from it.

Potential outcomes : a fictitious illustration

- Take the hospital example. Outcome=good health (=1) or bad health (=0) at the end of the year.
- Suppose there are four types of individuals, such that :
  - Type 1 (probability=1/2) does not go to hospital and is in good health. She would be also fine if she went to hospital.
  - Type 2 (probability=1/4) goes to hospital and is in good health. She would be in bad health if she did not go to hospital.
  - Type 3 (probability=1/8) does not go to hospital and is in bad health. She would be fine if she went to hospital.
  - Type 4 (probability=1/8) goes to hospital and is in bad health. She would be fine if she did not go to hospital (she got infected there).

Potential outcomes : a fictitious illustration

Table 1 – Unobserved and observed r.v. in the example

Types	Prob. of type	Unobserved r.v. $Y(0)$	$Y(1)$	Δ	Observed r.v. D	Y
1	1/2	1	1	0	0	1
2	1/4	0	1	1	1	1
3	1/8	0	1	1	0	0
4	1/8	1	0	-1	1	0

- In this example,

$$\delta = \frac{1}{2} \times 0 + \frac{1}{4} \times 1 + \frac{1}{8} \times 1 + \frac{1}{8} \times (-1) = \frac{1}{4},$$

$$\delta^T = \frac{1/4}{1/4 + 1/8} \times 1 + \frac{1/8}{1/4 + 1/8} \times (-1) = \frac{1}{3}$$

- Therefore, $\delta^T > \delta$ here (but this need not be the case in general).

- How can we estimate δ or δ^T, given that we only observe D and Y ?
- Issue : δ and δ^T involve unobserved quantities.
- The following proposition links δ^T to the coefficient β_D of the theoretical (population) linear regression of Y on D.

Proposition 1

$\beta_D = \delta^T + B$ with $B = E[Y(0)|D=1] - E[Y(0)|D=0]$. $\beta_D = \delta^T$ iff $B = 0$ or, equivalently, $Cov(D, Y(0)) = 0$.

Proof : recall that in the theoretical regression of Y on $D \in \{0,1\}$,

$$\beta_D = E[Y|D=1] - E[Y|D=0].$$

Then :

$$\begin{aligned} \delta^T &= E[Y(1)|D=1] - E[Y(0)|D=0] + E[Y(0)|D=0] - E[Y(0)|D=1] \\ &= \beta_D - B. \end{aligned}$$

The first and the second conclusion follow. To obtain the third, follow the same proof as in Slide 9 of Chapter 1 to show that

$$\frac{\text{Cov}(D, Y(0))}{V(D)} = E(Y(0)|D=1) - E(Y(0)|D=0) = B \ \square.$$

The selection problem

- Thus, $\widehat{\beta}$ does not converge to the causal effect δ^T in general.
- However, it does converge to δ^T if $B = 0$.
- B is called the selection bias.
- The name comes from that, in terms of $Y(0)$, the individuals who receive the treatment are different from those who do not receive the treatment, so that

$$E[Y(0)|D = 1] - E[Y(0)|D = 0] \neq 0.$$

- In the fictitious example above, $B = -7/15$ so $\beta_D = -2/15$, of opposite sign to $\delta^T > 0$.
- Unfortunately, with real data, we cannot compute B or test $\text{Cov}(D, Y(0)) = 0$, because we do not observe $Y(0)$ when $D = 1$!
- But we can often have an idea on the sign of B.

Example : hospitalization and health

- Let us look at real data, by using two questions of the US NHIS survey :
  1. During the last 12 months, have you been admitted to hospital for at least one night ?
  2. Would you say that in general, your health is excellent (5), very good (4), good (3), fair (2), bad (1) ?
- Results of the 2005 survey :

Table 2 – Hospitalization and health

Admitted	Sample size	Average health	Std. error
Yes	7,774	3.21	0.014
No	90,049	3.93	0.003

Notes : table taken from Angrist & Pischke, p.13.

