# Part 3: Treatment Effects

Chris Conlon

Microeconometrics

May 26, 2017

# Overview

This Lecture will cover (roughly) the following papers:
Theory:

- Angrist and Imbens (1994)
- Heckman Vytlacil (2005/2007)
- Abadie and Imbens (2006)

Empirics:

- Chandra Staiger (2007/2011)
- Conlon and Mortimer (2014).

# The Evaluation Problem

- The issue we are concerned about is identifying the effect of a policy or an investment or some individual action on one or more outcomes of interest
- This has become the workhorse approach of the applied microeconomics fields (Public, Labor, etc.)
- Examples may include:
  - The effect of taxes on labour supply
  - The effect of education on wages
  - The effect of incarceration on recidivism
  - The effect of competition between schools on schooling quality
  - The effect of price cap regulation on consumer welfare
  - The effect of indirect taxes on demand
  - The effects of environmental regulation on incomes
  - The effects of labour market regulation and minimum wages on wages and employment

# The Evaluation Problem

- Define an outcome variable $Y_i$ for each individual
- Two potential outcomes for each person $\{Y_i(1), Y_i(0)\}$ depending on whether they receive treatment or not.
- Call $Y_i(1) - Y_i(0) = \beta_i$ the Treatment effect.
- Two major problems:
  - All individuals have different treatment effects (heterogeneity).
  - We don't actually observe any one person's treatment effect ! (Missing Data problem)
- We need strong assumptions in order to recover $f(\beta_i)$ from data.
- Instead we have to characterize simpler functions such as $E[\beta_i]$ (ATE) or $E[\beta_i|T_i = 1]$ (ATT) or $E[\beta_i|T_i = 0]$ (ATC) with fewer restrictions.

# More Difficulties

What is hard here?

- Heterogeneous effect of $\beta_i$ in population.
- Selection in treatment may be endogenous. That is $T_i$ depends on $Y_i(1), Y_i(0)$.
- Fisher or Roy (1951) model:

$$Y_i = (Y_i(1) - Y_i(0))T_i + Y_i(0) = \alpha + \beta_i T_i + u_i$$

- Agents usually choose $T_i$ with $\beta_i$ or $u_i$ in mind.
- Can't necessarily pool across individuals since $\beta_i$ is not constant.

# Structural vs. Reduced Form

- Usually we are interested in one or two parameters of the distribution of $\beta_i$ (such as the average or marginal treatment effect).
- Most program evaluation approaches seek to identify one effect or the other effect. This leads to these as being described as reduced form or quasi-experimental.
- The structural approach attempts to recover the entire joint $f(\beta_i, u_i)$ distribution but generally requires more assumptions, but then we can calculate whatever we need.

# Start with Easy Cases

- Let's start with the easy cases: run OLS and see what happens.
- OLS compares mean of treatment group with mean of control group (possibly controlling for other $X$)

$$\begin{aligned}\beta^{OLS} &= E(Y_i|T_i=1) - E(Y_i|T_i=0) \\ &= \underbrace{E[\beta_i = T_i = 1]}_{\text{ATT}} + \left( \underbrace{E[u_i|T_i=1] - E[u_i|T_i=0]}_{\text{selection bias}} \right)\end{aligned}$$

- Even in absence of heterogeneity $\beta_i = \beta$ we can still have selection bias.
- $Y_i^0 = \alpha + u_i$ may vary within the population (this is quite common).

# Solutions

1. Matching
2. Instrumental Variables
3. Diference in Difference and Natural Experiments
4. RCTs
5. Structural Models

- Key distinction: the treatment effect of some program (a number) from understanding how and why things work (the mechanism).
- Models let us link numbers to mechanisms.

# Matching

- Compare treated individuals to un-treated individuals with identical observable characteristics $X_i$.
- Key assumption: everything about $Y_i(1) - Y_i(0)$ is captured in $X_i$; or $u_i$ is randomly assigned conditional on $X_i$.
- Basic idea: The treatment group and the control group don't have the same distribution of observed characteristics as one another.
- Re-weight the un-treated population so that it resembles the treated population.
- Once distribution of $X_i$ is the same for both groups $X_i|T_i \sim X_i$ then we assume all other differences are irrelevant and can just compare means.
- Matching assumes all selection is on observables.

# Matching

- Formally the key assumption is the Conditional Independence Assumption (CIA)

$$\{Y_i^1, Y_i^0\} \perp T_i | X_i$$

- Once we know $X_i$ allocation to treatment $T_i$ is as if it is random.
- The only difference between treatment and control is composition of the sample.

