Applied Tools:
Combining Theory and Data to Evaluate Policy

Owen Zidar
Princeton
Spring 2020

Lecture 1b

This lecture draws from several lectures in applied economics (i.e., lectures by David Card, Raj Chetty, Pat Kline, Magne Mogstad, Kevin Murphy, Jesse Shaprio, Alex Torgovitsky, Chris Walters), parts of which are reproduced here.

Outline

Preliminaries:
Research designs, DGP, and applied modeling

What is a research design (1/2)

- Consider the effect of a treatment (e.g., tax) T on outcome y

$$y_i = \alpha + \beta T_i + \varepsilon_i$$

- Treatment is assigned based on "selection" model

$$T_i = \alpha_T + \beta_T X_i + \eta_i$$

- Treatment may be non-random: $cov(X_i, \varepsilon_i) \neq 0, cov(\eta_i, \varepsilon_i) \neq 0$
- Traditional approach to accounting for confounding factors or selection: control for observables X_i when estimating treatment effect

$$y_i = \alpha + \beta T_i + \gamma X_i + \varepsilon_i$$

What is a research design (2/2)

- Problem with this approach is that we don't know the source of variation in T_i
- There must be some reason that one person or location got treated and another did not even if they are perfectly matched on observables (e.g., twins)
  - η_i must to correlated with T_i to have variation in $T_i|X_i$
- But that same unobserved factor could also affect outcome: no way to know if $cov(\eta_i, \varepsilon_i) = 0$
- A **research design** is a source of variation in η_i that is credibly unrelated to ε_i
  - E.g., a reform that affects people above age 65 but not below. People at age 64 and 65 likely to have similar outcomes $\Rightarrow cov(\eta_i, \varepsilon_i) = 0$

Interpreting regression results

It is easier to run a regression than to understand the results

- How were the data generated? And what (part of the) data do we actually observe?
- Is the effect big or small?
- What substantively is in the error term ε_i and η_i?
- How might outcomes change if we vary treatment?

Policy evaluation requires thinking about these questions. And connecting regressions with applied statistical and economic models is very helpful for thinking through these issues

How were the data generated?

- How were the data generated? A Data Generating Process (DGP) is a complete characterization of how the data were generated
  - Parametric example

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$
$$(X_i, \varepsilon_i) \sim \mathcal{N}\left(\begin{bmatrix}\mu_1 \\ \mu_2\end{bmatrix}, \begin{bmatrix}\sigma_1^2, \sigma_{12} \\ \sigma_{12}, \sigma_2^2\end{bmatrix}\right)$$
$$(\beta_0, \beta_1, \mu_1, \mu_2, \sigma_1, \sigma_{12}, \sigma_2) \in \mathbb{R}^7$$

  - Everything there is to know about the data in 7 numbers!
  - Important to ask what (part of the) data do we actually observe? And what underlying model led to what we are able to see?

Importance of DGPs for policy recommendations: WWII example

Importance of DGPs for policy recommendations: WWII example

WWII data: section of plane and bullet holes per square foot

- Engine 1.11
- Fuselage 1.73
- Fuel system 1.55
- Rest of plane 1.8

More bullet holes in the fuselage, not so many in the engines $\Rightarrow$ protect fuselage?

- Policy question – where should plane armor be allocated?
- Wald's insight: ask where are the missing holes. Don't see planes that didn't come back
- Intuition: if you go to the recovery room at the hospital, you'll see a lot more people with bullet holes in their legs than people with bullet holes in their chests; not because people don't get shot in the chest; it's because the people who get shot in the chest don't recover

Source: `https://medium.com/@penguinpress/an-excerpt-from-how-not-to-be-wrong-by-jordan-ellenberg-664e708cfc3d`

Connecting theory and empirics

Real World $\rightarrow$ Model $\rightarrow$ Real World

- (1) Make observations, (2) feed them into the model, and (3) use it to analyze key questions (e.g., what would happen if X happened?)
- What data can we look at to understand what's happening?
- How are these data generated? E.g., did agents choose to be treated?
- What predictions do different theories make?
- How does the model rationalize what we are observing?
- How should we change the model to better match the data?

The microeconomic approach to modeling

A model is a simplified representation of reality that gets to the essence of what is going on

To understand how and why something happens, look at individual behavior

- Who are the people making the choices?
- What does each person want?
- What decisions are optimal, given constraints?
- How do the decisions of different players interact?
- What adjusts if the choices aren't mutually consistent?

Applying the microeconomic approach

Main application: Markets

- Big picture: Supply and Demand
- Consumers choose what to buy
  - Maximize utility subject to prices & budgets
- Firms choose how much to produce, what prices to charge
  - Maximize profit subject to demand curves, costs
- Equilibrium: prices and quantities adjust to clear market

- What predictions do we make about the impact of supply/demand shifts, taxes, etc?
- What can we say about consumer and producer welfare of different policies?

Connecting theory and data

Overview

1. Connecting theory and data
   - Using supply and demand
   - Inequality: Katz and Murphy
   - Optimization and the value of police services
2. Potential Outcomes and Selection
3. Roy model and college choice example
4. Applied Metrics Tools

Connecting theory and data using supply and demand

Using Supply and Demand Outline

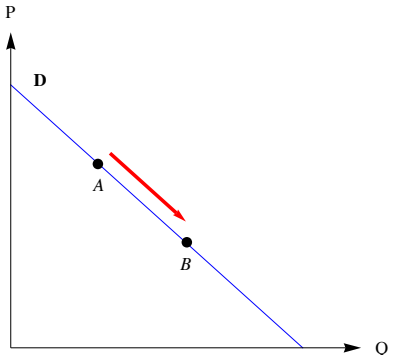
1. Quantitative supply and demand framework
2. Using supply and demand to study inequality (Katz Murphy)
3. Using supply and demand – tax example

Recall the two ways the quantity demanded can change

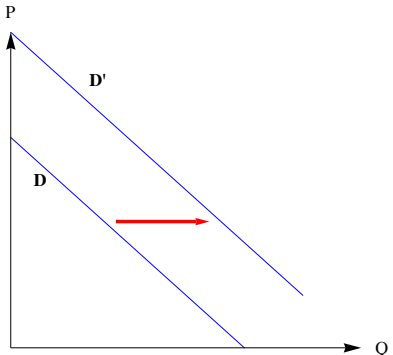
1. Moves *along* demand curve vs. 2. Shifts *of* demand curve

"Demand goes up" can mean one of two things.

Move along a demand curve:
Price falls, so quantity goes up



Shift of a demand curve:
Any P gives a higher Q



Two ways the quantity demanded can change (Math)

The quantity demanded can change in two ways:

$$\%\Delta Q^D = \underbrace{\%\Delta D}_{\text{Shift}} + \underbrace{\varepsilon^D \%\Delta P}_{\text{Movement Along}}$$

- $\%\Delta Q^D$ is the percentage change in the quantity demanded
- $\%\Delta D$ is the shift in demand in percentage terms
- $\%\Delta P$ is the percentage change in price
- ε^D is the elasticity of demand

Note that the shift and movement along are in terms of percent changes in **quantities**

Two ways the quantity supplied can change (Math)

Similarly, the quantity supplied can change in two ways:

$$\%\Delta Q^S = \underbrace{\%\Delta S}_{\text{Shift}} + \underbrace{\varepsilon^S \%\Delta P}_{\text{Movement Along}}$$

- $\%\Delta Q^S$ is the percentage change in the quantity supplied
- $\%\Delta S$ is the shift in supply in percentage terms
- $\%\Delta P$ is the percentage change in price
- ε^S is the elasticity of supply

Note that the shift and movement along are in terms of percent changes in **quantities**

Unified Framework

What do we know?

1. $\%\Delta Q^D = \%\Delta D + \varepsilon^D \%\Delta P$
2. $\%\Delta Q^S = \%\Delta S + \varepsilon^S \%\Delta P$

In equilibrium, the change in quantity demanded and supplied have to be the same:

$$\%\Delta Q^D = \%\Delta Q^S$$
$$\%\Delta D + \varepsilon^D \%\Delta P = \%\Delta S + \varepsilon^S \%\Delta P$$

Implications for Prices and Quantities

The magnitude of price changes reflect four forces:

$$\%\Delta P = \frac{\%\Delta D - \%\Delta S}{\varepsilon^S - \varepsilon^D}$$

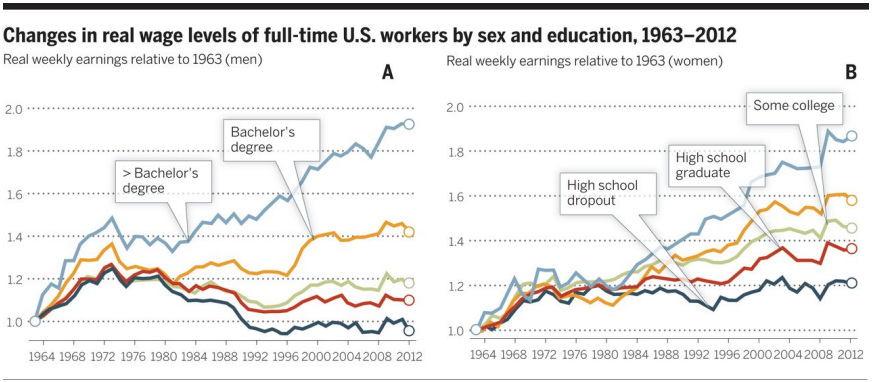
We can use this price change to determine the quantity change:

$$\%\Delta Q = \%\Delta S + \varepsilon^S \left(\frac{\%\Delta D - \%\Delta S}{\varepsilon^S - \varepsilon^D} \right)$$

$$\%\Delta Q = \frac{-\varepsilon^D \%\Delta S + \varepsilon^S \%\Delta D}{\varepsilon^S - \varepsilon^D}$$

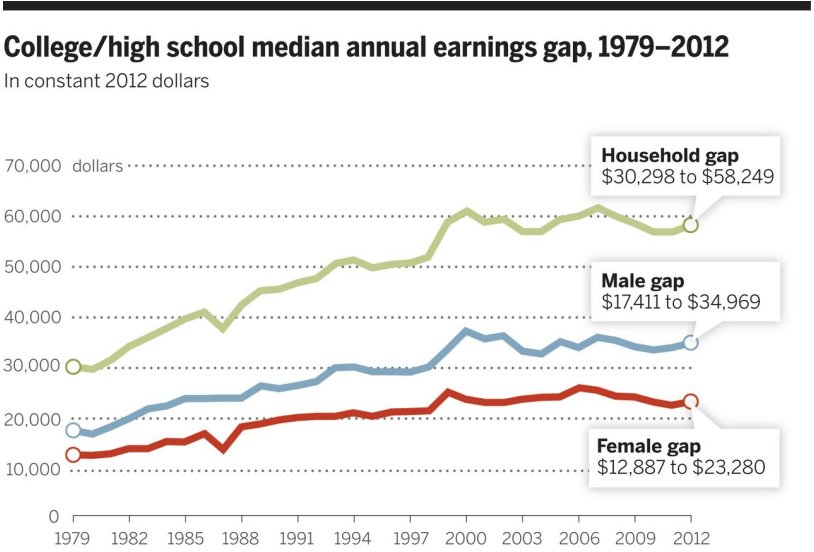
Bottom line: the quantity change is a an elasticity-weighted average of shifts in supply and demand

Application: Rise in Wage Inequality (from D. Autor)

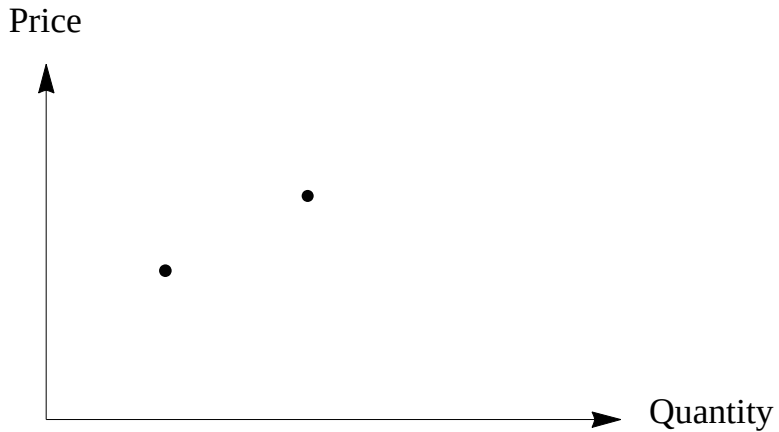


From David Autor. Science 23 May 2014: Vol. 344 no. 6186 pp. 843-851

College vs. High-school Gap in Median Earnings (D. Autor)

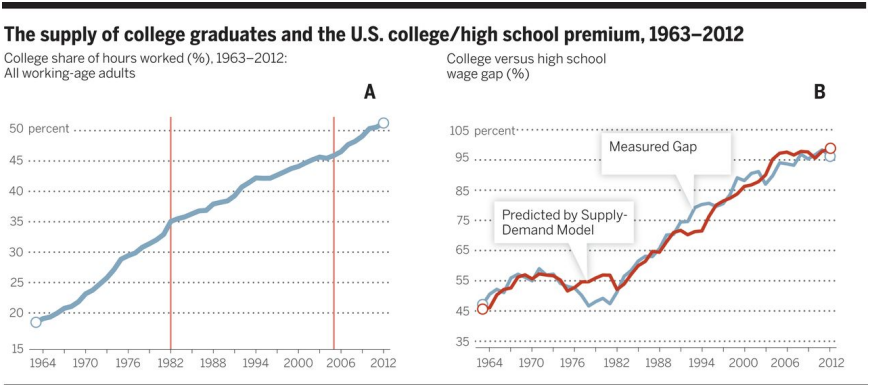


What do we actually observe (Katz-Murphy Example)



Supply has increased, but outpaced by demand

There's a "race between education and technology" (Goldin and Katz)



Katz Murphy

Overview

$$\log \frac{w_{1t}}{w_{2t}} = \alpha - \beta \log \frac{L_{1t}}{L_{2t}} + \delta_t + e_t \tag{1}$$

- $\frac{w_{1t}}{w_{2t}}$ is relative wages of college-educated workers
- $\frac{L_{1t}}{L_{1t}}$ is ratio of college to non-college workers
- Goal: Show how to use micro theory to interpret the relationship between relative wages and the supply and demand of college-educated workers

Setup

The standard framework of production given different sets of labor skills assumes

1. Output y is given by $y = f(K, h(L_1, L_2, ...))$, where $f(K, L) = AL^{\alpha}K^{1-\alpha}$
2. Return on capital r is exogenous
3. $h(L_1, L_2, ...)$ has a nested CES structure

- Under the first two assumptions, we have

$$\frac{\partial y}{\partial K} = (1-\alpha)AL^{\alpha}K^{-\alpha} = r$$

$$\Rightarrow K = L\left(\frac{(1-\alpha)A}{r}\right)^{1/\alpha} \text{ and } \frac{y}{K} = \frac{r}{(1-\alpha)} \tag{2}$$

- Equation 2 shows that K adjusts to match the overall supply of total labor units L, keeping y/K constant and keeping K/L on a trend path that is driven by the rate of growth of TFP.
- These assumptions are very plausible at a local level (or for "small open economies") that that the price of capital as exogenous.