- What is $\widehat{\beta}_D$ here ? Do we believe $\delta^T = \beta_D$ (i.e. that $\widehat{\beta}$ converges to δ^T) ? What is the likely sign of B ?

Example #2 : effect of class size on students' achievement

- Do students in small classrooms (e.g., ≤ 20) perform better at school ?
- Important question, since reducing class sizes is very costly.
- Let us use the French 1997 panel of pupils, and in particular :
  - Class sizes in 1st and 2nd grade ;
  - Scores at national tests in the beginning of 3rd grade.

Table 3 – Class size and students' achievement

Small class	Sample size	Average class size	Average test score (./100) French	Maths
Yes	1, 903	17.8 (0.046)	75.4 (0.37)	66.9 (0.40)
No	5, 433	24.0 (0.026)	76.9 (0.20)	67.4 (0.23)

Source : 1997 panel from the French Ministry of Education. Std. errors under parentheses.

- Should we increase class sizes ? What is the likely sign of B ?

An ideal set-up : randomized experiments

- In such experiments, individuals initially selected are randomly affected
  - either in the « treatment » group ($D = 1$)
  - or in the « control » group ($D = 0$).
- Very common in medicine, but also increasingly popular in social sciences.
- In this case, $D \perp\!\!\!\perp (Y(0), Y(1))$. Then, $\text{Cov}(D, Y(0)) = 0$ so $\beta_D = \delta^T$.
- Also,

$$\begin{aligned}\delta^T &= E(Y(1)|D = 1) - E(Y(0)|D = 1) \\ &= E(Y(1)) - E(Y(0)) \\ &= \delta.\end{aligned}$$

$\Rightarrow$ The selection bias disappears and $\beta_D = \delta = \delta^T$.

Example #2 (c'ed) : lesson from an experiment

- Tenessee launched in 1985 the project STAR to evaluate the effect of an important class size reduction in kindergarten.
- Children and teachers randomly assigned in three groups : « small » (13-17 pupils), « normal with help » (22-25 pupils) and « normal without help » .

Table 4 – Differences wrt to the "normal without help" group

Treatment	Sample size	Diff. in avg class size	Diff. in avg test score in 1st grade (./100)
Small	1 925	7.0	8.57 (1.97)
Normal with help	2 319	0.7	3.44 (2.05)

Source : Krueger (1999), "Experimental Estimates of Education Production Functions", *Quarterly Journal of Economics*. Std err. under parenth.

- Significant and large effect of a class size reduction !

$\Rightarrow$ in the French panel, the selection effect was negative and larger than the positive effect of a class size reduction.

Causal, linear representation

- We have

$$\begin{aligned} Y &= Y(0) + D\Delta \\ &= \underbrace{E(Y(0))}_{} + D\delta^T + \underbrace{\left[Y(0) - E(Y(0)) + D(\Delta - \delta^T)\right]}_{} \\ &= \quad \gamma \quad + D\delta^T + \quad \eta \end{aligned} \tag{1}$$

- By construction $E(\eta) = 0$. But in general $E(D\eta) \neq 0$.
- We thus have two linear representations, which differ in general :
  - A non-causal one : $Y = \alpha_0 + D\beta_D + \varepsilon$, with $E(\varepsilon) = E(D\varepsilon) = 0$ (simple linear projection) ;
  - A causal one : $Y = \gamma + D\delta^T + \eta$, with $E(\eta) = 0$ but $E(D\eta) \neq 0$ in general.
- However, if $\text{Cov}(D, Y(0)) = 0$, $(\gamma, \delta^T, \eta) = (\alpha_0, \beta_D, \varepsilon)$ and $E(D\eta) = 0$.
- If not, $\alpha_0 + \beta_D D$ is simply the best (linear) prediction of Y by D.