# Matching

Let $F^1(x)$ be the distribution of characteristics in the treatment group, we can define the ATE as

$$E[Y(1) - Y(0)|T = 1] = E_{F^1(x)}[E(Y(1) - Y(0)|T = 1, X)]$$
$$= E_{F^1(x)}[E(Y(1)|T = 1, X)] - E_{F^1(x)}[E(Y(0)|T = 1, X)] \text{ linearity}$$

The first part we observe directly:

$$= E_{F^1(x)}[E(Y(1)|T = 1, X)]$$

But the counterfactual mean is not observed!

$$= E_{F^1(x)}[E(Y(0)|T = 1, X)]$$

But conditional independence does this for us:

$$E_{F^1(x)}[E(Y(0)|T = 1, X)] = E_{F^1(x)}[E(Y(0)|T = 0, X)]$$

# A Matching Example

Here is an example where I found that matching was helpful in my own work with Julie Mortimer:

- We ran a randomized experiment where we removed Snickers bars from around 60 vending machines in office buildings in downtown Chicago.
- We have a few possible control groups:
  1. Same vending machine in other weeks (captures heterogeneous tastes in the cross section)
  2. Other vending machines in the same week (might capture aggregate shocks, ad campaigns, etc.)
- We went with #1 as #2 was not particularly helpful.

# A Matching Example

Major problem was that there was a ton of heterogeneity in the overall level of (potential) weekly sales which we call $M_t$.

- Main source of heterogeneity is how many people are in the office that week, or how late they work.
- Based on total sales our average over treatment weeks was in the 74th percentile of all weeks.
- This was after removing a product, so we know sales should have gone down!
- How do we fix this without running the experiment for an entire year!
- Can't use shares instead of quantities. Why?


Vends
6000
4000
2000
0
0
50
100
150
Week
Total Sales
Focal Sales

# A Matching Example

Ideally we could just observe $M_t$ directly and use that as our matching variable $X$

- We didn't observe it directly and tried a few different measures:
  - Sales at the soda machine next to the snack machine
  - Sales of salty snacks at the same machine (not substitutes for candy bars).
  - We used k-NN with $k = 4$ to select control weeks – notice we re-weight so that overall sales are approximately same (minus the removed product).
- We also tried a more structured approach:
  - Define controls weeks as valid IFF
  - Overall sales were weakly lower
  - Overall sales were not less than Overall Sales less expected sales less Snickers Sales.

| Product | Control Mean | Control %ile | Treatment Mean | Treatment %ile | Mean Difference | % Δ |
|---|---|---|---|---|---|---|
| *Vends* | | | | | | |
| Peanut M&Ms | 359.9 | 73.6 | 478.3* | 99.4 | 118.4* | 32.9 |
| Twix Caramel | 187.6 | 55.3 | 297.1* | 100.0 | 109.5* | 58.4 |
| Assorted Chocolate | 334.8 | 66.7 | 398.0* | 95.0 | 63.2* | 18.9 |
| Assorted Energy | 571.9 | 63.5 | 616.2 | 76.7 | 44.3 | 7.8 |
| Zoo Animal Cracker | 209.1 | 78.6 | 243.7* | 98.1 | 34.6* | 16.5 |
| Salted Peanuts | 187.9 | 70.4 | 216.3* | 93.7 | 28.4 | 15.1 |
| Choc Chip Famous Amos | 171.6 | 71.7 | 193.1* | 95.0 | 21.5* | 12.5 |
| Ruger Vanilla Wafer | 107.3 | 59.7 | 127.9 | 78.6 | 20.6* | 19.1 |
| Assorted Candy | 215.8 | 43.4 | 229.6 | 60.4 | 13.7 | 6.4 |
| Assorted Potato Chips | 279.6 | 64.2 | 292.4* | 66.7 | 12.8 | 4.6 |
| Assorted Pretzels | 548.3 | 87.4 | 557.7* | 88.7 | 9.4 | 1.7 |
| Raisinets | 133.3 | 66.0 | 139.4 | 74.2 | 6.1 | 4.6 |
| Cheetos | 262.2 | 60.1 | 260.5 | 58.2 | -1.8 | -0.7 |
| Grandmas Choc Chip | 77.9 | 51.3 | 72.5 | 37.8 | -5.4 | -7.0 |
| Doritos | 215.4 | 54.1 | 203.1 | 39.6 | -12.3* | -5.7 |
| Assorted Cookie | 180.3 | 61.0 | 162.4 | 48.4 | -17.9 | -10.0 |
| Skittles | 100.1 | 62.9 | 75.1* | 30.2 | -25.1* | -25.0 |
| Assorted Salty Snack | 1382.8 | 56.0 | 1276.2* | 23.3 | -106.7* | -7.7 |
| Snickers | 323.4 | 50.3 | 2.0* | 1.3 | -321.4* | -99.4 |
| Total | 5849.6 | 74.2 | 5841.3 | 73.0 | -8.3 | -0.1 |

Notes: Control weeks are selected through the-neighbor matching using four control observations for each treatment week. Percentiles are relative to the full distribution of control weeks.

