NOVA IMS INFORMATION MANAGEMENT SCHOOL

Causal Inference: Potential Outcomes and Difference-in-Differences

Statistic for Data Science

Bruno Damásio
bdamasio@novaims.unl.pt
@bmpdamasio

2025/2026

Table of contents

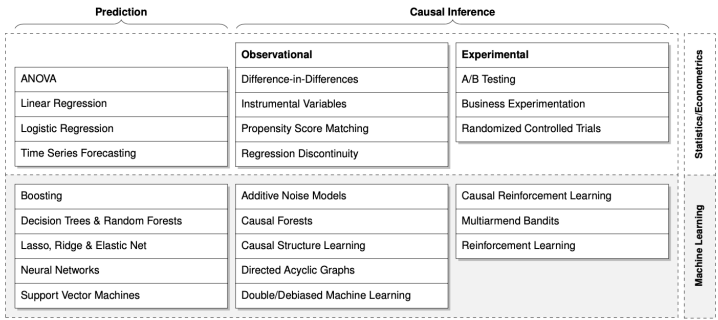
Motivation

Motivation

- So far, we have mainly focused on the study of the *ceteris paribus* relationship between a dependent variable and one (or more) independent variables.
- We inferred *associations* among variables, based on which we may do some *prediction*.
- **Causal inference** is about *counterfactual prediction*, predict what would have happened to the same units/subjects had they were exposed to a different (counterfactual) condition.

Causal Inference vs Prediction

Some examples of popular data analysis algorithms in statistics and econometrics, as well as machine learning and artificial intelligence, classified according to prediction and causal inference methods. Causal inference methods are further differentiated according to observational (based on ex-post observed data) and experimental approaches:



Causal Inference vs Prediction

Traditional prediction

- Traditional prediction seeks to detect patterns in data and fit functional relationships between variables with a high degree of accuracy
- "Does this person have heart disease?", "How many books will I sell?"
- It is not predictions of what effect a choice will have, though

Causal inference

- Causal inference is also a type of prediction, but it's a prediction of a *counterfactual* associated with a particular *choice taken*
- Causal inference takes that predicted (or imputed) counterfactual and constructs a causal effect that we hope tells us about a future in the event of a similar choice taken

#1: Correlation and causality are different concepts

Causal is one unit, correlation is many units

- Causal question: "If a doctor puts a patient on a ventilator (D), will her covid symptoms (Y) improve?"
- Correlation question:

$$\frac{\text{Cov}(D, Y)}{\sqrt{\text{Var}(D)}\sqrt{\text{Var}(Y)}}$$

- Error extends to predictive modeling that isn't based on causal frameworks

#2: Coming first may not mean causality!

- Every morning the rooster crows and then the sun rises
- Did the rooster cause the sun to rise? Or did the sun cause the rooster to crow?
- What if cat killed the rooster?
- *Post hoc ergo propter hoc*: "after this, therefore, because of this"

#3: Causality may mask correlations!

- A sailor is sailing her boat across the lake on the windy day. As the wind blows, she counters by turning the rudder in such a way so as to exactly offset the force of the wind.
- Back and forth she moves the rudder, yet the boat follows a straight line across the lake.
- A naive person with no knowledge of wind or boats might not see any relationship between the movement of the rudder and the direction of the boat.
- No correlation does not mean no causality.

Potential Outcomes

Naive causal inference

- Aliens estimate a model showing a systematic correlation between COVID deaths and ventilators
- They conclude doctors are killing patients with ventilators so they come to earth to liberate the patients, but it only makes things worse
- Their error was they confused correlation with causality, but deeper than that, they didn't understand how the world worked
- *We are the aliens in our research*

Potential Outcomes vs Realized Outcomes

- Econometrics traditionally modeled causality in terms of realized outcomes until recently (with some exceptions)
- We need to make a distinction between now the idea of data ("realized outcomes") and these hypothetical concepts ("potential outcomes").
- Potential outcomes is defined as a comparison between two states of the world.

Potential Outcomes vs Realized Outcomes

Example

- In the first state of the world (sometimes called the "actual" state of the world), we placed ventilators in hospitals where we have COVID patients.
- In the second state of the world (sometimes called the "counterfactual" state of the world), we don't place any ventilators in the same hospitals where we have the same COVID patients.
- The causal effect of the ventilators is the difference between these two states of the world.

Potential outcomes notation

Let the **treatment** be a binary variable:

$$D_{i,t} = \begin{cases} 1 \text{ if placed on ventilator at time } t \\ 0 \text{ if not placed on ventilator at time } t \end{cases}$$

where i indexes an individual observation, such as a person.
The treatment could be any particular intervention that can be manipulated.