Simplifying the production function

- Substituting for K, we get

$$y = AL^{\alpha}K^{1-\alpha} = A^{1/\alpha}\left(\frac{1-\alpha}{r}\right)^{\frac{1-\alpha}{\alpha}} L \tag{3}$$

  which is linear in L. Equation 3 shows that under assumptions 1-3, we can ignore capital.
- To analyze the effects of relative supply or relative technology changes (i.e., the part of technology embedded in $h()$), we need to specify the labor aggregator function.
- A good starting point is a 2-group CES model:

$$L = h(L_1, L_2) = (\theta_1 L_1^{\frac{\sigma-1}{\sigma}} + \theta_2 L_2^{\frac{\sigma-1}{\sigma}})^{\frac{\sigma}{\sigma-1}} \tag{4}$$

  where θ_1 and θ_2 are possibly trending over time.

Marginal Product of Each Group

- The marginal product of group 1 is

$$h_1(L_1, L_2) = \theta_1 L_1^{\frac{-1}{\sigma}} (\theta_1 L_1^{\frac{\sigma-1}{\sigma}} + \theta_2 L_2^{\frac{\sigma-1}{\sigma}})^{\frac{1}{\sigma-1}} = \theta_1 L_1^{\frac{-1}{\sigma}} L^{\frac{1}{\sigma}} \tag{5}$$

- Likewise,

$$h_2(L_1, L_2) = \theta_2 L_2^{\frac{-1}{\sigma}} L^{\frac{1}{\sigma}} \tag{6}$$

Setting Wages equal to marginal products

- Assuming $w_1/w_2 = h_1/h_2$ (i.e., MRTS=relative wage), we have:

$$\log \frac{w_1}{w_2} = \log \frac{\theta_1}{\theta_2} - \frac{1}{\sigma} \log \frac{L_1}{L_2} \tag{7}$$

- The slope of the relative demand curve is $-\frac{1}{\sigma}$, which is 0 if the two types are perfect substitutes, and something larger otherwise.
- This simple model is widely used to discuss "skill-biased technical change" (SBTC).

Katz Murphy

- In the traditional SBTC literature (e.g., Katz and Murphy, 1992), it is assumed that

$$\log \frac{\theta_{1t}}{\theta_{2t}} = a + bt + e_t \tag{8}$$

- leading to a model for the relationship of relative wages to relative supplies:

$$\log \frac{w_{1t}}{w_{2t}} = a + bt - \frac{1}{\sigma} \log \frac{L_{1t}}{L_{2t}} + e_t \tag{9}$$

Implementation

- Freeman (1976) and Katz and Murphy (1992) estimate models of this form, using 2 "types" of labor - high-school equivalents and college equivalents.
- Dropouts are assumed to be perfect substitutes for HS graduates with a relative efficiency of (roughly) 70%.
- Post-graduates are assumed to be perfect substitutes for college graduates with a relative efficiency of (roughly) 125%.
- People with 1-3 years of college are assumed to represent 1/2 unit of HS labor and 1/2 unit of college labor.
- (There are different conventions about whether supply should be based on the total numbers of adults in each education group, or total employees. There are also different ways to combine men and women).

Estimates

- The "magic number" is $\frac{1}{\sigma} = 0.7$, which implies $\sigma = 1.4$ (See KM, equation 19, page 69).
- It has turned out to be hard to get a model like (9) to work as well as it did in KM's study (and in Freeman, 1976) when the sample is extended to the 1990s and 2000's.
- Katz and Goldin (2008) present some estimates that have trend breaks in the last two decades and manage to get estimates in the range of $\frac{1}{\sigma} = 0.7$.

$$\log \frac{w_{1t}}{w_{2t}} = a + bt - \frac{1}{\sigma} \log \frac{L_{1t}}{L_{2t}} + e_t \tag{10}$$

See `https://economics.mit.edu/files/15391` for more discussion and extensions.

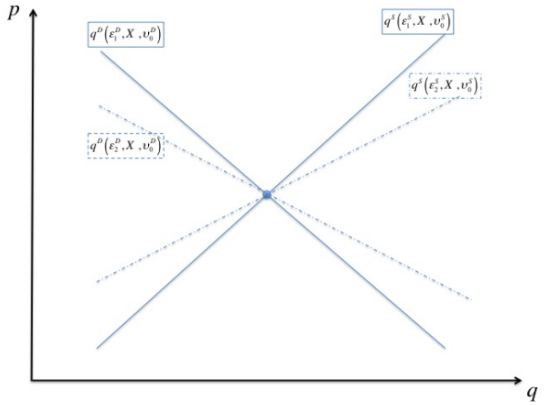
The Model: Supply and Demand

- Quantity traded and price are equilibrium outcomes from a system of simultaneous equations:

$$q_i^S = \epsilon^S p_i + \Gamma^S X_i + \nu_i^S$$
$$q_i^D = \epsilon^D p_i + \Gamma^D X_i + \nu_i^D$$

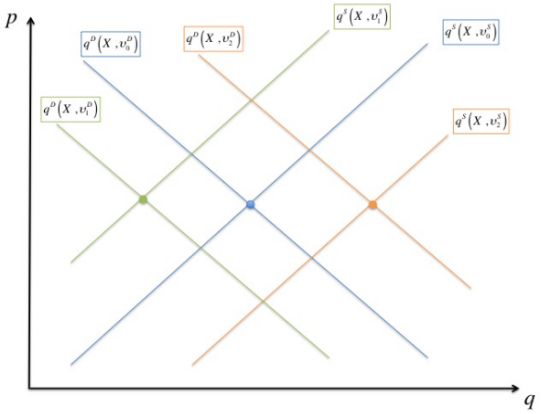
- Where:
  - i indexes different markets, S indexes supply, D indexes demand
  - q is log quantity, p is log price
  - X is a vector of (pre-determined) observable determinants of demand and supply (including a constant term)
  - $\left\{\nu^S, \nu^D\right\}$ are unobservable determinants of supply and demand.
- Target parameters: ϵ^S and $-\epsilon^D$

We only observe the equilibrium, not supply/demand



Solid and dashed lines represent two different supply/demand systems with different elasticities $\epsilon_1^D \neq \epsilon_2^D$ and $\epsilon_1^S \neq \epsilon_2^S$ yet observed equilibrium can be rationalized by both systems

Endogeneity



Endogeneity – equilibria across multiple markets $i \in \{1, 2, 3\}$ do not trace out either supply or demand

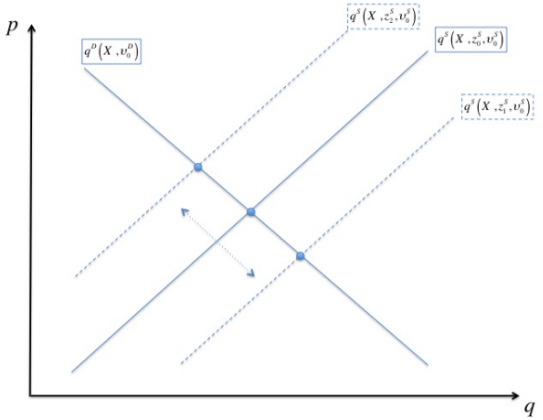
Exclusion Restrictions - Supply shifter

- Assume that we observe a variable (Z_i^S) that enters the supply equation but is excluded from the demand equation:

$$q_i^S = \epsilon^S p_i + \Gamma^S X_i + \theta^S Z_i^S + \nu_i^S$$
$$q_i^D = \epsilon^D p_i + \Gamma^D X_i + \nu_i^D$$

- We further assume:
  - $\theta^S \neq 0$ so that quantity supplied is a nontrivial function of Z_t^S
  - $Z_i^S \perp \nu_i^S, \nu_i^D | X_i$

Exclusion Restrictions - Supply shifter



Using variation in Z_i^S identifies the elasticity of demand by shifting supply along the demand curve.

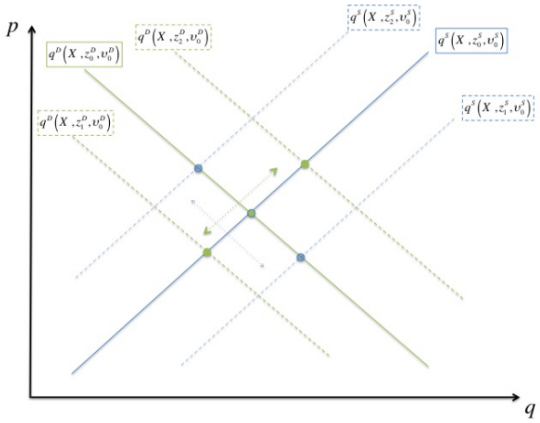
Exclusion Restrictions - Supply and Demand shifters

- Assume that in addition to the supply shifter (Z_i^S), we observe a variable (Z_i^D) that enters the demand equation but is excluded from the supply equation:

$$q_i^S = \epsilon^S p_i + \Gamma^S X_i + \theta^S Z_i^S + \nu_i^S$$
$$q_i^D = \epsilon^D p_i + \Gamma^D X_i + \theta^D Z_i^D + \nu_i^D$$

- We further assume:
  - $\theta^D \neq 0$ so that quantity demanded is a nontrivial function of Z_t^D
  - $Z_i^D \perp \nu_i^S, \nu_i^D | X_i$

Exclusion Restrictions - Supply and Demand shifters



Variation in Z_i^D (holding Z_i^S constant) identifies the elasticity of supply.
Variation in Z_i^S (holding Z_i^D constant) identifies the elasticity of demand.

Supply and Demand shifters - Reduced Form

- Solving equations for the equilibrium quantity and price on each market i, we obtain:

$$q_i = \frac{\epsilon^S \Gamma^D - \epsilon^D \Gamma^S}{\epsilon^S - \epsilon^D} X_i + \frac{\epsilon^S \theta^D Z_i^D - \epsilon^D \theta^S Z_i^S}{\epsilon^S - \epsilon^D} + \frac{\epsilon^S \nu_i^D - \epsilon^D \nu_i^S}{\epsilon^S - \epsilon^D}$$

$$p_i = \frac{\Gamma^D - \Gamma^S}{\epsilon^S - \epsilon^D} X_i + \frac{\theta^D Z_i^D - \theta^S Z_i^S}{\epsilon^S - \epsilon^D} + \frac{\nu_i^D - \nu_i^S}{\epsilon^S - \epsilon^D}$$

- Denote by q^* and p^* the residual variation in q and p after partialling out variation in X_i.
- Note: $q_i^* = \frac{\epsilon^S \theta^D Z_i^D - \epsilon^D \theta^S Z_i^S}{\epsilon^S - \epsilon^D} + \frac{\epsilon^S \nu_i^D - \epsilon^D \nu_i^S}{\epsilon^S - \epsilon^D}$ and $p_i^* = \frac{\theta^D Z_i^D - \theta^S Z_i^S}{\epsilon^S - \epsilon^D} + \frac{\nu_i^D - \nu_i^S}{\epsilon^S - \epsilon^D}$

IV estimates

$$\beta^{IV,D} = \frac{\text{Cov}\left(q_i^*, Z_i^S\right)}{\text{Cov}\left(p_i^*, Z_i^S\right)} = \frac{-\epsilon^D \theta^S}{-\theta^S} = \epsilon^D$$

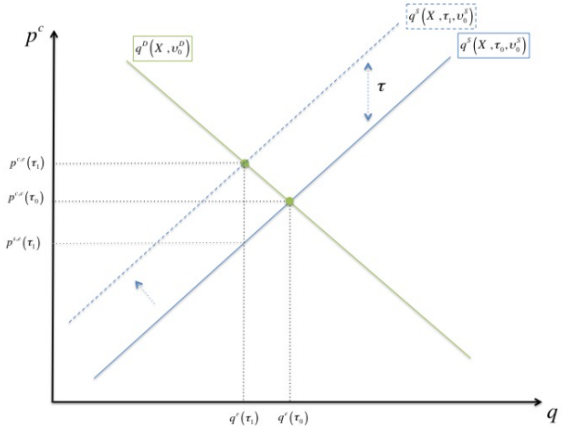
$$\beta^{IV,S} = \frac{\text{Cov}\left(q_i^*, Z_i^D\right)}{\text{Cov}\left(p_i^*, Z_i^D\right)} = \frac{\epsilon^S \theta^D}{-\theta^D} = \epsilon^S$$

- IV recovers the elasticities. In general, we need one instrument for each elasticity.
- An interesting exception: When tax rate is an instrument $\Rightarrow$ a single instrument (tax rate) recovers both elasticities (Gavrilova, Zoutman and Hopland 2018)

Using tax rates as an instrument

- Assume there is an ad valorem tax rate t_i imposed on producers. We define $\tau_i = \log(1 + t_i)$.
- We also denote by p_i^c the price paid by consumers and by $p_i^s = p_i^c - \tau_i$ the price received by suppliers.
- We assume $\tau_i \perp \nu_i^S, \nu_i^D | X_i$
- Because the tax is on producers, it does not enter the demand equation $\Rightarrow \epsilon^D$ is identified via standard exclusion restriction.
- Economic theory generates an additional exclusion restriction: Ramsey Exclusion Restriction (see GZH 2018)

Identification of Demand



The tax is a "supply shifter" – it allows identification of ϵ^D

Tax Rate as an Instrument

- The system of equation becomes:

$$q_i^D = \epsilon^D p_i^c + \Gamma^D X_i + \nu_i^D$$
$$q_i^S = \underbrace{\epsilon^S p_i^c + \underbrace{\theta^S Z_i^S}_{=-\epsilon^S \tau_i}}_{=\epsilon^S (p_i^c - \tau_i)} + \Gamma^S X_i + \nu_i^S$$

- Note: We impose an additional restriction – extremely common in public finance – that suppliers respond to the tax the same way they would respond to a cost shock ($\theta^S = -\epsilon^S$). This directly follows from an assumption of profit maximization.

Tax Rate as an Instrument - Reduced Form

- Solving the previous system of equations for the equilibrium quantity and price on each market i, we obtain:

$$q_i = \frac{\epsilon^S \Gamma^D - \epsilon^D \Gamma^S}{\epsilon^S - \epsilon^D} X_i + \frac{\epsilon^S \epsilon^D}{\epsilon^S - \epsilon^D} \tau_i + \frac{\epsilon^S \nu_i^D - \epsilon^D \nu_i^S}{\epsilon^S - \epsilon^D}$$

$$p_i^c = \frac{\Gamma^D - \Gamma^S}{\epsilon^S - \epsilon^D} X_i + \frac{\epsilon^S}{\epsilon^S - \epsilon^D} \tau_i + \frac{\nu_i^D - \nu_i^S}{\epsilon^S - \epsilon^D}$$

- Denote by q^* and p^{c*} the residual variation in q and p^c after partialling out variation in X_i.