# Higher Dimensions

So matching works great in dimension 1. But what if $dim(X) > 1$?

- True high-dimensional matching may be infeasible. There may be no set of weights such that:
  $f(X_i|T_i = 1) \equiv \int w_i f(X_i|T_i = 0) \partial w_i$.
- One solution is the nearest-neighbor approach in Abadie Imbens (2006).
- This is still cursed in that our nearest neighbors get further away as the dimension grows.
- Suppose instead we had a sufficient statistic

# Propensity Score

- Rosenbaum and Rubin propose the propensity score

$$P(T_i = 1|X_i) \equiv P(X_i)$$

- They prove that the propensity score and any function of $X$, $b(X)$ which is finer serves as a balancing score.
- Finer implies that:

$$\begin{aligned} b(X^1) = b(X^2) &\implies P(X^1) = P(X^2) \\ P(X^1) = P(X^2) &\not\implies b(X^1) = b(X^2) \end{aligned}$$

# Propensity Score

- Main result: If treatment assignment is strongly ignorable conditional on $X$ (CIA) then it is strongly ignorable $Y(1), Y(0) \perp T|X$ given any balancing score $b(X)$ including the propensity score:

$$\begin{aligned} Pr(T = 1|Y(1), Y(0), P(X)) &= E[Pr(T = 1|Y(1), Y(0), X)|P(X)] \\ &= E[Pr(T = 1|x)|P(X)] = P(X) \end{aligned}$$

- Also we require that $0 < P(X) < 1$ at each $X$ which is known as the support condition.
- The theorem implies that given $P(X)$ we have as if random assignment.

# Propensity Score

- Instead of matching on $K$ dimensional $X$ we can now match on a one-dimensional propensity score
- Thus the propensity score provides dimension reduction
- We still have to estimate the propensity score which is a high dimensional problem without *ad-hoc* parametric restrictions.
- Let us begin by assuming a can-opener.

# Propensity Score

Just like in the matching case the problem arises because we do not observe the counterfactual mean:

$$E_{F^1(x)}[E(Y(0)|T=1,X)]$$

With conditional independence and the propensity score:

$$\begin{aligned} E_{F^1(x)}[E(Y(0)|T=1,X)] &= E_{F^1(x)}[E(Y(0)|T=0,X)] \\ &= E_{F^1(x)}[E(Y(0)|T=0,P(X))] \end{aligned}$$

# Kernel Matching

How do we implement?

- Kernels are an obvious choice

$$\widehat{ATT} = \frac{1}{N_1} \sum_{i \in T=1} \left[ Y_i - \frac{\sum_{j \in T=0} Y_j K\left(P(X_i) - P(X_j)\right)}{\sum_{s \in T=0} K\left(P(X_i) - P(X_s)\right)} \right]$$

  where $N_1$ is the sample size of the treatment group
  and $K(u)$ is a valid Kernel weight (people tend to use Gaussian Kernels here)
- As your propensity score gets further away from observation $i$ you get less weight
- As $h \to 0$ (or $\sigma_h$) the window gets smaller and we use fewer neighbors.

# Kernel Matching

- The usual caveats apply: $h$ determines the bias-variance tradeoff
- Choice of Kernel effects finite-sample properties
- Here the common support is important. We can only learn about cases where $P(X) \neq 1$ and $P(X) \neq 0$. If you always get treated (or not-treated) we cannot learn from this observation.
- We also have to be careful in choosing $X$ so as not to violate CIA (too many $X$'s , too few $X$'s) $\rightarrow$ have to think carefully!
- If you use propensity scores you will need a slide convincing us you have thought about why CIA holds for you!

# Gotcha!

Under CIA we know

$$G(Y(1), Y(0)|X, T) = G(Y(1), Y(0)|X)$$

Suppose we add in $Z$, then we require that:

$$G(Y(1), Y(0)|X, Z, T) = G(Y(1), Y(0)|X, Z)$$

However,

$$\begin{aligned} G(Y(1), Y(0)|X, T) &= \int G(Y(1), Y(0)|X, Z, T) dF(Z|X, T) \\ &= G(Y(1), Y(0)|X) \end{aligned}$$

where the last part follows by CIA.

- Thus each element can depend on $T$ conditional on $Z, X$ but the average may not.
- Mindless applications of matching can give you biased results!

# Matching and OLS

- Recall that OLS is a special case of Kernel regression (and hence matching!)
- Think about

$$Y = \alpha + \beta T_i + u$$

- Assume that $E(u|T, X) = E(u|X)$ which is a conditional mean independence assumption
- The we can get $\beta$ consistently (but not other variables) by running the following:

$$Y = \alpha + \beta T_i + \gamma X + v$$

- Again we are in the homogenous treatment world

# What about IV

So what does IV do?