- We cannot test $\text{Cov}(D, Y(0)) = 0$, because we do not observe $Y(0)$ when $D = 1$.
- Yet, we can test close conditions.
- Idea : if D is *uncorrelated* with X, and X is a variable *correlated* with $Y(0)$, then we have some evidence supporting $\text{Cov}(D, Y(0)) = 0$.
- We can test $\text{Cov}(D, X) = 0$, as both D and X are observed.

Example #2 (c'ed) : testing the absence of selection

- We check here that randomization was done correctly.
- Let X be socio-demographic characteristics of a child. We must have $\text{Cov}(X, D) = 0$ or equivalently, $E(X|D = 1) = E(X|D = 0)$.

Table 5 – Differences in socio-demo. background between groups

Variable	Groups			p-value of the equality test
	Small	Normal with help	Normal without help	
Free meal	0.47	0.50	0.48	0.09
White or Asian	0.68	0.66	0.67	0.26
Age in 1985	5.44	5.42	5.43	0.32

Source : Krueger (1999), "Experimental Estimates of Education Production Functions", *Quarterly Journal of Economics*.

- Conclusion ?

Outline

Causal model with a non-binary D

- We now consider a treatment D that is non-binary but ordered and cardinal.
- Examples : class sizes, years of schooling, years of experience, unemployment insurance benefits, prices...
- We still let $Y(d)$ be the potential outcome if the treatment is equal to d. Again, we only observe $Y = Y(D)$.
- For simplicity, we suppose hereafter a linear effect of D, i.e. for some d_0

$$Y(d) = Y(d_0) + \Delta(d - d_0). \qquad \text{(Lin. effects)}$$

- Remarks :
  1. The choice of d_0 is irrelevant : if (Lin. effects) holds for some d_0, it will hold as well for any other $d_1 \neq d_0$;
  2. Δ is random $\Rightarrow$ the slope may vary from one individual to another ;
  3. When D is continuous, Δ may be interpreted as the marginal effect of the treatment, $\Delta = \partial Y(d)/\partial d$.

Accounting for nonlinearities

- (Lin. effects) is violated if $Y(d)$ depends nonlinearly on d.
- But it can be made valid by considering transformations of $Y(d)$ or d.
- Example of wages : we commonly assume that education or experience have a multiplicative effect on wages.
- If $Y(d) = Y(d_0)\exp(\Delta(d - d_0))$, then (Lin. effects) holds by considering $\ln(Y(d))$ instead of $Y(d)$.
- Another common model : the log-log model :

$$Y(d) = A \times d^{\Delta} \tag{2}$$

  $\Rightarrow$ (Lin. effects) holds on $\ln(Y(d))$ and $\ln(d)$.

- Example for (2) : $Y(d)$ = demand for a given product at a price d in an area (e.g., city). Then (2) assumes a constant elasticity per city.

$\Rightarrow$ we only need a linear relationship between a known transform of $Y(d)$ and a known transform of d.

Linear regression with a non-binary D

- Let $W = (D - E(D))^2 / V(D)$. W may be seen as a random weight, i.e. $W \geq 0$ and $E(W) = 1$.
- Let us then consider the causal effect $\delta^W = E[W\Delta]$.
- $\delta^W =$ weighted average marginal effect, where individuals far from the treatment mean have a larger weight.
- Let β_D still denote the coefficient of the theoretical (population) regression of Y on D (i.e. β_D is the probability limit of the OLS estimator $\widehat{\beta}$).
- We have the following analog of Proposition 1, for a non-binary D :

Proposition 2

Suppose (Lin. effects) *and* $Cov(D, Y(d)) = 0$ *for all* d. *Then* $\beta_D = \delta^W$.