Tax Rate as an Instrument - IV estimate

$$\beta_{\tau}^{IV,D} = \frac{\text{Cov}\left(q_i^*, \tau_i\right)}{\text{Cov}\left(p_i^{c*}, \tau_i\right)} = \epsilon^D$$

- This directly follows from slide 43 and the fact that the tax is excluded from the demand equation (Standard Exclusion Restriction)
- Can we identify more than just ϵ^D?
- Yes, it is the role of the additional restriction that suppliers respond to the tax the same way they would to an increase in marginal cost ($\theta^S = -\epsilon^S$). $\Rightarrow$ Key implication is that the passthrough of the tax (to consumers) is $\frac{dp^c}{d\tau} = \frac{\epsilon^S}{\epsilon^S - \epsilon^D}$

Tax Rate as an Instrument - Identifying ϵ^S

- Because 1) ϵ^D is identified and 2) we can estimate the passthrough $\frac{dp^c}{d\tau}$, which is a function of two elasticities, we can recover ϵ^S.
- GZH 2018 recommend using the following IV estimator:

$$\beta_{\tau}^{IV,S} = \frac{\text{Cov}\left(q_i^*, \tau_i\right)}{\text{Cov}\left(p_i^{s*}, \tau_i\right)} = \epsilon^S$$

Interpreting regression results with optimizing agents

Outline: Optimizing agents and regression results

1. Moneyball
2. The value of police services (example of "adding economics" to a nice reduced-form paper)

Firm Optimization

- Inputs X with prices w
- Output $Y = F(X)$ with value V
- Firm optimization gives us the following marginal condition:

$$VF_i = w_i$$

- Value of output $\times$ Marginal product $=$ marginal cost of input

Implications

$$VF_i = w_i$$
$$V = \frac{w_i}{F_i}$$

- Cost per marginal unit of output must be the same for all factors

Application #1: Moneyball

Baseball team

- Output is wins
- V is value of a win
- X are player attributes
- w are prices of player attributes

Table 1

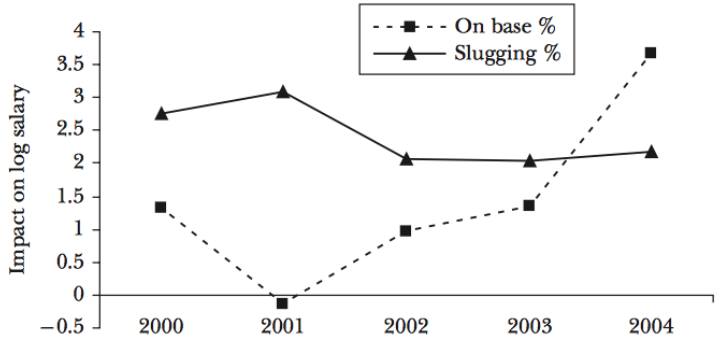
The Impact of On-Base and Slugging Percentage on Winning

	Model			
	1	*2*	*3*	*4*
Constant	0.508	0.612	0.502	0.500
	(0.114)	(0.073)	(0.099)	(0.005)
On-Base	3.294		2.141	2.032
	(0.221)		(0.296)	(0.183)
On-Base against	−3.317		−1.892	−2.032[R]
	(0.196)		(0.291)	
Slugging		1.731	0.802	0.900
		(0.122)	(0.149)	(0.106)
Slugging against		−1.999	−1.005	−0.900[R]
		(0.112)	(0.152)	
Number of observations	150	150	150	150
R^2	.825	.787	.885	.884

Hypothesis test of model 4, H^0: On-Base = Slugging
$F(1, 147) = 16.74$, p-value = 0.0001

Source: Hakes and Sauer (2006). Example from Jesse Shapiro.

Labor Market Returns to On-Base and Slugging Percentage Over Time



Source: Hakes and Sauer (2006). Example from Jesse Shapiro.

#2 Optimization and the Value of Police Services

Mello (2018) "More COPS, Less Crime"

1. **Policy relevant question:** What is the effect of additional police on local criminal activity?

2. **Nice variation:** ARRA police funding index increased COPS in some locations

3. **Interesting Results:**
   - Average grant increased police by 0.7 per 10,000 residents (or 6% increase in police)
   - Each officer reduces 4.3 violent crimes and 15.4 property crimes
   - Benefit of $35 per resident vs $29 cost

Economic Questions

This paper focuses on reduced-form effects, and provides a good example that shows how we could think about optimization and the following questions:

1. What is the value of a marginal police officer?
2. How many police officers should we hire?
3. How should they be allocated? Should police focus more on violent crime?

Economic Framework

Local governments produce safety

$$y = f(L)$$

- y are units of safety
- L is number of police officers

Local governments maximize:

$$\max_L pf(L) - wL$$

- p is the value of a unit of safety
- w is wage of police officers

Should we hire more police officers?

FOC

$$pf'(L) = w$$

- $pf'(L)$ is the marginal value of safety
- w is the marginal cost of safety

Estimates suggest that $pf'(L) > w$

- Estimate of marginal benefit from Table 2 is \$35.2 per 10K residents
- Direct cost is roughly \$29 per 10K residents

$\Rightarrow L < L^*$

Keep hiring police until these are equal! (But also need to account for cost of raising funds).

Economic framework with two types of crimes

Two types of safety y:

- safety from violent crime y_1
- safety from property crime y_2

Local governments maximize:

$$\max_{L_1, L_2} p_1 f(L_1) + p_2 g(L_2) - w(L_1 + L_2)$$

- p_1 is the value of a unit of safety from violent crime
- p_2 is the value of a unit of safety from property crime
- L_1 is number of police officers allocated to reducing violent crime
- L_2 is number of police officers allocated to reducing property crime
- Note main outcome in paper is approx $\$68,000 \times y_1 + \$4,000 \times y_2$

Optimal policing of violent crime?

FOC for violent crimes:

$$p_1 f'(L_1) = w$$

- p_1 is approx \$68,000
- $f'(L_1) = 4.3$, i.e., hiring one more officer reduces # of violent crimes by 4.3
- Marginal benefit is $4.3 \times \$68,000 \approx \$292,400$

If local governments are optimizing, then

$$\underbrace{f'(L_1)}_{\text{Marginal product}} = \frac{w}{\$68,000}$$

Optimal policing of property crime?

FOC for property crimes:

$$p_2 g'(L_2) = w$$

- p_2 is approx \$4,000
- $f'(L_2) = 15.4$, i.e., hiring one more officer reduces # of property crimes by 15.4
- Marginal benefit is $15.4 \times \$4,000 \approx \$61,600$

If local governments are optimizing, then

$$\underbrace{g'(L_2)}_{\text{Marginal product}} = \frac{w}{\$4,000}$$

Should police focus more on violent crime reduction?

FOCs for violent and property crimes:

$$p_1 f'(L_1) = w$$
$$p_2 g'(L_2) = w$$

But $p_1 f'(L_1) = \$292,000 > p_2 g'(L_1) = \$62,000$

If local governments are optimizing, then

$$\underbrace{\$68,000}_{\text{Value of output}} = \underbrace{\frac{w}{f'(L_1)}}_{\text{cost per marginal unit of output}}$$

$$\underbrace{\$4,000}_{\text{Value of output}} = \underbrace{\frac{w}{g'(L_2)}}_{\text{cost per marginal unit of output}}$$

Seems like police should focus more on violent crime given p_1 and p_2

FOC

$$p_c f'(L_c) = w_c$$

- $p_c f'(L_c)$ is the marginal value of safety in location c
- w_c is the marginal cost of safety in location c

Would be interesting to analyze heterogeneity based on variation in

- Initial force size L_c varies (so can trace out $f'(L_c)$)
- Local cost of safety w_c
- Local value for safety p_c can vary

Demand for safety

Where do the estimates of p_1 and p_2 come from?

Resident utility depends on level of safety and other consumption:

$$max_{x,y} U(x, y) \;\; s.t. \;\; p_y y + p_x x = M$$

- y is units of safety
- x is a composite of other goods
- M is income (and λ is MU of income)

FOC: $\frac{\partial U}{\partial y} = \lambda p_y$ suggests that:

- Marginal utility of safety depends on level of safety (so level of L)
- Value of safety $\frac{\frac{\partial U}{\partial y}}{\lambda}$ is increasing in income (since λ is decreasing in M)
- Thus, p_y should depend on level of L and local incomes

Other thoughts

Estimates are interesting inputs for welfare analysis of an important non-traded good

1. **Welfare analysis** Could think about effective cost w that includes overhead and MCPF that would rationalize current hiring levels
2. **Time allocation** Could weigh into debates about how police spend their time (violent crime vs property crime)
3. **Supply side** Could learn more about production function of safety $f(L)$ and $g(L)$
4. **Demand side** Could think more about value of unit of safety and the efficiency vs equity considerations of how police spending is allocated
5. **Evaluating current police spending** What social welfare function and/or cost of public funds are consistent with the level and allocation?

Potential Outcomes and Selection

Causality, Potential Outcomes, and Selection

1. Potential Outcomes
2. Selection Bias
3. Example: return to selective colleges

Motivation

- Counterfactual questions are everywhere
  - What would happen if a job training program were expanded
  - What would happen to prices/welfare if two firms marged?
  - What would different monetary policy do to real output?
  - What effect would this medication have on heart disease
  - What will happen to global temps if emissions decrease?
- Causal inference
  - Thinking about a counterfactual requires thinking about **causality**
  - Theory alone might (*might*) tell us the direction of causality
    - Even when it does, it will rarely tell us the magnitude
  - **Causal inference** uses data to address counterfactuals

Potential Outcome Notation

- Also known as the Neyman-Fisher-Roy-Quandt-Rubin causal model
- $\mathcal{D}$ is a mutually exclusive and exhaustive set of states ("treatments") e.g. training/no training $\mathcal{D} = \{0, 1\}$, prices $\mathcal{D} = [0, +\infty)$, etc.
- For each $d \in \mathcal{D}$, there is a **potential outcome** Y_d (a random variable)
- Y_d is what *would have* happened if the state were d
- We observe the actual state, a random variable $D \in \mathcal{D}$
- We also observe an outcome Y, related to potential outcomes as

$$Y = \sum_{d \in \mathcal{D}} Y_d \mathbf{1}[D = d] = Y_D$$

$$\text{binary case:} \quad Y = DY_1 + (1 - D)\, Y_0(\text{"switching regression"})$$

- $Y = Y_D$ is observed, but Y_d for $d \neq D$ are unobserved

The fundamental problem of causal inference (Holland, 1986)

$Y = DY^1 + (1 - D)\, Y^0$

Causal inference requires assumptions/restrictions on the "missing" potential outcomes

i	Y_i^0	Y_i^1	D_i
1	32	20	1
2	200	4	0
3	0	0	0
4	80	80	1
5	53	88	0
6	0	90	1

What Do We Want to Measure?

- We are interested in counterfactuals, Y_d for $d \neq D$
- These variables capture the "what if" aspect of causality
- Since they are random variables, they can be summarized in many ways
- That is, there are many possible **parameters of interest**

Example: Program Evaluation

- Suppose $d \in \{0, 1\}$ indicates participation in a job training program
- Y is a scalar labor market outcome such as earnings
- If $D = 1$ we observe Y_1 (but not Y_0) and if $D = 0$ we observe Y_0
- There are many possible questions one could ask:
  - What would be average earnings if everyone were trained, i.e. $\mathbb{E}[Y_1]$?
  - What is the average effect of the program, i.e. $\mathbb{E}[Y_1 - Y_0]$?
  - What about only for those who are trained, i.e. $\mathbb{E}[Y_1 - Y_0|D = 1]$?
- What is useful depends on what question we want to answer!

Latent Variable Notation

- Many empirical models in economics look like a special case of:

$$Y = g(D, U),$$

  where g is a function and U are unobservable variables
- A causal interpretation of this model is implicitly saying:

$$Y_d = g(d, U) \quad \text{for every } d \in \mathcal{D}$$

- This could impose assumptions, depending on what g and U are

Going from potential outcomes to parameters of interest

Three considerations/ pillars of econometrics:

- Identification (what can be learned?)
- Estimation (how best to learn it?)
- Inference (how certain am I?)

Identification

- The parameter of interest is a function of the unobservables. In the PO model, this is some function of $\{Y_d\}_{d \in \mathcal{D}}$
- What could we learn about this function from the observables, (Y, D)?
- This is the question of **identification**
- Identification can be seen as *the link between data and theory*
- By **theory** here I mean the way in which we believe the world works
- This can include what we think of as "economic theory"
- But it need not be so formal – a synonym would be **assumptions**
- Theory is often ambiguous – otherwise we wouldn't need data
- Identification *always* involves a (possibly simple) **model**
- This model encodes our assumptions (theory)

Identification vs. Statistical Inference

- In practice, we only see a finite **sample** of the observables $\{\{Y_i, D_i\}\}_{i=1}^n$
- From this we know the **sample distribution**
- However, we don't know the **population distribution** of (Y, D)
- **Statistical inference** is using the sample to learn about the population
- It is useful to separate identification from statistical inference:

$$\text{sample} \underbrace{\rightarrow}_{\text{statistical inference}} \text{population} \underbrace{\rightarrow}_{\text{identification}} \text{unobserved parameters}$$

- The second arrow is logically the first thing to consider
- Can't recover a parameter when we know the population distribution? Then you also couldn't recover it with the sample distribution!
- Informally, identification is sometimes seen as having "infinite data"

Identification is Prior to Statistical Inference

- Return to the example of job training and earnings
- Suppose we care about the average effect of the program on participants:

$$ATT = \mathbb{E}[Y_1 - Y_0 | D = 1] = \underbrace{\mathbb{E}[Y|D=1]}_{\text{fnc. of pop. dist.}} - \underbrace{\mathbb{E}[Y_0|D=1]}_{\text{fnc. of unobs.}}$$

- An important ingredient in a decision to continue or end the program
- The first term is a function of the population distribution
- Using the sample to understand this from data is the domain of statistics
- The question of identification is about the second term
- What can we say about $\mathbb{E}[Y_0|D=1]$ under different assumptions?
- Must answer this question *before* we can construct an estimate of ATT

Selection Bias

Selection

- There is selection into the treatment state D if

$$\underbrace{Y_d|D=d}_{\text{observable}} \text{ is distributed differently from } \underbrace{Y_d|D=d'}_{\text{unobserved}} \text{ for } d' \neq d$$

- This is not the case under the random assignment assumption
- Expected to occur if agents choose D with knowledge of $\{Y_d\}_{d\in\mathcal{D}}$
- Selection is common
  - Particularly concerning if you are trained in neoclassical economics
  - Agents choose a job training program ($D \in \{0,1\}$) to max utility
  - Utility will incorporate expected future earnings (Y_0, Y_1)
  - Agents who choose job training might do so because of low Y_0
  - Data typically supports this story ("Ashenfelter's (1978) dip")
  - Alternatively, might choose $D=0$ because of high Y_0