- Let's assume a binary instrument $Z_i = 1$
- $Y_i(1), Y_i(0)$ depends on the value of $T_i$
- But now we endogenize $T_i(1), T_i(0)$ where the argument is the value of $Z_i$.
- We observe $\{Z_i, T_i = T_i(Z_i), Y_i = Y_i(T_i(Z_i))\}$.

# IV Assumptions

So what does IV do?

**Independence** $Z_i \perp Y_i(1), Y_i(0), T_i(1), T_i(0)$. Instrument is as if randomly assigned and does not directly affect $Y_i$

This is not implied by random assignment. In that case there would be four potential outcomes $Y_i(z, t)$

**Random Assignment**

$$Z_i \perp Y_i(0,0), Y_i(0,1), Y_i(1,0), Y_i(1,1), T_i(1), T_i(0).$$

**Exclusion Restriction** $Y_i(z, t) = Y_i(z', t)$ for all $z, z', t$.

Thus we require both RA and ER to guarantee Independence. The second assumption is a substantive one.

We only observe $(Z_i, T_i)$ not the pair $T_i(0), T_i(1)$ so we cannot determine compliance types directly! (See the picture)

# IV Assumptions

Table 1: Compliance Types

| | | $W_i(0)$ = 0 | $W_i(0)$ = 1 |
|---|---|---|---|
| $W_i(1)$ | 0 | never-taker | defier |
| | 1 | complier | always-taker |

# IV Assumptions

We are stuck without further assumptions, so we assume:

Monotonicity/No Defiers $T_i(1) \geq T_i(0)$

- Works in many applications (classical drug compliance).
- Implied by many latent index models with constant coefficients
- Works as long as sign of $\pi_{1,i}$ doesn't change

$$T_i(z) = 1[\pi_0 + \pi_1 z + \varepsilon_i > 0]$$

# IV Assumptions

Table 2: Compliance Type by Treatment and Instrument

| | | $Z_i$ = 0 | $Z_i$ = 1 |
|---|---|---|---|
| $W_i$ | 0 | complier/never-taker | never-taker/defier |
| | 1 | always-taker/defier | complier/always-taker |

# IV Assumptions

Table 3: Compliance Type by Treatment and Instrument given Monotonicity

| | | $Z_i$ = 0 | $Z_i$ = 1 |
|---|---|---|---|
| $W_i$ | 0 | complier/never-taker | never-taker |
| | 1 | always-taker | complier/always-taker |

# LATE Derivation

- We can derive the expression for $\beta_{IV}$ as:

$$\beta_{IV} = \frac{E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]}{E[T_i|Z_i = 1] - E[W_i|Z_i = 0]} = E[Y_i(1) - Y_i(0)|complier]$$

- We can derive the expression for $\pi_c$ (the fraction of compliers):

$$\pi_c = E[T_i|Z_i = 1] - E[T_i|Z_i = 0]$$

- Proof see Angrist and Imbens

# How Close to ATE?

Angrist and Imbens give some idea how close to the ATE the LATE is:

- $E[Y_i(0)|never - taker]$ and $E[Y_i(1)|always - taker]$ can be estimated form the data
- Compare these to their respective compliers $E[Y_i(0)|complier]$, $E[Y_i(1)|complier]$.
- When these are close then possibly $ATE \approx LATE$.

# How Close to ATE?

Angrist and Imbens give some idea how close to the ATE the LATE is:

$$\widehat{\beta}_1^{TSLS} \rightarrow^p \frac{E[\beta_{1i}\pi_{1i}]}{E[\pi_{i1}]} = LATE$$

$$LATE = ATE + \frac{Cov(\beta_{1i}, \pi_{1i})}{E[\pi_{1i}]}$$

- Weighted average for people with large $\pi_{1i}$.
- Late is treatment effect for those whose probability of treatment is most influenced by $Z_i$.
- If you always (never) get treated you don't show up in LATE.

# How Close to ATE?

- With difference instruments you get different $\pi_{1i}$ and TSLS estimators!
- Even with two valid $Z_1, Z_2$
  - Can be influential for different members of the population.
  - Using $Z_1$, TSLS will estimate the treatment effect for people whose probability of treatment $X$ is most influenced by $Z_1$
  - The LATE for $Z_1$ might differ from the LATE for $Z_2$
  - A J-statistic might reject even if both $Z_1$ and $Z_2$ are exogenous! (Why?).

# Example: Cardiac Catheterization

- $Y_i =$ surival time (days) for AMI patients
- $X_i =$ whether patient received cadiac catheterization (or not) (intensive treatment)
- $Z_i =$ differential distance to CC hospital

$$\begin{aligned} SurvivalDays_i &= \beta_0 + \beta_{1i} CardCath_i + u_i \\ CardCath_i &= \pi_0 + \pi_{1i} Distance_i + v_i \end{aligned}$$

- For whom does distance have the great effect on probability of treatment?
- For those patients what is their $\beta_{1i}$?