Proof : we have, for any $d \neq d_0$,

$$\begin{aligned}\text{Cov}(D, \Delta) &= \text{Cov}\left(D, \frac{Y(d) - Y(d_0)}{d - d_0}\right) \\ &= 0.\end{aligned}$$

As a result,

$$\begin{aligned}\text{Cov}(D, Y) =& \text{Cov}(D, Y(d_0) + \Delta(D - d_0)) \\ =& \text{Cov}(D, \Delta D) - d_0 \text{Cov}(D, \Delta) \\ =& E[(D - E(D))D\Delta] \\ =& E[(D - E(D))^2 \Delta] \\ & \text{(since } E[(D - E(D))E(D)\Delta] = E(D)\text{Cov}(D, \Delta) = 0\text{)} \\ =& V(D)E[W\Delta].\end{aligned}$$

The result follows since $\beta_D = \text{Cov}(D, Y)/V(D)$ $\square$

Linear regression with a non-binary D

- In the binary case, the condition $\text{Cov}(D, Y(d)) = 0$ for all d implies
$$\delta^W = \delta^T = \delta\,.$$

- As in the binary case, β_D does not have a causal interpretation when $\text{Cov}(D, Y(d)) \neq 0$ in general.

- We can nevertheless always interpret $\alpha_0 + \beta_D D$ as the best linear prediction of Y by D, or the best linear approximation of $E(Y|D)$.

Exemple #2 (c'ed) : linear regression

- Krueger (1999) also considers a linear regression of Y on D =class size.
- He obtains $\widehat{\beta} = -0.85$ (std error=0.13).
- So removing one student per class increases average achievement by 0.85 (out of 100)
- Here, the average is weighted by $W = (D - E(D))^2/V(D)$.
- The linearity assumption may not be too bad here, since the vast majority of classes has between 13 and 26 students.

Outline

The absence of selection bias assumption

- Thus, the OLS estimator of Y on D converges to a causal effect (δ^T or δ^W) if $\text{Cov}(Y(d), D) = 0$ for all d.
- Unfortunately, this assumption is rarely credible outside randomized experiments.
- Even in such experiments, it may not be satisfied either.
- For instance, when $D \in \{0, 1\}$, the probability of being treated or not may vary according to certain characteristics.
- In this last case, $\text{Cov}(Y(d), D|G) = 0$ but in general $\text{Cov}(Y(d), D) \neq 0$.

Example : effect of training on unemployment

- Black et al. (2003) study the effect of a training program for unemployed people on their unemployment duration.
- Unemployed people were divided in groups depending on their characteristics and their employment history.
- Training was given to eligible unemployed people...
- ... but if unemployed people from a given group were too many, they were drawn randomly to determine who would get the training program.
- Then $\text{Cov}(Y(d), D|G) = 0$: conditionally on $G = g$, unemployed people were drawn randomly.
- But if G is correlated with $Y(d)$, $\text{Cov}(Y(d), D) \neq 0$.

The absence of conditional selection

- More generally, we now consider the assumption $\text{Cov}(Y(d), D|G) = 0$.
- G is a vector of random variables, not only necessary the list of group dummies : « control variables » .
- This condition neither implies nor is implied by $\text{Cov}(Y(d), D) = 0$. But it is often more credible.

Proposition 3

If $Cov(Y(d), D|G) = 0$ *and* $Y(d) = Y(d_0) + \Delta(d - d_0)$ *for some* d_0*, then*

$$E\left[\frac{Cov(Y, D|G)}{V(D|G)}\right] = \delta_G^W := E\left[W_G \Delta\right],$$

with $W_G = (D - E(D|G))^2 / V(D|G)$.

- This proposition directly extends Proposition 2 above.

$\Rightarrow$ If G is discrete, we can estimate δ_G^W by : (i) making separate regressions for each group $g \Rightarrow \widehat{\beta}_g$; (ii) forming the sample average $\widehat{\beta} = \sum_{i=1}^n \widehat{\beta}_{G_i} / n$.