Selection Bias

- Consider the simple treatment/control mean contrast under selection
- This contrast would be the ATE under random assignment
- Decompose the contrast into a causal effect and selection bias:

$$\begin{aligned}
&\mathbb{E}[Y|D=1] - \mathbb{E}[Y|D=0] \\
&= \underbrace{\mathbb{E}[Y_1|D=1] - \mathbb{E}[Y_0|D=1]}_{\text{ATT}} + \underbrace{\mathbb{E}[Y_0|D=1] - \mathbb{E}[Y_0|D=0]}_{\text{selection bias}} \\
&= \underbrace{\mathbb{E}[Y_1|D=0] - \mathbb{E}[Y_0|D=0]}_{\text{ATU}} + \underbrace{\mathbb{E}[Y_1|D=1] - \mathbb{E}[Y_1|D=0]}_{\text{selection bias}}
\end{aligned}$$

- First term is causal effect for those who were treated/untreated
  - Under random assignment would have ATT = ATU = ATE
- Second term is how the treated would have been different anyway
  - Under random assignment this would be 0
- The first expression is more natural if thinking of $D = 0$ as baseline

Selection on Observables

- A simple relaxation of random assignment is **selection on observables**
- Suppose that we observe (Y, D, X) where X are covariates
- The selection on observables assumption is that

$$\{Y_d\}_{d \in \mathcal{D}} \perp D | X$$

- Says: Conditional on X, treatment is *as-good-as* randomly assigned
- Other terms: **unconfoundedness, ignorable treatment assignment**
- Underlies causal interpretations of linear regression
  - We will look into this connection more later

Pre-Determined Covariates

Advice - only condition on predetermined observables

- For selection on observables to be plausible, X should be **predetermined**
- In particular, D should not have a causal effect on X
- Usually this really is a temporal issue (measured before vs. after D)
- Intuition is clear – we want to condition on selection *into* treatment

Simple but trivial example

- Suppose we accidentally included Y as part of X
- Then clearly we aren't going to have $(Y_0, Y_1) \perp D|X$

Less trivial examples

- Don't include earnings 1 year after the program in X
- Don't include employment after the program in X
- Don't include marital status after the program in X

Imputation Estimator for the ATE

- Suppose $D \in \{0, 1\}$ and recall the first point identification result:

$$\text{ATE} = \underbrace{E}_{\text{over X}} \left[\mathbb{E}\left[Y|D=1, X\right] - \mathbb{E}\left[Y|D=0, X\right]\right]$$

- Suppose we have an i.i.d. sample of data $\{(Y_i, D_i, X_i)\}_{i=1}^N$
- A natural **imputation estimator** is given by:

$$\widehat{\text{ATE}} = \frac{1}{N} \sum_{i=1}^{N} \hat{\mu}_1\left(X_i\right) - \hat{\mu}_0\left(X_i\right)$$

where:

- $\hat{\mu}_1\left(x\right)$ is an estimator of $\mu_1\left(x\right) \equiv \mathbb{E}\left[Y|D=1, X=x\right]$
- $\hat{\mu}_0\left(x\right)$ is an estimator of $\mu_0\left(x\right) \equiv \mathbb{E}\left[Y|D=0, X=x\right]$

- Estimate conditional means, then take the sample analog

Imputation Estimator for the ATT

- The ATT is actually easier to estimate, since

$$\begin{aligned}\text{ATT} &= \mathbb{E}\left[Y_1 - Y_0 | D = 1\right] \\ &= \mathbb{E}\left[Y | D = 1\right] - \mathbb{E}\left[\mu_0\left(X\right) | D = 1\right] = \mathbb{E}\left[Y - \mu_0\left(X\right) | D = 1\right]\end{aligned}$$

- An imputation estimator here would be

$$\widehat{\text{ATT}} = \frac{1}{N_1} \sum_{i:D_i=1} Y_i - \hat{\mu}_0\left(X_i\right) \quad \text{with } N_1 = \sum_{i=1}^{N} D_i$$

- Imputation of the first term is simplified (sample average)
- Using control group to impute control outcomes for treated group
  - No need to do the opposite, so don't need to estimate $\hat{\mu}_1$
- Notice that the sample average weights X according to the treated group
  - So, same issue remains (how to estimate μ_0), but a bit simpler

A Slightly More Complicated Imputation Estimator

- Using insight from the ATT imputation, we could estimate ATE with

$$\begin{aligned}\widehat{\text{ATE}} &= \frac{1}{N}\sum_{i=1}^{N} D_i\left(Y_i - \hat{\mu}_0\left(X_i\right)\right) + \left(1 - D_i\right)\left(\hat{\mu}_1\left(X_i\right) - Y_i\right) \\ &= \underbrace{\frac{N_1}{N}}_{\approx \mathbb{P}[D=1]} \underbrace{\left[\frac{1}{N_1}\sum_{i:D_i=1} Y_i - \hat{\mu}_0\left(X_i\right)\right]}_{\equiv \widehat{\text{ATT}}} + \underbrace{\frac{N_0}{N}}_{\approx \mathbb{P}[D=0]} \underbrace{\left[\frac{1}{N_0}\sum_{i:D_i=0} \hat{\mu}_1\left(X_i\right) - Y_i\right]}_{\equiv \widehat{\text{ATU}}}\end{aligned}$$

- Makes it clear that we know $Y_i = Y_{di}$ for $D_i = d$
- No need to impute Y_{1i} for $D_i = 1$ or Y_{0i} for $D_i = 0$
- The two forms of imputation will be numerically identical if:

$$\frac{1}{N_d}\sum_{i:D_i=d} \hat{\mu}_d\left(X_i\right) = \frac{1}{N_d}\sum_{i:D_i=d} Y_i \qquad \text{– often the case, e.g. linear regression}$$

- In either form, primary problem is estimating $\hat{\mu}_0$ and/or $\hat{\mu}_1$

Parametric Estimators

- We could also just use the good ol' fashioned parametric model, e.g.,

$$\mu_d(x) = \alpha_d + \beta_d' x$$

- Estimate α_d, β_d by regressing Y on X among $D = d$ subpopulations then

$$\begin{aligned}\widehat{\text{ATE}} &= \frac{1}{N}\sum_{i=1}^{N}\left[\widehat{\alpha_1} + \widehat{\beta_1}'X_i - \widehat{\alpha_0} - \widehat{\beta_0}'X_i\right] \\ &= \underbrace{\bar{Y}_1 - \bar{Y}_0}_{\text{Naive contrast}} + \underbrace{\left(\frac{N_1}{N}\hat{\beta}_0 + \frac{N_0}{N}\hat{\beta}_1\right)'(\bar{X}_0 - \bar{X}_1)}_{\text{regression adjustment}}\end{aligned}$$

- which can be shown after noting $\widehat{\alpha}_d = \bar{Y}_d - \widehat{\beta}_d'\bar{X}_d$ and rearranging terms
- Widely used approach, but considered poor taste by many
- Concern about functional forms driving results via extrapolation

Connection to Linear Regression

- Nevertheless, linear regression is the most widely used approach
- In fact, by far the most widely used specification looks like this:

$$\mu_d(x) = \alpha_d + x'\beta \equiv \alpha_0 + (\alpha_1 - \alpha_0)\, d + x'\beta$$

- Under selection on observables, this implies that

$$\mathbb{E}[Y|D, X] = D\mu_1(X) + (1 - D)\,\mu_0(X) = \alpha_0 + (\alpha_1 - \alpha_0)\, D + X'\beta$$

- This is restrictive and implies **constant treatment effects**:

$$\mathbb{E}[Y_1 - Y_0|X = x] = \alpha_1 - \alpha_0 \qquad \text{does not depend on } x$$

- In contrast, the specification on the previous slide had

$$\mathbb{E}[Y_1 - Y_0|X = x] = \alpha_1 - \alpha_0 + (\beta_1 - \beta_0)'x \qquad \text{still depends on } x$$

- Can be implemented as a single regression of Y on 1, D, X and DX

Example: returns to selective colleges

Dale and Krueger (2002)

Question, empirical setting

- The impact of selective colleges on labor market outcomes
- Compare outcomes of those who attend selective vs. non-selective?
  - *Clearly* unlikely to be a causal effect
- Even after conditioning on observables such as GPA
  - Unobservables play a large role in college admissions

Methodology

- Match (condition) on the *set* of colleges to which amitted
- Key variable which they use to argue causality
- Assumption is selection on a (carefully-chosen) observable
- Use linear regression (without interactions)

The Empirical Strategy

Stylized idea

- UPenn (selective) vs. Penn State (less selective)
- Condition on students with similar observables
  - GPA, sex, race, athlete, high school rank
- Find students who were only *admitted* to UPenn and Penn State
- Compare log wages of those who *attended* UPenn vs. Penn State

Practical challenges

- Defining selectivity $\rightarrow$ they use average SAT
- Implementing the grouping above more generally
- Main estimates use average SAT instead of school identity...

Grouping Scheme

TABLE I

ILLUSTRATION OF HOW MATCHED-APPLICANT GROUPS WERE CONSTRUCTED

		Student applications to college							
		Application 1		Application 2		Application 3		Application 4	
Student	Matched-applicant group	School average SAT	School admissions decision	School average SAT	School admissions decision	School average SAT	School admissions decision	School average SAT	School admissions decision
Student A	1	1280	Reject	1226	Accept*	1215	Accept	na	na
Student B	1	1280	Reject	1226	Accept	1215	Accept*	na	na
Student C	2	1360	Accept	1310	Reject	1270	Accept*	1155	Accept
Student D	2	1355	Accept	1316	Reject	1270	Accept*	1160	Accept
Student E	2	1370	Accept*	1316	Reject	1260	Accept	1150	Accept
Student F	Excluded	1180	Accept*	na	na	na	na	na	na
Student G	Excluded	1180	Accept*	na	na	na	na	na	na
Student H	3	1360	Accept	1308	Accept*	1260	Accept	1160	Accept
Student I	3	1370	Accept*	1311	Accept	1255	Accept	1155	Accept
Student J	3	1350	Accept	1316	Accept*	1265	Accept	1155	Accept
Student K	4	1245	Reject	1217	Reject	1180	Accept*	na	na
Student L	4	1235	Reject	1209	Reject	1180	Accept*	na	na
Student M	5	1140	Accept	1055	Accept*	na	na	na	na
Student N	5	1145	Accept*	1060	Accept	na	na	na	na
Student O	No match	1370	Reject	1038	Accept*	na	na	na	na

* Denotes school attended.
na = did not report submitting application.
The data shown on this table represent hypothetical students. Students F and G would be excluded from the matched-applicant subsample because they applied to only one school (the school they attended). Student O would be excluded because no other student applied to an equivalent set of institutions.

- Also report estimates that group using school identities directly
- And estimates that group using *Barron*'s selectivity ranking (1-5)

Main Results

TABLE III
LOG EARNINGS REGRESSIONS USING COLLEGE AND BEYOND SURVEY, SAMPLE OF MALE AND FEMALE FULL-TIME WORKERS

	Model					
	Basic model: no selection controls		Matched-applicant model	Alternative matched-applicant models		
	Full sample	Restricted sample	Similar school-SAT matches*	Exact school-SAT matches**	*Barron's* matches***	Self-revelation model
Variable	1	2	3	4	5	6
School-average SAT score/100	0.076	0.082	−0.016	−0.106	0.004	−0.001
	(0.016)	(0.014)	(0.022)	(0.036)	(0.016)	(0.018)
Predicted log(parental income)	0.187	0.190	0.163	0.232	0.154	0.161
	(0.024)	(0.033)	(0.033)	(0.079)	(0.028)	(0.025)
Own SAT score/100	0.018	0.006	−0.011	0.003	−0.005	0.009
	(0.006)	(0.007)	(0.007)	(0.014)	(0.005)	(0.006)
Female	−0.403	−0.410	−0.395	−0.476	−0.400	−0.396
	(0.015)	(0.018)	(0.024)	(0.049)	(0.017)	(0.014)
Black	−0.023	−0.026	−0.057	−0.028	−0.057	−0.034
	(0.025)	(0.052)	(0.052)	(0.049)	(0.029)	(0.025)

- Columns 1-2: Raw coefficient is positive and large
- Columns 3-5: Match on schools admitted to
- Column 6: Match on schools *applied* to – they argue upward bias here

Possible Explanations and Critiques

Explanations

1. More likely to enter academia/public sector at a more selective college
   - They check this by including controls for occupation
   - Not a good strategy – occupation is an *outcome* (not predetermined)
2. "Big fish, small pond" phenomenon
   - Supported by changing outcome variable to college class rank
   - Heterogeneity by parental income – find benefis for low income

Critiques

- Not interacting groups with SAT imposes constant effects
  - Selectivity has the same labor market impact for any choice set?
- SAT is too coarse a measure of school quality
  - Replacing SAT by dummies for schools they strongly reject 0 effect
- Students might still be sorting on unobservables within groups

Roy Model and Selection Corrections

Roy Model of Self-Selection

- Roy (1951) sought to understand the influence of occupational choice on the observed distribution of earnings
- Consider individuals indexed by i choosing a binary variable $D_i \in \{0, 1\}$, e.g. hunting vs. fishing
- Y_{i1} and Y_{i0} are i's potential outcomes (e.g. earnings) with and without treatment
- Realized outcome is $Y_i = Y_{i0} + (Y_{i1} - Y_{i0})D_i$
- Pure Roy (1951) model: Individuals want to maximize Y_i, so choose the alternative with the best potential outcome:

$$D_i = 1\{Y_{i1} > Y_{i0}\}$$

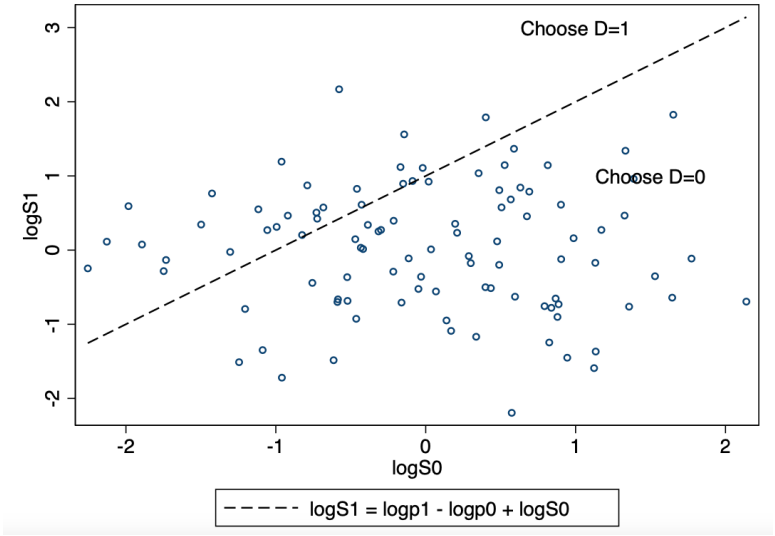
Roy Model of Self-Selection

- Questions of interest:
  - Will the best hunters hunt?
  - Will the best fishermen/women fish?
- Suppose potential outcomes are given by

$$Y_{id} = p_d S_{id}, \ d \in \{0, 1\}$$

- S_{id} is skill in occupation d, and p_d is price of output
- A worker who is indifferent between the two occupations satisfies

$$\log S_{i1} = \log p_0 - \log p_1 + \log S_{i0}$$

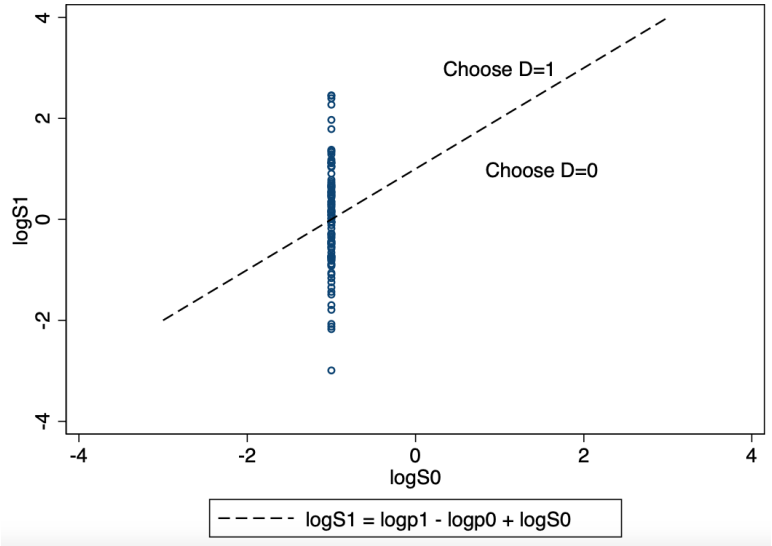

Choose D=1
Choose D=0
logS1
logS0
3
2
1
0
-1
-2
-2
-1
0
1
2
logS1 = logp1 - logp0 + logS0