# Example: Cardiac Catheterization

- IV estimates causal effect for patients whose value of $X_i$ is most heavily influenced by $Z_i$
  - Patients with small positive benefit from CC in the expert judgement of EMT will receive CC if trip to CC hospital is short (compliers)
  - Patients that need CC to survive will always get it (always-takers)
  - Patients for which CC would be unnecessarily risky or harmful will not receive it (never-takers)
  - Patients for who would have gotten CC if they lived further from CC hospital (hopefully don't see) (defiers)
- We mostly weight towards the people with small positive benefits.

# Diversion Example

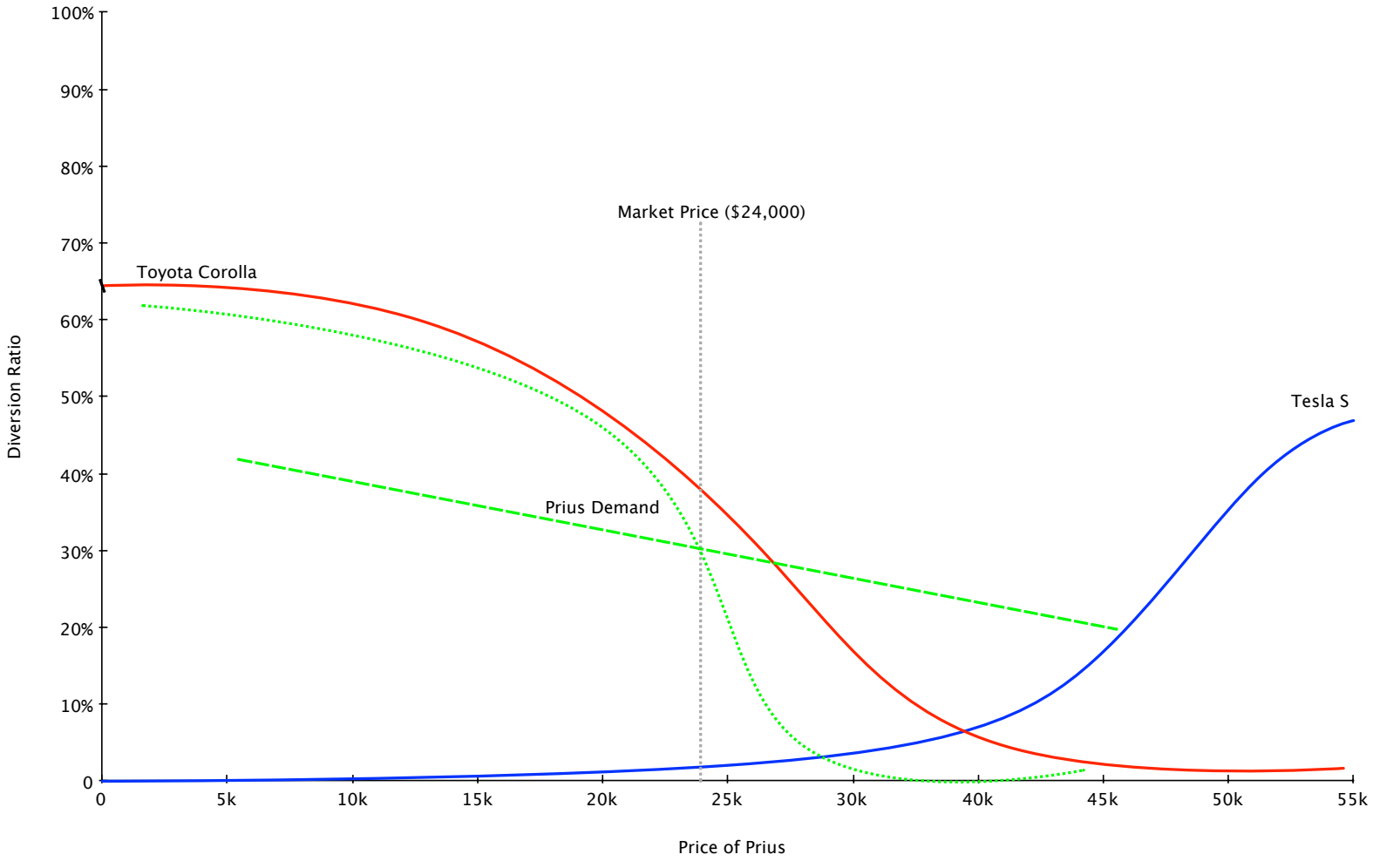
I have done some work trying to bring these methods into merger analysis.