Linear, causal model

- But what if G is not discrete ?
- We now rely on a stronger assumption than just $\text{Cov}(Y(d), D|G) = 0$.
- Specifically, we assume that for some d_0,

$$\begin{cases} Y(d) &= Y(d_0) + \Delta(d - d_0), & E(\Delta|D, G) = \delta_0, \\ Y(d_0) &= \zeta_0 + G'\gamma_0 + \eta, & E[\eta|D, G] = 0. \end{cases} \qquad \text{(Lin. mod. 1)}$$

- Note that (Lin. mod. 1) implies $\text{Cov}(Y(d), D|G) = 0$:

$$\begin{aligned} &\text{Cov}\left(\zeta_0 + G'\gamma_0 + \eta + \Delta(d - d_0), D|G\right) \\ =&\text{Cov}(\eta + \Delta(d - d_0), D|G) \\ =&E[(\eta + (d - d_0)\Delta)D|G] - (E[\eta|G] + (d - d_0)E[\Delta|G])E[D|G] \\ =&E[E[\eta + (d - d_0)\Delta|D, G]D|G] - (d - d_0)\delta_0 E(D|G) \\ =&E[(d - d_0)\delta_0 D|G] - (d - d_0)\delta_0 E(D|G) \\ =&0. \end{aligned}$$

Discussion and extensions of (Lin. mod. 1)

- As in (Lin. effects), the value of d_0 does not matter : if (Lin. mod. 1) holds for some d_0, it holds for any other $d_1 \neq d_0$, up to modifying ζ_0 and η appropriately.
- $E[\eta|D, G] = 0$ implies that once we control for G in $Y(d_0)$, we have *no selection* into treatment ($\text{Cov}(D, Y(d_0)|G) = 0$).
- $E(\Delta|D, G) = \delta_0$ implies that the average treatment effect is the same for all « groups » defined by G.
- The latter may be too restrictive. We can then consider a model with interactions :

$$Y(d) = Y(d_0) + (\Delta_0 + \Delta_1 G)(d - d_0).$$

Discussion and extensions of (Lin. mod. 1)

- Other limitation : D may be multivariate.
- Example : project STAR. There were 3 possible treatments : "normal without help", "normal with help", "small". Then

$$D = (\mathbb{1}\{\text{normal with help}\}, \mathbb{1}\{\text{small}\}).$$

- Even if D is univariate, we may want to allow for nonlinear effects.
- Example : D =class size. Then, we may have for instance

$$|Y(11) - Y(10)| > |Y(31) - Y(30)|.$$

- Not compatible with (Lin. mod. 1), since Δ is supposed to not depend on d.

$\Rightarrow$ Add for instance d^2 to the model on $Y(d)$.

Discussion and extensions of (Lin. mod. 1)

- If $D \in \mathbb{R}^m$, $m > 1$, we simply allow d and Δ to be multivariate in (Lin. mod. 1) :

$$\left\{ \begin{array}{rcll} Y(d) & = & Y(d_0) + \Delta'(d - d_0), & E(\Delta|D, G) = \delta_0, \\ Y(d_0) & = & \zeta_0 + G'\gamma_0 + \eta, & E[\eta|D, G] = 0. \end{array} \right. \quad \text{(Lin. mod. 2)}$$

- Then, each component D_j of D is assumed to have a marginal effect of $\partial Y(d)/\partial d_j = \Delta_j$, with $\Delta = (\Delta_1, ..., \Delta_m)'$.
- To allow for nonlinear effects, we replace the first equation of (Lin. mod. 1) by :

$$Y(d) = Y(d_0) + \Delta'(f(d) - f(d_0)), \quad E(\Delta|D, G) = \delta_0,$$

where $f(\cdot)$ is a known function (e.g., $f(d) = (d, d^2)'$).