Roy Model of Self-Selection

- Suppose there is no variation in the non-treated outcome, so $S_{i0} = \bar{S}_0 \ \forall i$
- In this case the decision rule is

$$D_i = 1\left\{S_{i1} \geq \left(\frac{p_0}{p_1}\right)\bar{S}_0\right\}$$

- Those with the most skill in sector 1 choose $D_i = 1$
- Everyone with $D_i = 1$ earns more than anyone with $D_i = 0$


Choose D=1
Choose D=0
logS1
logS0
4
2
0
-2
-4
-4
-2
0
2
4
logS1 = logp1 - logp0 + logS0

Roy Model of Self-Selection

- Suppose we have perfect correlation between $\log S_{i0}$ and $\log S_{i1}$:

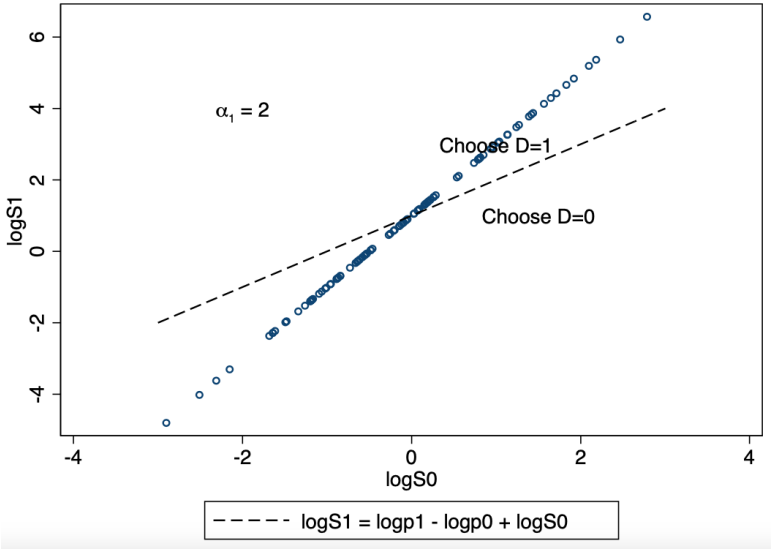
$$\log S_{i1} = \alpha_0 + \alpha_1 \log S_{i0}, \ \alpha_1 > 0$$

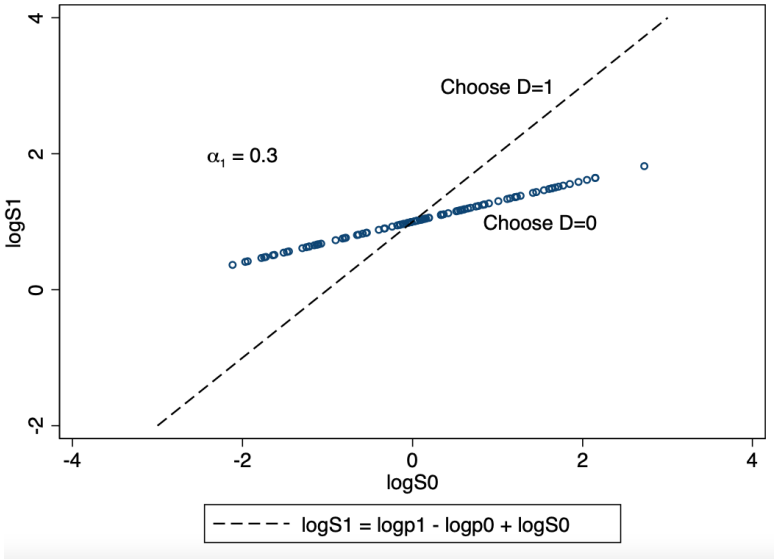
- This is a one-factor model
- Decision rule:

$$D_i = 1\{\alpha_0 + \alpha_1 \log S_{i0} \geq \log p_0 - \log p_1 + \log S_{i0}\}$$

$$= 1\{(\alpha_1 - 1) \log S_{i0} \geq \log p_0 - \log p_1 - \alpha_0\}$$

- Higher skilled choose $D_i = 1$ iff $\alpha_1 \geq 1$
- Note that $Var(\log S_{i1}) = \alpha_1^2 Var(\log S_{i0})$. Higher skilled choose the sector with higher variance of earnings


$\alpha_1 = 2$
Choose D=1
Choose D=0
logS1
logS0
6
4
2
0
-2
-4
-4
-2
0
2
4
logS1 = logp1 - logp0 + logS0


Choose D=1
α_1 = 0.3
Choose D=0
logS1
logS0
4
2
0
-2
-4
-2
0
2
4
logS1 = logp1 - logp0 + logS0

Roy Model: Parametric example

- Suppose $(\log S_{i1}, \log S_{i0})$ are bivariate normal:

$$(\log S_{i1}, \log S_{i0}) \sim N\left((\mu_1, \mu_0), \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_0 \\ \rho\sigma_1\sigma_0 & \sigma_0^2 \end{bmatrix}\right)$$

- Implies skill $\log S_{i1}$ and difference in log earnings $\log Y_{i1} - \log Y_{i0}$ are also bivariate normal:

$$(\log S_{i1}, \log Y_{i1} - \log Y_{i0}) \sim$$
$$N\left((\mu_1, \log p_1 + \mu_1 - \log p_0 - \mu_0), \begin{bmatrix} \sigma_1^2 & \sigma_1^2 - \rho\sigma_1\sigma_0 \\ \sigma_1^2 - \rho\sigma_1\sigma_0 & \sigma_1^2 + \sigma_0^2 - 2\rho\sigma_1\sigma_0 \end{bmatrix}\right)$$

- Average skill in sector 1 for those who choose sector 1:

$$E\left[\log S_{i1} | D_i = 1\right] = E\left[\log S_{i1} | \log Y_{i1} - \log Y_{i0} > 0\right]$$

$$= \mu_1 + \left(\frac{\sigma_1^2 - \rho\sigma_1\sigma_0}{\sigma_1^2 + \sigma_0^2 - 2\rho\sigma_1\sigma_0}\right) \times \frac{\phi\left(\frac{\log p_1 + \mu_1 - \log p_0 - \mu_0}{\sqrt{\sigma_1^2 + \sigma_0^2 - 2\rho\sigma_1\sigma_0}}\right)}{\Phi\left(\frac{\log p_1 + \mu_1 - \log p_0 - \mu_0}{\sqrt{\sigma_1^2 + \sigma_0^2 - 2\rho\sigma_1\sigma_0}}\right)}.$$

Roy Model: Parametric example

$$E\left[\log S_{i1} | D_i = 1\right] = \mu_1 + \left(\frac{\sigma_1^2 - \rho\sigma_1\sigma_0}{\sigma_1^2 + \sigma_0^2 - 2\rho\sigma_1\sigma_0}\right) \times \frac{\phi\left(\frac{\log p_1 + \mu_1 - \log p_0 - \mu_0}{\sqrt{\sigma_1^2 + \sigma_0^2 - 2\rho\sigma_1\sigma_0}}\right)}{\Phi\left(\frac{\log p_1 + \mu_1 - \log p_0 - \mu_0}{\sqrt{\sigma_1^2 + \sigma_0^2 - 2\rho\sigma_1\sigma_0}}\right)}.$$

- Those who choose sector 1 have above average skill in sector 1 iff $\sigma_1^2 - \rho\sigma_1\sigma_0 > 0$, or equivalently, $(\sigma_1/\sigma_0) > \rho$.

- Suppose you are interested in the benefit of College Education ($D = 1$) relative to not having College Education ($D = 0$)
- For each individual you observe realized wage:

$$Y = DY_1 + (1 - D)\, Y_0$$

- Where:

$$\begin{aligned} Y_1 &= X\beta_1 + U_1 \\ Y_0 &= X\beta_0 + U_0 \\ D &= \mathbf{1}\left(Y_1 > Y_0\right) \\ \begin{pmatrix} U_1 \\ U_0 \end{pmatrix} &\sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{pmatrix}\right) \end{aligned}$$

Using Roy model to think about returns to college

- Note:

$$U_0 - U_1 \sim \mathcal{N}\left(0, \sigma^2 + 1 - 2\rho\sigma\right)$$
$$\text{Cov}\left(U_1, U_0 - U_1\right) = \rho\sigma - \sigma^2$$
$$\text{Cov}\left(U_0, U_0 - U_1\right) = 1 - 2\rho\sigma$$

- Decision rule:

$$\begin{aligned} D &= \mathbf{1}\left(Y_1 > Y_0\right) \\ &= \mathbf{1}\left(X\beta_1 + U_1 > X\beta_0 + U_0\right) \\ &= \mathbf{1}\left(X\left(\beta_1 - \beta_0\right) > U_0 - U_1\right) \end{aligned}$$

- Implies:

$$\mathbb{P}\left(D = 1|X\right) = \mathbb{P}\left(X\left(\beta_1 - \beta_0\right) > U_0 - U_1\right) = \Phi\left(\frac{X\left(\beta_1 - \beta_0\right)}{\sqrt{\sigma^2 + 1 - 2\rho\sigma}}\right)$$

$$
\begin{aligned}
\text{ATE} &= \mathbb{E}\left(Y_1 - Y_0 | X\right) = X\left(\beta_1 - \beta_0\right) \\
\text{ATT} &= \mathbb{E}\left(Y_1 - Y_0 | X, D = 1\right) \\
&= \mathbb{E}\left(X\beta_1 + U_1 - X\beta_0 - U_0 | X, X\left(\beta_1 - \beta_0\right) > U_0 - U_1\right) \\
&= X\left(\beta_1 - \beta_0\right) - \mathbb{E}\left(U_0 - U_1 | X, U_0 - U_1 < X\left(\beta_1 - \beta_0\right)\right) \\
&= X\left(\beta_1 - \beta_0\right) + \underbrace{\sqrt{\sigma^2 + 1 - 2\rho\sigma}\frac{\phi\left(\frac{X(\beta_1-\beta_0)}{\sqrt{\sigma^2+1-2\rho\sigma}}\right)}{\Phi\left(\frac{X(\beta_1-\beta_0)}{\sqrt{\sigma^2+1-2\rho\sigma}}\right)}}_{>0}
\end{aligned}
$$

- Intuition: Those who select into college benefit from it

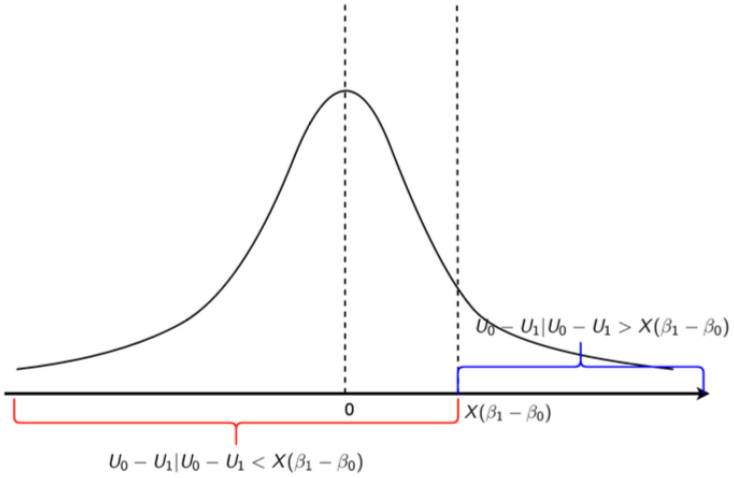
The basic Roy model, selection and target parameters

$$\begin{aligned}\text{ATU} &= X(\beta_1 - \beta_0) - \mathbb{E}(U_0 - U_1 | X, U_0 - U_1 \geq X(\beta_1 - \beta_0)) \\ &= X(\beta_1 - \beta_0) \underbrace{- \sqrt{\sigma^2 + 1 - 2\rho\sigma} \frac{\phi\left(\frac{X(\beta_1 - \beta_0)}{\sqrt{\sigma^2 + 1 - 2\rho\sigma}}\right)}{1 - \Phi\left(\frac{X(\beta_1 - \beta_0)}{\sqrt{\sigma^2 + 1 - 2\rho\sigma}}\right)}}_{<0}\end{aligned}$$

- Intuition: individuals do not select into college because they do not benefit from it

The basic Roy model, selection and target parameters

- Graphical intuition for the sign of the selection bias (the expectation of truncated normal):



What is a "Policy Relevant" Parameter?