- Key quantity: Diversion Ratio as I raise my price, how much do people switch to a particular competitor's product

$$D_{12} = \left| \frac{\partial q_2}{\partial p_1}(p) / \frac{\partial q_1}{\partial p_1}(p) \right|$$

- The treatment is leaving good 1.
- The $Y_i$ is increased sales of good 2.
- The $Z_i$ is the price of good 1.
- The key is that all changes in sales of 2 come through people leaving good 1 (no direct effects).

# Diversion for Prius (FAKE!)



# Diversion Example

What is the point?

- We might want to think about raising the price to choke price (or eliminating the product from the consumers choice set)
- Demand for Prius is steep: everyone leaves right away so $ATE \approx MTE$
- If Demand for Prius is not steep: now have to worry that at some prices it competes with Corolla and others it competes with the Tesla Model S.

# Local Average Treatment Effect

So how is this useful?

- It shows why IV can be meaningless when effects are heterogeneous.
- It shows that if the monotonicity assumption can be justified, IV estimates the effect for a particular subset of the population.
- In general the estimates are specific to that instrument and are not generalisable to other contexts.
- As an example consider two alternative policies that can increase participation in higher education.
  - Free tuition is randomly allocated to young people to attend college ($Z_1 = 1$ means that the subsidy is available).
  - The possibility of a competitive scholarship is available for free tuition ($Z_1 = 1$ means that the individual is allowed to compete for the scholarship).

# Local Average Treatment Effect

- Suppose the aim is to use these two policies to estimate the returns to college education. In this case, the pair $\{Y^1, Y^0\}$ are log earnings, the treatment is going to college, and the instrument is one of the two randomly allocated programmes.
- First, we need to assume that no one who intended to go to college will be discouraged from doing so as a result of the policy (monotonicity).
- This could fail as a result of a General Equilibrium response of the policy; for example, if it is perceived that the returns to college decline as a result of the increased supply, those with better outside opportunities may drop out.

# Local Average Treatment Effect

- Now compare the two instruments.
- The subsidy is likely to draw poorer liquidity constrained students into college but not necessarily those with the highest returns.
- The scholarship is likely to draw in the best students, who may also have higher returns.
- It is not a priori possible to believe that the two policies will identify the same parameter, or that one experiment will allow us to learn about the returns for a broader/different group of individuals.

# Local Average Treatment Effect

Finally, we need to understand what monotonicity means in terms of restrictions on economic theory.

- To quote from Vytlacil (2002) Econometrica:
  *"The LATE assumptions are not weaker than the assumptions of a latent index model, but instead impose the same restrictions on the counterfactual data as the classical selection model if one does not impose parametric functional form or distributional assumptions on the latter."*
- This is important because it shows that the LATE assumptions are equivalent to whatever economic modeling assumptions are required to justify the standard Heckman selection model and has no claim to greater generality.
- On the other hand there are no magical solutions to identifying effects when endogeneity/selection is present; this problem is exacerbated when the effects are heterogeneous and individuals select into treatment on the basis of the returns.

# Further approaches to evaluation of programme effects:

## Difference in Differences

- Sometimes we may feel we can impose more structure on the problem.
- Suppose in particular that we can write the outcome equation as

$$Y_{it} = \alpha_i + d_t + \beta_i T_{it} + u_{it}$$

- In the above we have now introduced a time dimension $t = \{1, 2\}$.
- Now suppose that $T_{i1} = 0$ for all $i$ and $T_{i2} = 1$ for a well defined group of individuals in our population. Denote the groups by $G$.
- This framework allows us to identify the ATT effect under the assumption that the growth of the outcome in the non-treatment state is independent of treatment allocation:

$$E[Y_{i2}^0 - Y_{i1}^0 | T] = E[Y_{i2}^0 - Y_{i1}^0]$$

# Difference in Differences

We have that

$$\begin{aligned} E[Y_{i2} - Y_{i1}|T_{i2} = 1] &= E[Y_{i2}^1 - Y_{i1}^1|T_{i2} = 1] \\ &= d_2 - d_1 + E[\beta_i|T_{i2} = 1] \end{aligned}$$

The above is the basis for the *before* and *after* estimator. However, this is not consistent because in general the time effects $d_t$ differ with $t$.Thus, if there is overall growth in the outcome variable this will be attributed to the treatment erroneously.

# Difference in Differences

Now consider:

$$E[Y_{i2}^0 - Y_{i1}^0 | T_{i2} = 1] = E[Y_{i2}^0 - Y_{i1}^0 | T_{i2} = 0]$$
$$E[Y_{i2} - Y_{i1} | T_{i2} = 0] = d_2 - d_1$$

We need an extra assumption: We can then obtain immediately an estimator for ATT as

$$E[\beta_i | T_{i2} = 1] = E[Y_{i2} - Y_{i1} | T_{i2} = 1] - E[Y_{i2} - Y_{i1} | T_{i2} = 0]$$

which can be estimated by the difference in the growth between the treatment and the control group.