- Then $\partial Y(d)/\partial d = \Delta' \partial f/\partial d(d)$ depends on d in general. We can then consider the average marginal effect :

$$E\left[\frac{\partial Y}{\partial d}(D)\right] = E\left[\Delta' \frac{\partial f}{\partial d}(D)\right] = \delta_0' E\left[\frac{\partial f}{\partial d}(D)\right].$$

Outline

- Let $X = (1, G', D)'$ and β_0 be the coefficients of the theoretical (population) linear regression of Y on X (i.e. probability limit of $\widehat{\beta}$).
- Then, β_D can receive a causal interpretation if (Lin. mod. 2) holds :

Proposition 4

Suppose (Lin. mod. 2) *holds and* $E(XX')$ *invertible. Then* $\beta_0 = (\zeta_0 - d_0'\delta_0, \gamma_0', \delta_0')'$.

Proof : (Lin. mod. 2) implies that

$$Y(d) = \zeta_0 - d_0'\delta_0 + G'\gamma_0 + d'\delta_0 + (\Delta - \delta_0)'d + \eta.$$

Then, defining $b_0 := (\zeta_0 - d_0'\delta_0, \gamma_0', \delta_0')'$ and $\nu := (\Delta - \delta_0)'D + \eta$, we have

$$Y = Y(D) = X'b_0 + \nu.$$

Moreover, $E(\nu|X) = E(\nu|D, G) = 0$. Thus, $E(XY) = E(XX')b_0$ and then

$$b_0 = E(XX')^{-1}E(XY) = \beta_0 \ \square$$

- In general, the equation

$$Y = X'\beta_0 + \varepsilon,\ E[X\varepsilon] = 0 \tag{3}$$

representing the theoretical (population) linear regression of Y on X does not have a causal meaning.

- It is not a causal linear model, just a linear projection of Y on X.
- Still, the coefficient β_0 in (3) has a causal interpretation if (Lin. mod. 2) holds.
- A key condition behind (Lin. mod. 2) is the absence of (conditional) selection $E(\eta|D, G) = 0$, which unfortunately is not directly testable.
- With the same reasoning as in Proposition 4, we can also estimate consistently γ_0 and δ_0 in the model with nonlinear effects :

$$\begin{cases} Y(d) &= Y(d_0) + \Delta'(f(d) - f(d_0)), & E(\Delta|D, G) = \delta_0, \\ Y(d_0) &= \zeta_0 + G'\gamma_0 + \eta, & E[\eta|D, G] = 0. \end{cases}$$

Omitted variable bias

- What if (Lin. mod. 1) holds but we consider the simple reg. of Y on $D\,(\in \mathbb{R})$?
- As before, we let :
  - β_0^S denote the slope coeff. of the theoretical reg. of Y on D ;
  - λ_0 denote the vector of slope coeffs. of D in the theoretical regressions of G^j on D, where $G = (G^1, \ldots, G^p)$
- In general the OLS estimator will not be consistent for δ_0 :

Proposition 5

Suppose (Lin. mod. 2) *holds with* $E(XX')$ *invertible. Then* $\beta_0^S = \delta_0 + \lambda_0'\gamma_0$.

Proof : by Proposition 7 in Chapter 1, $\beta_0^S = \beta_D + \lambda_0'\beta_G$, where β_D (resp. β_G) is the coefficient of D (resp. G) in the theoretical regression of Y on D and G. By Proposition 4 above, $(\beta_D, \beta_G) = (\delta_0, \gamma_0)$. The result follows $\square$

Omitted variable bias

- $\lambda_0'\gamma_0$ (or its empirical version, cf. Proposition 4 Chapter 1) is called « the omitted variable bias » .
- If $\dim(G) = 1$, this bias is 0 only if $\gamma_0 = 0$ (no effect of G on $Y(d)$) or $\lambda_0 = 0$ (D is not correlated with G).
- Example : effect of class size on students' achievement :

Table 6 – Effect of class size when varying G

	List of controls G		
	None	score_grade1	score_grade1, girl, dip_f
Coeff. of D	0.132	0.056	-0.048
R^2	0.001	0.336	0.344

Source : 1997 panel from the French Ministry of Education. score_grade1=test score in grade 1. girl=1 if student is girl, 0 otherwise. dip_f : highest diploma of the father.