- The MTE framework partitions all agents in a clear way
- Provides a foundation for thinking about "ideal" treatment effects
- The "ideal" treatment effect clearly depends on the question
- The ATE receives a lot of attention in the literature
  - But not very useful for policy – can agents still *choose* D?
- The ATT is somewhat clearer in this regard
  - Loss in benefit to treated group from discontinuing $D = 1$
- Perhaps more relevant is changing the agent's *choice problem*
- For example, $D \in \{0, 1\}$ is attending a four-year college
- Average effect of forcing college/no college (ATE) is not interesting
- Nor is the effect on college-goers of shutting down college (ATT)
- More interesting are the effects via D of adjusting tuition Z

Decomposition methods and the Pay Gap

Decomposition Methods

- Decomposition methods are traditionally used to separate differences between groups into a component explained by observables and an unexplained component
- This is conceptually similar to decomposing a treatment/control difference into a component explained by controls (bias) and an unexplained component (treatment effect)

Oaxaca-Blinder Decompositions

- Standard tool in labor economics: the Oaxaca (1973)-Blinder (1973) decomposition
- Let's briefly review how OB decompositions work
- There are two groups, M's and W's
- Average outcomes are $\bar{Y}_M$ and $\bar{Y}_W$, where $\bar{Y}_g = E\left[Y_{ig} | group_i = g\right]$
- We hope to explain group differences with an observed covariate vector X_i

Oaxaca-Blinder Decompositions

- Quantity to be explained:

$$\Delta \equiv \bar{Y}_M - \bar{Y}_W$$

- Suppose we run a regression for each group

$$Y_{ig} = X_{ig}'\beta_g + \epsilon_{ig}$$

- X_{ig} includes a constant
- OLS coefficient vector:

$$\beta_g = E\left[X_{ig}X_{ig}'|group_i = g\right]^{-1} E\left[X_{ig}Y_{ig}|group_i = g\right]$$

Oaxaca-Blinder Decompositions

- By construction OLS fits group means:

$$\bar{Y}_g = \bar{X}_g' \beta_g.$$

- Therefore we can write

$$\Delta = \bar{X}_M' \beta_M - \bar{X}_W' \beta_W$$

$$= \left(\bar{X}_M - \bar{X}_W\right)' \beta_M + \bar{X}_W' \left(\beta_M - \beta_W\right)$$

- The OB decomposition splits Δ into a component explained by X's, and a component explained by β's
- First term answers the question: How much more would M's make than W's if both groups were paid like M's for observables?
- Second term answers the question: How much of Δ is due to differences in earnings for M's and W's with the same characteristics?

Oaxaca-Blinder Decompositions

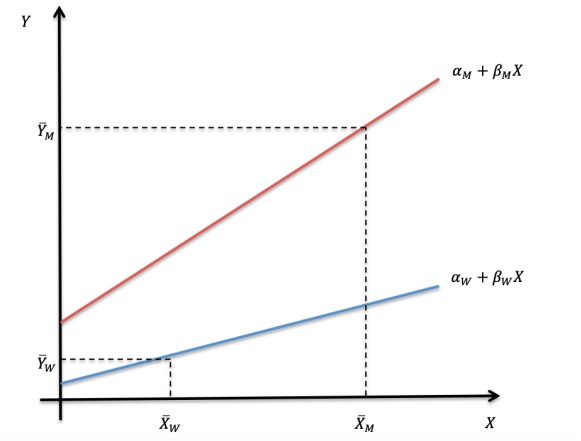
$$\Delta = \left(\bar{X}_M - \bar{X}_W\right)' \beta_M + \bar{X}_W' \left(\beta_M - \beta_W\right)$$

- Can also write the alternative decomposition:

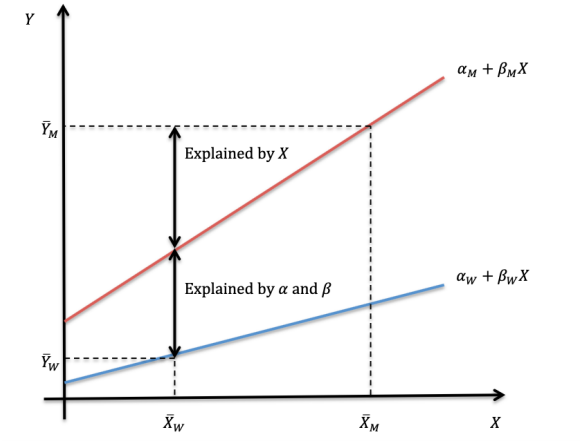
$$\Delta = \left(\bar{X}_M - \bar{X}_W\right)' \beta_W + \bar{X}_M' \left(\beta_M - \beta_W\right)$$

- New first term answers the question: How much more would M's make than W's if both groups were paid like W's for observables?
- Second term measures much of Δ is explained by the β's, weighting with M's characteristics

Oaxaca-Blinder Decompositions

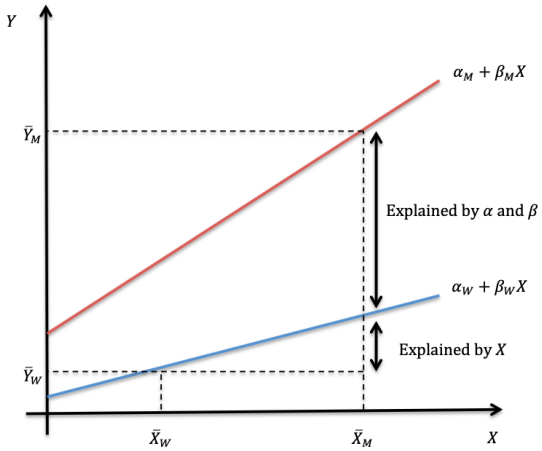


Oaxaca-Blinder Decompositions



Decomposition 1: What if women had male return to X?

Oaxaca-Blinder Decompositions



Decomposition 2: What if men had female return to X?

APPENDIX:
Example of pay-gap between natives and immigrants

Oaxaca-Blinder Decomposition Example from David Card

Example: let's look at our 2012 sample from the CPS. Here we will focus on men, age 30-35, and consider group a = natives and group b = immigrants. Some relevant information:

Natives:

mean log wage = 3.0129

mean education = 14.092 years

Immigrants:

mean log wage = 2.7660

mean education = 12.409 years

9

Oaxaca-Blinder Decomposition Example

	Pooled Model: Fit to Natives and Immigrants	
	(1)	(2)
Constant	3.013	1.546
	(0.006)	(0.025)
Immigrant	-0.247	-0.072
	(0.013)	(0.013)
Education (yrs)	--	0.104
		(0.002)
MSE	0.757	0.695
Adj. R-sq	0.018	0.173
Sample Size	19,092	19,092

Notes: Fit to data for males age 30-45 in March 2012 CPS. Dependent variable is log average hourly wage. Mean and standard deviation are 2.959 (0.764). Standard errors in parentheses.

Difference in mean wages

Difference in mean wages after "controlling" for education

Oaxaca-Blinder Decomposition Example

Let's perform the decomposition. We have $K = 2$, with the second variable being education.

$$\overline{y}^b - \overline{y}^a = 2.766 - 3.013 = -0.247$$

From the model in column 2 of the table, we have that

$$\overline{y}^b - \overline{y}^a = \sum_{j=2}^{K} (\overline{x}_j^b - \overline{x}_j^a)\widehat{\beta}_j + \widehat{\beta}_{K+1}$$

$$-0.247 = (12.409 - 14.092) \times 0.1041 - 0.0718$$

So the "effect of education" is $-1.683 \times 0.1041 = -0.175$ which is 70.9% of the wage gap. The remainder (29.1%) is "unexplained"

Oaxaca-Blinder Decomposition Example

	Pooled Model: Fit to Natives and Immigrants		Model for Natives	Model for Immigrants
	(1)	(2)	(3)	(4)
Constant	3.013	1.546	1.365	1.676
	(0.006)	(0.025)	(0.033)	(0.035)
Immigrant	-0.247	-0.072	--	--
	(0.013)	(0.013)	Coefficients are NOT the same	
Education (yrs)	--	0.104	0.117	0.088
		(0.002)	(0.002)	(0.002)
MSE	0.757	0.695	0.689	0.707
Adj. R-sq	0.018	0.173	0.146	0.208
Sample Size	19,092	19,092	14,921	4,141

Notes: Fit to data for males age 30-45 in March 2012 CPS. Dependent variable is log average hourly wage. Mean and standard deviation are: for overall sample, 2.959 (0.764); for natives 3.013 (0.746); for immigrants 2.766 (0.795). Standard errors in parentheses.

Oaxaca-Blinder Decomposition Example

Let's apply this to our example. Here we have

$$\begin{aligned} \hat{\beta}_2^a &= 0.117 \\ \hat{\beta}_2^b &= 0.088 \\ (\overline{x}_2^b - \overline{x}_2^a) &= 12.409 - 14.092 = 1.683 \end{aligned}$$

And we know $\overline{y}^b - \overline{y}^a = -0.247$. So if we use the coefficient for natives we have:

$$(\overline{x}_2^b - \overline{x}_2^a)\hat{\beta}_2^a = -0.197$$

$$\overline{x}_2^b(\hat{\beta}_2^b - \hat{\beta}_2^a) = -0.360$$

Whereas if we use the coefficient for immigrants we have

$$\begin{aligned} (\overline{x}_2^b - \overline{x}_2^a)\hat{\beta}_2^b &= -0.148 \\ \overline{x}_2^a(\hat{\beta}_2^b - \hat{\beta}_2^a) &= -0.409 \end{aligned}$$

Oaxaca-Blinder Decomposition Example

This shows a couple of important things. First, we have 2 estimates of the contribution of the difference in mean education:

$$\begin{aligned}(\bar{x}_2^b - \bar{x}_2^a)\hat{\beta}_2^a &= -0.197 \\ (\bar{x}_2^b - \bar{x}_2^a)\hat{\beta}_2^b &= -0.148\end{aligned}$$

Usually people interpret this as meaning that the effect is somewhere between -0.15 and -0.20 out of the total -0.247 wage gap. But what do we make out of the other term?

$$\begin{aligned}\bar{x}_2^b(\hat{\beta}_2^b - \hat{\beta}_2^a) &= -0.360 \\ \bar{x}_2^a(\hat{\beta}_2^b - \hat{\beta}_2^a) &= -0.409\end{aligned}$$

In either case we are "over-explaining" the wage gap (by a lot). If you look back at the fitted models you can see what is happening

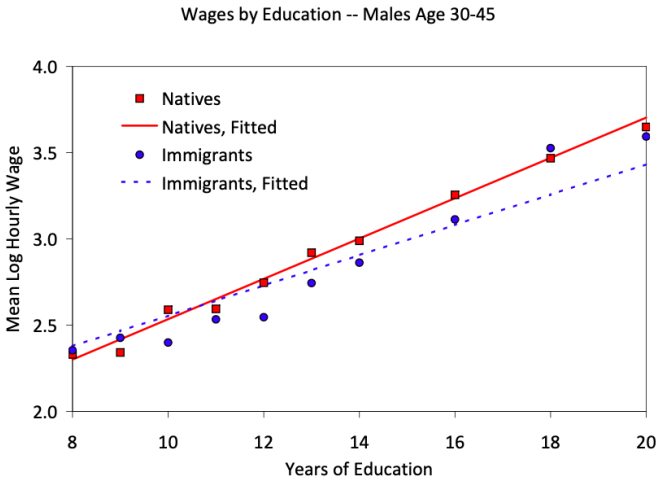
Oaxaca-Blinder Decomposition Example

	Pooled Model: Fit to Natives and Immigrants		Model for Natives	Model for Immigrants
	(1)	(2)	(3)	(4)
Constant	3.013	1.546	1.365	1.676
	(0.006)	(0.025)	(0.033)	(0.035)
Immigrant	-0.247	-0.072	--	--
	(0.013)	(0.013)		
Education (yrs)	--	0.104	0.117	0.088
		(0.002)	(0.002)	(0.002)
MSE	0.757	0.695	0.689	0.707
Adj. R-sq	0.018	0.173	0.146	0.208
Sample Size	19,092	19,092	14,921	4,141

Estimated intercepts are much different

Notes: Fit to data for males age 30-45 in March 2012 CPS. Dependent variable is log average hourly wage. Mean and standard deviation are: for overall sample, 2.959 (0.764); for natives 3.013 (0.746); for immigrants 2.766 (0.795). Standard errors in parentheses.

Oaxaca-Blinder Decomposition Example



Oaxaca-Blinder Decomposition Example

The decomposition is multiplying the difference in estimated "returns to education" – which is $0.088 - 0.117 = -0.029$ by numbers like 12 or 14, which "explains" a quite large difference in wages. The estimated constants are offsetting this so the total explained difference is always exactly -0.247.

We can see from this example that the part of the Oaxaca decomposition attributed to the difference in coefficients has to be interpreted carefully.

Oaxaca-Blinder Decomposition Example

Let's probe this a little more. Suppose instead of measuring education in "years," we measured in in "years of high school or more" i.e., we subtracted 8 from all measures of education.

$$\begin{aligned}\overline{y}^a &= \hat{\beta}_1^a + \hat{\beta}_2^a \overline{x}_2^a \\ &= \hat{\beta}_1^a + \hat{\beta}_2^a(\overline{x}_2^a - 8) + 8\hat{\beta}_2^a \\ &= (\hat{\beta}_1^a + 8\hat{\beta}_2^a) + \hat{\beta}_2^a(\overline{x}_2^a - 8)\end{aligned}$$

If we were to measure education as years of high school or more, we would get *exactly the same coefficient* on education, but the constant would be bigger (by exactly $8\hat{\beta}_2^a$). Likewise for group b :

$$\overline{y}^b = \hat{\beta}_1^b + \hat{\beta}_2^b \overline{x}_2^b = (\hat{\beta}_1^b + 8\hat{\beta}_2^b) + \hat{\beta}_2^b(\overline{x}_2^b - 8)$$

Oaxaca-Blinder Decomposition Example

If we examined the "difference in $x's$" part of the Oaxaca decompositiopn, we would compare differences in renormalized education:

$$(\overline{x}_2^b - 8) - (\overline{x}_2^a - 8) = \overline{x}_2^b - \overline{x}_2^a$$

multiplying by $\widehat{\beta}_2^a$ or $\widehat{\beta}_2^b$ – so we would get the same answer as before. But for the "difference in coefficients" part of the decomposition, we would look at

$$(\widehat{\beta}_2^b - \widehat{\beta}_2^a) \times (\overline{x}_2^b - 8)$$

or

$$(\widehat{\beta}_2^b - \widehat{\beta}_2^a) \times (\overline{x}_2^a - 8)$$

Oaxaca-Blinder Decomposition Example

Returning to our example:

$$\begin{aligned} \bar{x}_2^a &= 14.09 \\ \bar{x}_2^a &= 12.41 \\ \hat{\beta}_2^a &= 0.117 \\ \hat{\beta}_2^b &= 0.088 \end{aligned}$$

So if we use the renormalized mean for immigrants we have:

$$(\bar{x}_2^b - 8)(\hat{\beta}_2^b - \hat{\beta}_2^a) = 4.41 \times -0.029 = -0.128$$

Whereas if we use renormalized mean for natives we have:

$$(\bar{x}_2^a - 8)(\hat{\beta}_2^b - \hat{\beta}_2^a) = 6.09 \times -0.029 = -0.177$$

Which still "over-explains" the immigrant-native wage gap!