# Difference in Differences

Now consider the following problem:

- Suppose we wish to evaluate a training programme for those with low earnings. Let the threshold for eligibility be $B$.
- We have a panel of individuals and those with low earnings qualify for training, forming the treatment group.
- Those with higher earnings form the control group.
- Now the low earning group is low for two reasons
  1. They have low permanent earnings ($\alpha_i$ is low) - this is accounted for by diff in diffs.
  2. They have a negative transitory shock ($u_{i1}$ is low) - this is not accounted for by diff in diffs.

# Difference in Differences

- #2 above violates the assumption $E[Y_{i2}^0 - Y_{i1}^0|T] = E[Y_{i2}^0 - Y_{i1}^0]$.
- To see why note that those participating into the programme are such that $Y_{i0}^0 < B$. Assume for simplicity that the shocks $u$ are $iid$. Hence $u_{i1} < B - \alpha_i - d_1$. This implies:

$$E[Y_{i2}^0 - Y_{i1}^0|T = 1] = d_2 = d_1 - E[u_{i1}|u_{i1} < B - \alpha_i - d_1]$$

For the control group:

$$E[Y_{i2}^0 - Y_{i1}^0|T = 1] = d_2 = d_1 - E[u_{i1}|u_{i1} > B - \alpha_i - d_1]$$

- Hence

$$E[Y_{i2}^0 - Y_{i1}^0|T = 1] - E[Y_{i2}^0 - Y_{i1}^0|T = 0] =$$
$$E[u_{i1}|u_{i1} > B - \alpha_i - d_1] - E[u_{i1}|u_{i1} < B - \alpha_i - d_1] > 0$$

- This is effectively regression to the mean: those unlucky enough to have a bad shock recover and show greater growth relative to those with a good shock. The nature of the bias depends on the stochastic properties of the shocks and how individuals select into training.

# Difference in Differences

Ashefelter (1978) was one of the first to consider difference in differences to evaluate training programmes.

TABLE 1.—MEAN EARNINGS PRIOR, DURING, AND SUBSEQUENT TO TRAINING FOR 1964 MDTA CLASSROOM TRAINEES AND A COMPARISON GROUP

| | White Males | | Black Males | | White Females | | Black Females | |
|---|---|---|---|---|---|---|---|---|
| | Trainees | Comparison Group | Trainees | Comparison Group | Trainees | Comparison Group | Trainees | Comparison Group |
| 1959 | $1,443 | $2,588 | $ 904 | $1,438 | $ 635 | $ 987 | $ 384 | $ 616 |
| 1960 | 1,533 | 2,699 | 976 | 1,521 | 687 | 1,076 | 440 | 693 |
| 1961 | 1,572 | 2,782 | 1,017 | 1,573 | 719 | 1,163 | 471 | 737 |
| 1962 | 1,843 | 2,963 | 1,211 | 1,742 | 813 | 1,308 | 566 | 843 |
| 1963 | 1,810 | 3,108 | 1,182 | 1,896 | 748 | 1,433 | 531 | 937 |
| 1964 | 1,551 | 3,275 | 1,273 | 2,121 | 838 | 1,580 | 688 | 1,060 |
| 1965 | 2,923 | 3,458 | 2,327 | 2,338 | 1,747 | 1,698 | 1,441 | 1,198 |
| 1966 | 3,750 | 4,351 | 2,983 | 2,919 | 2,024 | 1,990 | 1,794 | 1,461 |
| 1967 | 3,964 | 4,430 | 3,048 | 3,097 | 2,244 | 2,144 | 1,977 | 1,678 |
| 1968 | 4,401 | 4,955 | 3,409 | 3,487 | 2,398 | 2,339 | 2,160 | 1,920 |
| 1969 | $4,717 | $5,033 | $3,714 | $3,681 | $2,646 | $2,444 | $2,457 | $2,133 |
| Number of Observations | 7,326 | 40,921 | 2,133 | 6,472 | 2,730 | 28,142 | 1,356 | 5,192 |

# Difference in Differences

Ashenfelter (1978) reports the following results.

TABLE 2.—CRUDE ESTIMATES (AND ESTIMATED STANDARD ERRORS), ASSUMING $B=0$ AND $\beta_j'=0$ FOR $j>1$, OF THE EFFECT OF TRAINING ON EARNINGS DURING AND AFTER TRAINING, WHITE MALE MDTA 1964 CLASSROOM TRAINEES

| Effect in (value of $t$) | Value of Effects for $t-s=1963$ | $t-s=1962$ | $t-s=1961$ |
|---|---|---|---|
| 1962 | — | — | 91<br>(13) |
| 1963 | — | −179<br>(14) | −88<br>(17) |
| 1964 | −426<br>(16) | −605<br>(18) | −514<br>(20) |
| 1965 | 763<br>(20) | 584<br>(22) | 675<br>(23) |
| 1966 | 697<br>(25) | 518<br>(27) | 609<br>(28) |
| 1967 | 833<br>(28) | 655<br>(30) | 746<br>(31) |
| 1968 | 745<br>(34) | 566<br>(35) | 657<br>(36) |
| 1969 | 984<br>(37) | 805<br>(39) | 896<br>(40) |