« Included variable » bias

- The previous discussion could make us believe it is always better to add control variables.
- This is not the case : if a control variable G is itself influenced by the treatment D, then including it may induce a bias.
- Intuition : by including G, we capture part of the causal effect of D on Y.
- Suppose $D \in \{0,1\}$ and let $G(0) \in \{0,1\}$ (resp. $G(1) \in \{0,1\}$) be the potential control variable corresponding to $D = 0$ (resp. $D = 1$).
- We only observe $Y = Y(D)$ and $G = G(D)$.
- Suppose D is randomly allocated, as in an experiment :

$$D \perp\!\!\!\perp (Y(0), Y(1), G(0), G(1)).$$

- Example : $D = \mathbb{1}\{\text{small class in kindergarten}\}$, $G(d) = \mathbb{1}\{\text{retention in grade 1 if allocated in a class of type } d\}$, $Y(d)$=test score in grade 2 if allocated to a class of type d.

« Included variable » bias

- Since $D \perp\!\!\!\perp (Y(0), Y(1))$, we have $\text{Cov}(D, Y)/V(D) = E[\Delta]$: the OLS estimator in the simple linear reg. converges to the average treatment effect.
- Now consider $E[\beta_G]$, with $\beta_g = \text{Cov}(Y, D|G = g)/V(D|G = g)$, coefficient of the reg. of Y on D for those such that $G = g$.
- If $\text{Cov}(Y(d), D|G) = 0$, $E[\beta_G]$ would identify a causal effect by Proposition 3. But here :

$$\begin{aligned}
\beta_g =& E(Y|D = 1, G = g) - E(Y|D = 0, G = g) \\
=& E(Y(1)|D = 1, G(1) = g) - E(Y(0)|D = 0, G(0) = g) \\
=& E(Y(1)|G(1) = g) - E(Y(0)|G(0) = g) \\
=& \underbrace{E(Y(1) - Y(0)|G(1) = g)}_{\text{Causal effect}} + \underbrace{E(Y(0)|G(1) = g) - E(Y(0)|G(0) = g)}_{\text{Selection bias}}.
\end{aligned}$$

- Example of class size and grade retention : plausibly,

$$E(Y(0)|G(1) = g) < E(Y(0)|G(0) = g).$$

Example : returns to schooling

- Let D =education and $Y(d)$ = log(potential wage) with d years of education.
- The condition $\text{Cov}(D, Y(d)) = 0$ is not credible (why ?).
- We then decide to add in the regression :
  1. Age ;
  2. Parental education ;
  3. Race and state of residence ;
  4. A test on cognitive skills unaffected by education.
- We can also add the type of profession (blue vs white collar).

Example : returns to schooling

- We use here the NLSY79, which includes a measure of cognitive skills at 16 years old (AFQT).
- Could we expect the decrease of the coeff. when moving from (1) to (4) ?
- Should we add the type of profession here ?

Table 7 – Estimates of the returns to schooling when varying the controls

Specification	(1)	(2)	(3)	(4)	(5)
Controls	None	Age dummies	(2)+ additional controls*	(3)+ AFQT	(4) + type of prof.
Coeff. of D	0, 132	0, 131	0, 114	0, 087	0, 066
	(0,007)	(0,007)	(0,007)	(0,009)	(0,010)

Source : Table 3.2.1 in Angrist & Pischke (2009). Standard errors under parentheses. * : namely parental education, race and state of residence.

Summary

- Neyman-Rubin's causal model.
- Selection bias : definition and interpretation.
- Validity of regressions without controls under $\text{Cov}(Y(d), D) = 0$.
- Absence of conditional selection : $\text{Cov}(Y(d), D|G) = 0$.
- Causal linear models (Lin. mod. 1) and its extensions.
- Link between causal effects and linear regressions.
- Omitted and included variable bias.