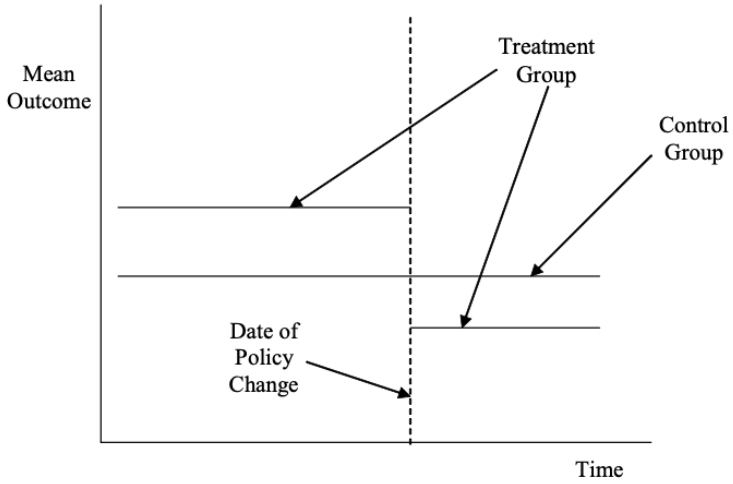
Outline

1. Preliminaries: research designs, DGP, and applied modeling
2. Connecting theory and data
   - Using supply and demand
   - Interpreting regression results with optimizing agents
3. Potential Outcomes and Selection
   - Selection Bias
   - Example: returns to selective colleges
4. The Roy Model and Selection Corrections
5. Decomposition methods and the Pay Gap
6. Review of 3 Applied Econometrics Tools
   - Difference-in-differences
   - Event Studies
   - Discrete Choice

Difference-in-differences

Difference-in-Differences

Idea is to use population unaffected by a program as a longitudinal control group for a treated population



Difference-in-Differences

$$Y_{it} = \mu + \alpha D_i + \gamma P_t + \beta D_i P_t + \varepsilon_{it}$$

- D_i - indicator for treatment group
- P_t - indicator for "post" period
- Parameter of interest is:

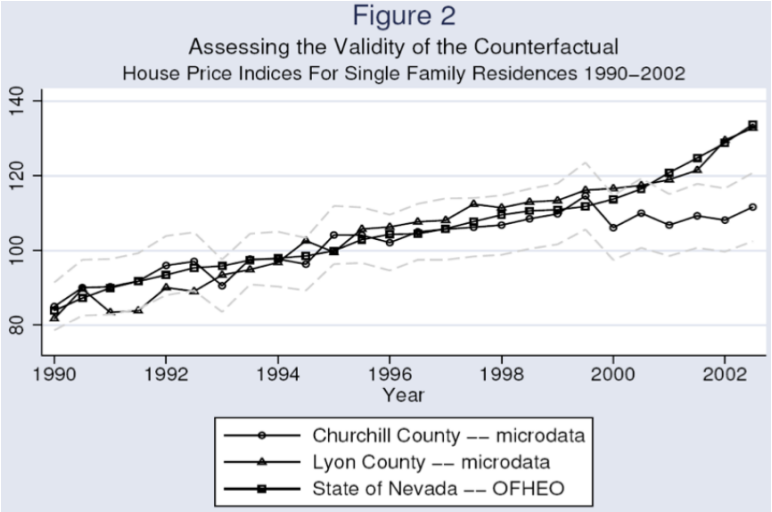
$$\begin{aligned}\beta = & \underbrace{\mathbb{E}\left[Y_{it}|D_i = 1, P_t = 1\right] - \mathbb{E}\left[Y_{it}|D_i = 1, P_t = 0\right]}_{\text{treatment change}} \\ & - \underbrace{\mathbb{E}\left[Y_{it}|D_i = 0, P_t = 1\right] - \mathbb{E}\left[Y_{it}|D_i = 0, P_t = 0\right]}_{\text{control change}}\end{aligned}$$

Difference-in-Differences

- The identifying assumption is that the change in the control group outcomes serves as a valid proxy for the change in the treatment group outcomes
- For this reason, the most convincing DDs exploit data from many periods
- Consider an example of Davis (2004) studying the effects of the emergence of a cancer Cluster in Churchill County in 2010

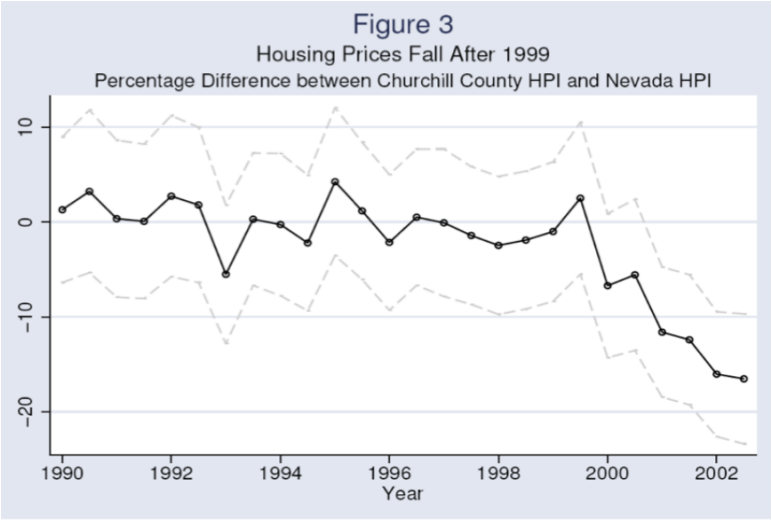
Treatment versus Control (Davis 2010)

Control groups (Lyon County and NV) rarely deviated from treatment group



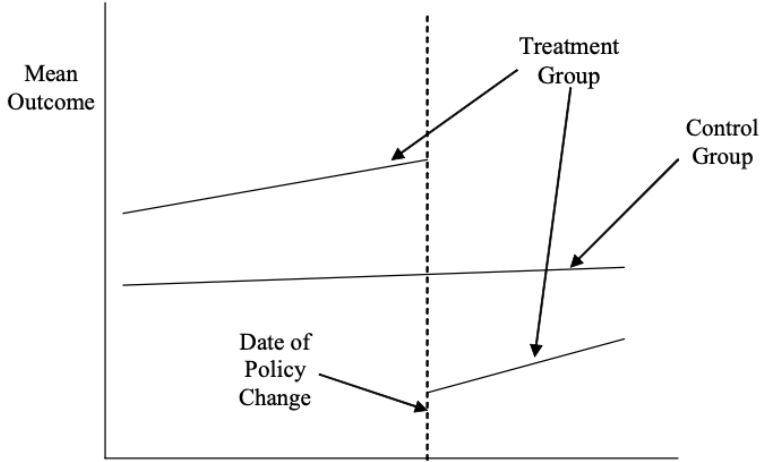
Difference-in-Differences (Davis 2010)

Compelling design because T-C difference has same long run mean and stable pre-treatment behavior



Difference-in-Differences: Issues with common trends

We worry about source of the trend (which is not observable) and whether it is stable



Difference-in-Differences: formalizing common trend assumption

We worry about source of the trend (which is not observable) and whether it is stable

Data with a time dimension

- The central problem in causal inference is inferring counterfactuals
- We observe Y_D, but what *would have* happened for $d \neq D$?
- Exploiting variation in treatment over time is a natural strategy
- Could be with **panel data** or **repeated cross sections**

$\rightarrow$ Distinction is a bit artificial (depends on unit of aggregation)

A simple, stylized, and useful setting

- Suppose that we have data from two (or more) time periods $T \in \{0, 1\}$
- Repeated cross section of two groups $G \in \{0, 1\}$
- Binary treatment $D \in \{0, 1\}$ and outcome of interest Y
- $D = GT$ — only group 1 gets treatment and only in time period 1
- Fix ideas: G is a (U.S.) state, T is a year, D is a policy change
- Data looks like $\{(Y_i, G_i, T_i)\}_{i=1}^{N}$, but gets aggregated to (G, T) cells

Difference-in-Differences: Potential outcome assumption

$$\overbrace{\underbrace{\mathbb{E}[Y_0|G=0,T=1]}_{\text{data}} - \underbrace{\mathbb{E}[Y_0|G=0,T=0]}_{\text{data}}}^{\text{time trend for } G=0}$$

$$= \overbrace{\underbrace{\mathbb{E}[Y_0|G=1,T=1]}_{\text{counterfactual}} - \underbrace{\mathbb{E}[Y_0|G=1,T=0]}_{\text{data}}}^{\text{time trend for } G=1}$$

$\rightarrow$ Says the effect of time for the two groups *would have* been the same

- Implies $\mathbb{E}[Y_0|G=1,T=1]$ is point identified, hence also

$$\underbrace{\mathbb{E}[Y_1 - Y_0|G=1,T=1]}_{\text{ATE for } G=1,T=1} = \underbrace{\mathbb{E}[Y_1|G=1,T=1]}_{\text{data}} - \underbrace{\mathbb{E}[Y_0|G=1,T=1]}_{\text{point identified}}$$

- Since $D = GT$, the left-hand side is just $\mathbb{E}[Y_1 - Y_0|D=1] = \text{ATT}$

Difference-in-Differences

- The name DID comes from rearranging the implied ATT:

$$\begin{aligned}\text{ATT} = & \underbrace{(\mathbb{E}[Y|G=1, T=1] - \mathbb{E}[Y|G=1, T=0])}_{\text{observed time difference, group 1}} \\ & - \underbrace{(\mathbb{E}[Y|G=0, T=1] - \mathbb{E}[Y|G=0, T=0])}_{\text{observed time difference, group 0}}\end{aligned}$$

- DID removes the time trend, assumed to be the same for both groups

- Alternatively, could arrange the expression as:

$$\begin{aligned}\text{ATT} = & \underbrace{(\mathbb{E}[Y|G=1, T=1] - \mathbb{E}[Y|G=0, T=1])}_{\text{observed group difference, time 1}} \\ & - \underbrace{(\mathbb{E}[Y|G=1, T=0] - \mathbb{E}[Y|G=0, T=0])}_{\text{observed group difference, time 0}}\end{aligned}$$

- Alternative statement of common trends is bias stays constant over time

Interpretation in Functional Form

- (G, T) takes four values, so write without loss of generality

$$\mathbb{E}[Y_0|G, T] = \pi_0 + \pi_T T + \pi_G G + \pi_{GT} GT$$

- The time trend for group 0 is:

$$\mathbb{E}[Y_0|G = 0, T = 1] - \mathbb{E}[Y_0|G = 0, T = 0] = \pi_T$$

- The time trend for group 1 is:

$$\mathbb{E}[Y_0|G = 1, T = 1] - \mathbb{E}[Y_0|G = 1, T = 0] = \pi_T + \pi_{GT}$$

- The common trends assumption is equivalent to assuming that $\pi_{GT} = 0$

$\rightarrow$ Then the common trend becomes π_T

- This interpretation is useful for setting up a regression...

Regression Implementation

Setup the regression

- Note again that $D = GT$, i.e. treatment only for group 1, time 1

$\rightarrow$ So $Y = Y_0 + GT(Y_1 - Y_0)$ and $\text{ATT} = \mathbb{E}[Y_1 - Y_0 | G = 1, T = 1]$

- Using the expression on the previous slide, we have

$$\begin{aligned}\mathbb{E}[Y|G,T] &= \mathbb{E}[Y_0|G,T] + D\,\mathbb{E}[Y_1 - Y_0|G,T] \\ &= \pi_0 + \pi_T T + \pi_G G + \pi_{GT} GT + D\,\mathbb{E}[Y_1 - Y_0|G,T] \\ &= \pi_0 + \pi_T T + \pi_G G + (\text{ATT} + \pi_{GT}) \times GT\end{aligned}$$

Implementation and interpretation

- If common trends holds, then $\pi_{GT} = 0$

$\rightarrow$ Regress Y on $1, G, T, (GT = D)$ — interaction coefficient estimates ATT

- To the extent that common trends fails, the coefficient picks up π_{GT} too

Event studies

Event studies

- Event studies are generalizations of the DD design where different units are treated at different times
- Started in finance. Typically looked at excess returns, which are deviations of stock returns over a level implied by some stock market index (in practice, one usually runs a regression of individual returns on a market average and takes a residual)
- Simply examined what happens to the mean value of excess returns in the neighborhood of a financial event
- The basic idea — of re-ordering a panel into event time — spilled over to evaluation literature

Event studies

- Let $e_i \in \{1, ..., T\}$ give date a unit is treated
- Define "event time" dummies:

$$D_{it}^k = 1\{t = e_i + k\}$$

- Even if all units are eventually treated, can pool diffs in diffs together using parametric linear model

$$Y_{it} = \alpha_i + \gamma_t + X_{it}'\phi + \sum_{k \in \mathcal{K}} \beta_k D_{it}^k + \varepsilon_{it}$$

- Typically choose two-sided distributed lag $\mathcal{K} = \{\underline{k}, ..., -2, 0, ..., \bar{k}\}$ where $(\underline{k}, \bar{k})$ are "binned" endpoints
  - Need to normalize one $\beta_k = 0$ b/c event time is a linear function of calendar time!
  - Showing leads allows one to look for pre-trends

Event studies

- The $\hat{\beta}_k$ can then be plotted over time and provide estimates of mean outcomes in "event time" after having taken out the individual and year specific effects.
- Jacobson, LaLonde, and Sullivan (1993) is a famous example. Look at effects of job loss on earnings
- JLS embellish the model to include individual-specific trends and allow for interactions between individual level characteristics and the event dummies

Event studies: Jacobson, LaLonde, and Sullivan (1993)

Jacobson, Lalonde, and Sullivan (1993, AER)

The figure that launched 1,000 dissertations..