# Difference in Differences

- The assumption on growth of the non-treatment outcome being independent of assignment to treatment may be violated, but it may still be true conditional on $X$.
- Consider the assumption

$$E[Y_{i2}^0 - Y_{i1}^0 | X, T] = E[Y_{i2}^0 - Y_{i1}^0 | X]$$

- This is just matching assumption on a redefined variable, namely the growth in the outcomes. In its simplest form the approach is implemented by running the regression

$$Y_{it} = \alpha_i + d_t + \beta_i T_{it} + \gamma_t' X_i + u_{it}$$

which allows for differential trends in the non-treatment growth depending on $X_i$. More generally one can implement propensity score matching on the growth of outcome variable when panel data is available.

# Difference in Differences with Repeated Cross Sections

- Suppose we do not have available panel data but just a random sample from the relevant population in a pre-treatment and a post-treatment period. We can still use difference in differences.
- First consider a simple case where $E[Y_{i2}^0 - Y_{i1}^0|T] = E[Y_{i2}^0 - Y_{i1}^0]$.
- We need to modify slightly the assumption to

$$E[Y_{i2}^0|\text{Group receiving training}] - E[Y_{i1}^0|\text{Group receiving training in the next period}] = E[Y_{i2}^0 - Y_{i1}^0]$$

  which requires, in addition to the original independence assumption that conditioned on particular individuals that population we will be sampling from does not change composition.
- We can then obtain immediately an estimator for ATT as

$$\begin{aligned} &E[\beta_i|T_{i2} = 1] \\ &= E[Y_{i2}|\text{Group receiving training}] - E[Y_{i1}|\text{Group receiving training next period}] \\ &- \{E[Y_{i2}|\text{Non-trainees}] - E[Y_{i1}|\text{Group not receiving training next period}]\} \end{aligned}$$

# Difference in Differences with Repeated Cross Sections

- More generally we need an assumption of conditional independence of the form

$$E[Y_{i2}^0|X, \text{Group receiving training}] - E[Y_{i1}^0|X, \text{Group receiving training next period}]$$
$$= E[Y_{i2}^0|X] - E[Y_{i1}^0|X]$$

- Under this assumption (and some auxiliary parametric assumptions) we can obtain an estimate of the effect of treatment on the treated by the regression

$$Y_{it} = \alpha_g + d_t + \beta T_{it} + \gamma' X_{it} + u_{it}$$

# Difference in Differences with Repeated Cross Sections

- More generally we can first run the regression

$$Y_{it} = \alpha_g + d_t + \beta(X_{it})T_{it} + \gamma' X_{it} + u_{it}$$

  where $\alpha_g$ is a dummy for the treatment of comparison group, and $\beta(X_{it})$ can be parameterized as $\beta(X_{it}) = \beta' X_{it}$. The ATT can then be estimated as the average of $\beta' X_{it}$ over the (empirical) distribution of $X$.
- A non parametric alternative is offered by Blundell, Dias, Meghir and van Reenen (2004).

# Difference in Differences and Selection on Unobservables

- Suppose we relax the assumption of *no selection* on unobservables.
- Instead we can start by assuming that

$$E[Y_{i2}^0|X,Z] - E[Y_{i1}^0|X,Z] = E[Y_{i2}^0|X] - E[Y_{i1}^0|X]$$

  where $Z$ is an instrument which determines training eligibility say but does not determine outcomes in the non-training state. Take $Z$ as binary (1,0).
- Non-Compliance: not all members of the eligible group ($Z = 1$) will take up training and some of those ineligible ($Z = 0$) may obtain training by other means.
- A difference in differences approach based on grouping by $Z$ will estimate the impact of being allocated to the eligible group, but not the impact of training itself.

# Difference in Differences and Selection on Unobservables

- Now suppose we still wish to estimate the impact of training on those being trained (rather than just the effect of being eligible)
- This becomes an IV problem and following up from the discussion of LATE we need stronger assumptions
  - Independence: for $Z = a, \{Y_{i2}^0 - Y_{i1}^0, Y_{i2}^1 - Y_{i1}^1, T(Z = a)\}$ is independent of Z.
  - Monotonicity $T_i(1) \geq T_i(0) \, \forall \, i$
- In this case LATE is defined by

$$[E(\Delta Y|Z = 1) - E(\Delta Y|Z = 0)]/[Pr(T(1) = 1) - Pr(T(0) = 1)]$$

assuming that the probability of training in the first period is zero.