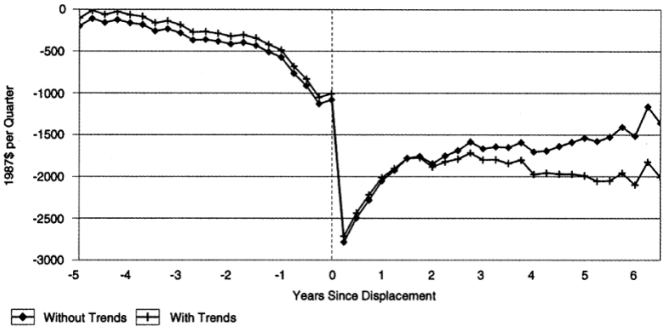


FIGURE 2. EARNINGS LOSSES FOR SEPARATORS IN MASS-LAYOFF SAMPLE

Event studies

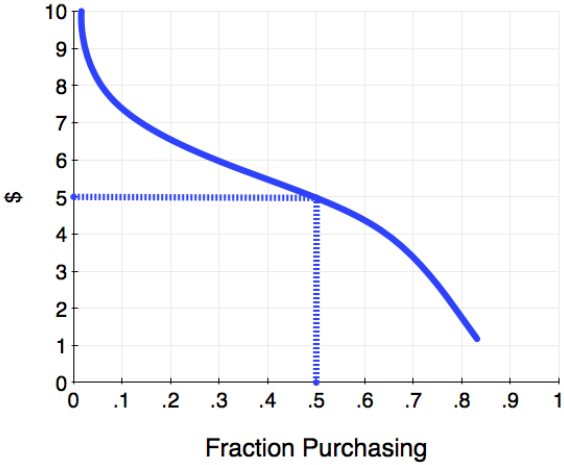
- Note that with a common event data $e_i = \bar{e}$, then the event study specification simplifies to a standard difference in differences model and the coefficients $\hat{\beta}_k$ merely plot the behavior of outcomes in the treatment group relative to the control group before and after treatment
- When treatment dates vary, then the ES approach is potentially a more efficient means of pooling together several different DDs, even in cases where all the units get treated.
- ES compares changes in outcomes of treated groups to both units that have not yet been treated and units that will never be treated – good to check if those two sets of controls are in fact exchangeable. (one can do this by re-estimating the model without the never treated units and seeing how estimates change
- Good practice to start simple and compare means of treatment and control groups in event time. Then add controls

Event studies: other tips

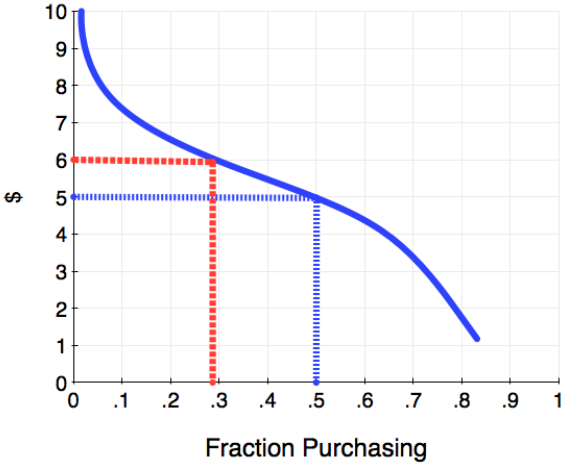
- Units that are treated early will have fewer pre-treatment obs and units that are treated late will have few post treatment obs, so that's why binning end points can make sense
- Alternatively, sometimes people focus on a balanced panel
- Note if you do not have any never treated controls, you will not be able to include all dummies so will have to normalize one of the event coefficients to zero, which JLS do (and -1 in event time is standard)
- It is common to find one thing in levels and another in logs because (in part) these are parametric models and linearity assumptions are not innocuous. Important to try different specifications of the time effects γ_t
- Greenstone Hornbeck and Moretti (2010) compute ES estimates on a treatment sample (which wins MDP) and a control sample (which narrowly lose them) and then take the difference to obtain more credible estimates of impacts. This extra difference is roughly equivalent to a DDD model

Discrete choice

Consumers decide whether or not to buy



Consumers decide whether or not to buy



Consumers decide whether or not to buy

- The first graph shows the share of consumers buying a product is 50% when it's price is $5
- The second graph shows the share of consumers buying a product is 30% when it's price is $6
- How can we think about how responsive demand will be to changes in price when consumers are making discrete (i.e., buy or not) choices?

Analytical Setup

- Suppose that individual i buys if her value exceeds the price, i.e., buy if $v_i > P$
- This value can be a function of common things (e.g., income, credit conditions, etc) or idiosyncratic tastes but at this stage, specifying what is in v_i doesn't matter. The fraction of people who buy is:

$$\text{Prob}(Q = 1) = P(v_i > P) \tag{11}$$
$$= 1 - F(P) \tag{12}$$

- where $F(x)$ is the c.d.f. of v_i. Note this is why the demand curve looks like a CDF rotated clockwise 90 degrees
- A c.d.f. describes the probability that a real-valued random variable X with a given probability distribution will be found to have a value less than or equal to x

Elasticity of Demand

- What is the elasticity of this curve?

$$Q(P) = N(1 - F(P)) \tag{13}$$

- where N is the size of the population (e.g., number of potential consumers in your market)

$$\varepsilon^D = \frac{dQ(P)}{dP}\frac{Q}{P} \tag{14}$$

Elasticity of Demand

- What is the derivative?

$$\frac{dQ(P)}{dP} = -Nf(P) \tag{15}$$

- where N is the size of the population (e.g., first time home buyers in an area)
- f(x) is the probability density function (p.d.f.)

Elasticity of Demand

$$\varepsilon^D = \frac{dQ(P)}{dP}\frac{P}{Q} \tag{16}$$

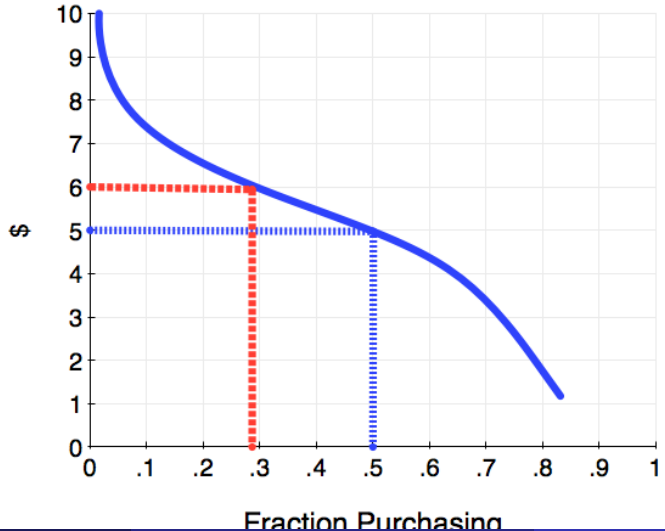
$$= -Nf(P)\frac{P}{N(1-F(P))} \tag{17}$$

$$= \frac{-f(P)}{1-F(P)}P \tag{18}$$

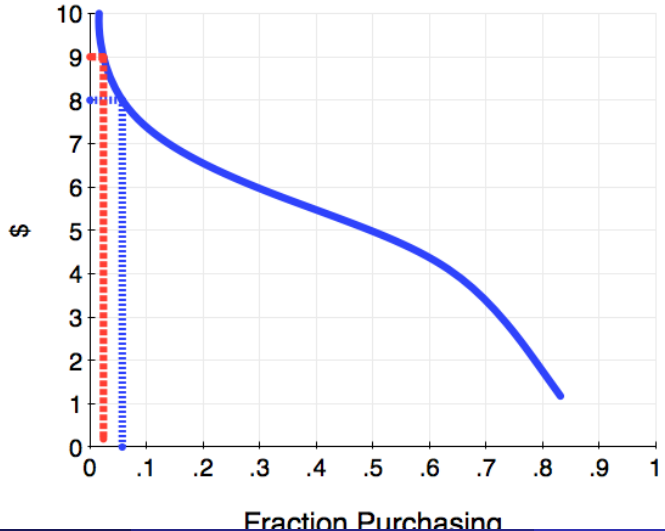
What matters for responsiveness?

- Fraction of people at the margin $f(P)$
- Fraction of people already buying $1 - F(P)$

From $5, a $1 dollar increase in price ⇓ demand by 20%



From $8, a $1 dollar increase in price ⇓ demand by 2%

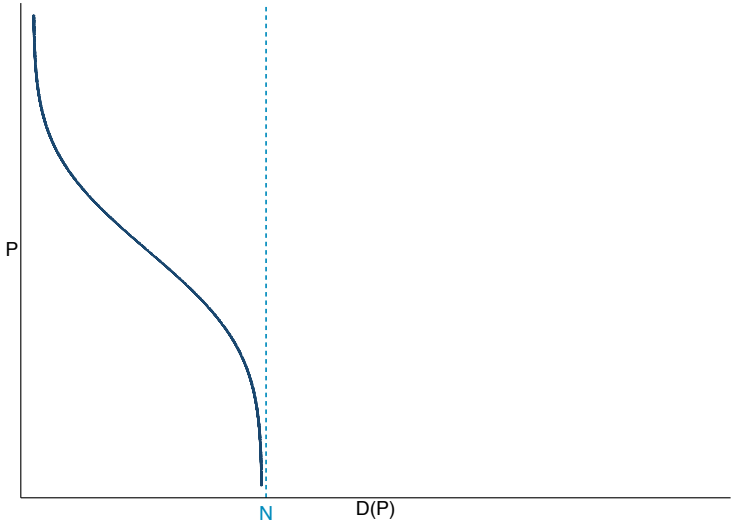


Elasticity of Demand: In words

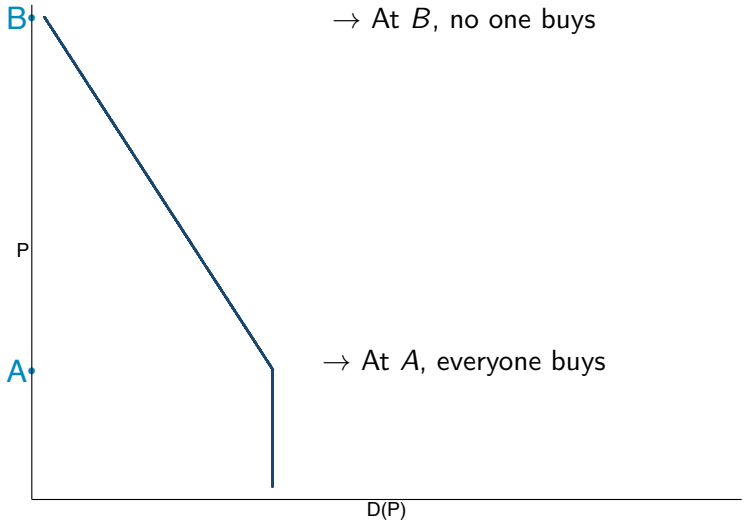
Takeaways:

- For very homogeneous populations, you'll have very elastic demand
- If tastes are more spread out, you'll see smaller responses
- At the extreme in which everyone is the same, demand will be a step function, so there is some price above which no one will buy and below which everyone will buy.
- In this case, things will be very inelastic at high prices, but very elastic near the price, and then unresponsive at very low prices
- Thinking about consumer choice in this way will be helpful for evaluating how effective sales can be

Demand if $V \sim N(\mu, \sigma)$



Demand if $V \sim U(A, B)$



$\rightarrow$ At B, no one buys

$\rightarrow$ At A, everyone buys

Discrete Choice more generally: $D_i = \text{argmax}_j U_{ij}$

- Model the indirect utility of alternative j as a function of observed individual characteristics X_i, alternative characteristics Z_j, and unobservable ϵ_{ij}:

$$U_{ij} = \mu_j(X_i, Z_j) + \epsilon_{ij}$$

$$\equiv V_{ij} + \epsilon_{ij}$$

- V_{ij} is the "observable" component of i's valuation
- Often $V_{ij} = Z_j'\gamma_i$, $\gamma_i = \gamma_0 + \gamma_1 X_i$
- ϵ_{ij} is the unobservable component of i's valuation

See Discrete Choice Methods with Simulation by Ken Train
`https://eml.berkeley.edu/books/choice2.html`.

Discrete Choice more generally: $D_i = \text{argmax}_j U_{ij}$

- The probability that i chooses j is:

$$Pr\left[D_i = j | X_i, Z_j\right] = Pr\left[V_{ij} + \epsilon_{ij} > V_{ik} + \epsilon_{ik},\ \forall k \neq j\ | V_{i1} ... V_{iJ}\right]$$

- In general this involves a $J-1$-dimensional integral. With two alternatives and the ϵ_{ij}'s independent of the V_{ij}'s:

$$Pr\left[D_i = 1 | V_{i1}, V_{i2}\right] = Pr\left[V_{i1} + \epsilon_{i1} > V_{i2} + \epsilon_{i2} | V_{i1}, V_{i2}\right]$$

$$= Pr\left[\epsilon_{i2} - \epsilon_{i1} < V_{i1} - V_{i2}\right]$$

$$= F_{\epsilon_2 - \epsilon_1}(V_{i1} - V_{i2})$$

- This requires us to evaluate the CDF of $\epsilon_{i2} - \epsilon_{i1}$

Discrete Choice more generally:

$$Pr\left[D_i = j | X_i, Z_j\right] = Pr\left[\epsilon_{ik} - \epsilon_{ij} < V_{ij} - V_{ik},\ \forall k \neq j\right]$$

- Observations:
  - Only relative utilities matter. Adding a constant to all V_{ij}'s has no effect
  - Scale doesn't matter. Multiplying all V_{ij}'s and ϵ_{ij}'s by $\lambda > 0$ has no effect
- Leads to location and scale normalizations
  - Location: Typically normalize one of the V_{ij}'s to 0
  - Scale: Typically fix variance of one (or more) of the ϵ_{ij}'s

Discrete Choice: probit (normal distribution)

$$Pr\left[D_i = j | X_i, Z_j\right] = Pr\left[\epsilon_{ik} - \epsilon_{ij} < V_{ij} - V_{ik},\ \forall k \neq j\right]$$

- Observations:
  - Only relative utilities matter. Adding a constant to all V_{ij}'s has no effect
  - Scale doesn't matter. Multiplying all V_{ij}'s and ϵ_{ij}'s by $\lambda > 0$ has no effect
- Leads to location and scale normalizations
  - Location: Typically normalize one of the V_{ij}'s to 0
  - Scale: Typically fix variance of one (or more) of the ϵ_{ij}'s

Discrete Choice: logit (logistic distribution)

- Logit-based choice models remain tractable even with many alternatives
- The logit model uses independent extreme value type I (EVI) distributions for the ϵ_{ij}'s
- Properties of the EVI distribution:
  - CDF: $F(\epsilon) = \exp(-\exp(-\epsilon))$
  - PDF: $f(\epsilon) = \exp(-\epsilon - \exp(-\epsilon))$
  - Mean: $\tau \approx 0.577$ (Euler's constant)
  - Variance: $\pi^2/6$
  - Difference in two independent EVI variables follows a logistic distribution
  - Special case of the Gumbel distribution – arises naturally as distribution of maxima (maximum of Gumbels is Gumbel)

Discrete Choice: logit (logistic distribution)

- With two alternatives, $\epsilon_{ij} \sim iid\ EVI$, we have the binary logit model:

$$Pr\left[D_i = 1 | V_{i1}, V_{i2}\right] = \frac{\exp(V_{i1} - V_{i2})}{1 + \exp(V_{i1} - V_{i2})}$$

$$\equiv \Lambda(V_{i1} - V_{i2}).$$

- This implies

$$\log\left(\frac{Pr\left[D_i = 1 | V_{i1}, V_{i2}\right]}{1 - Pr\left[D_i = 1 | V_{i1}, V_{i2}\right]}\right) = V_{i1} - V_{i2}$$

- When $V_{i1} - V_{i2} = X_i'\gamma$, coefficient γ is interpretable as a marginal effect on the log odds ratio (logit)

- Also known as "logistic regression" because $\epsilon_{i2} - \epsilon_{i1}$ follows a logistic distribution, and $\Lambda(\cdot)$ is the logistic CDF

Ken Train has an excellent book on discrete choice

- For more detail, see Discrete Choice Methods with Simulation by Ken Train
- The book is very readable and is freely available online:
  `https://eml.berkeley.edu/books/choice2.html`