Potential outcomes notation

Potential outcomes:

$$Y^{j}_{i,t} = \begin{cases} 1 \text{ health if placed on ventilator at time } t \\ 0 \text{ health if not placed on ventilator at time } t \end{cases}$$

where j indexes a potential treatment status for the same i person at the same t point in time.

- This means we have two separates states of the world for the exact same person at the exact same time.
- The state of the world where no treatment occurred is the control state.

Realized vs potential outcomes

- Potential outcome Y^1 refers to the "a priori fixed but unknown" outcomes associated with different possible treatment assignments
- Realized outcome Y refers to the "posterior and known" outcome associated with a specific treatment assignment
- Potential outcomes become realized outcomes through treatment assignment generated by an assignment mechanism like randomization or rationality

Treatment Assignment

- Treatment assignment *mechanism* drives the entire effort to identify causal effects as some make it easy and some make it potentially *impossible*.

Example

The ventilators will not be randomly distributed through hospitals, the assignment could be determined by the number of COVID patients in the hospitals, proportion of the population at risk in the area of the hospital, etc.

Important definitions

Definition 1: Individual treatment effect

The individual treatment effect, δ_i, associated with a ventilator is equal to $Y_i^1 - Y_i^0$.

Important definitions

Definition 2: Switching equation

An individual's realized health outcome, Y_i, is determined by treatment assignment, D_i which selects one of the potential outcomes:

$$Y_i = D_i Y_i^1 + (1 - D_i) Y_i^0$$

$$Y_i = \begin{cases} Y_i^1 \text{ if } D_i = 1 \\ Y_i^0 \text{ if } D_i = 0 \end{cases}$$

Not the same as treatment assignment mechanism. Treatment assignment mechanism describes how D was assigned, not whether it was assigned.

Missing data problem

Definition 3: Fundamental problem of causal inference

If you need both potential outcomes to know causality with certainty, then since it is impossible to observe both Y_i^1 and Y_i^0 for the same individual, δ_i, is *unknowable*.

Missing data problem

- Fundamental problem of causal inference is deep and impossible to overcome – not even with more data (you will always with more data be missing one of the potential outcomes)
- Causal inference is a missing data problem
- All of causal inference involves imputing missing counterfactuals and not all imputations are equal

Average Treatment Effects

Definition 4: Average treatment effect (ATE)

The average treatment effect is the population average of all i individual treatment effects

$$\begin{aligned} E[\delta_i] &= E[Y_i^1 - Y_i^0] \\ &= E[Y_i^1] - E[Y_i^0] \end{aligned}$$

Aggregate parameters based on individual treatment effects are *summaries* of individual treatment effects

Cannot be calculated because Y_i^1 and Y_i^0 do not exist *for the same unit i* due to switching equation

Conditional Average Treatment Effects

Definition 5: Average Treatment Effect on the Treated (ATT)

The average treatment effect on the treatment group is equal to the average treatment effect conditional on being a treatment group member:

$$\begin{aligned} E[\delta|D=1] &= E[Y^1 - Y^0|D=1] \\ &= E[Y^1|D=1] - E[Y^0|D=1] \end{aligned}$$

Cannot be calculated because Y_i^1 and Y_i^0 do not exist *for the same unit i* due to switching equation.

Conditional Average Treatment Effects

Definition 6: Average Treatment Effect on the Untreated (ATU)

The average treatment effect on the untreated group is equal to the average treatment effect conditional on being untreated:

$$\begin{aligned} E[\delta|D=0] &= E[Y^1 - Y^0|D=0] \\ &= E[Y^1|D=0] - E[Y^0|D=0] \end{aligned}$$

Cannot be calculated because Y_i^1 and Y_i^0 do not exist *for the same unit i* due to switching equation

Average Treatment Effects are Simple Summaries

- Notice how in all three of these, all we did was take the defined treatment effect at the individual and aggregate
- Because aggregate causal parameters are *summaries* of individual treatment effects, each of which cannot be calculated, the aggregates cannot be calculated either
- Missing data in this context isn't missing your car keys – it's missing unicorns and fire breathing dragons (fictional vs real data)
- While we cannot measure average causal effects, we can estimate them, but only in situations and we review one – randomization

Example

Perfect Doctor

- Let's consider the Perfect Doctor Example.
- The doctor knows each person's treatment effects, despite counterfactuals, and assigns treatment based on whether gains are positive or not.
- In this example, Y is the health outcome and D the treatment (binary variable).

Example - Perfect Doctor

	Potential Outcomes	
	Y^0	Y^1
	13	14
	6	0
	4	1
	5	2
	6	3
	6	1
	8	10
	8	9
True Averages	7	5

Example - Perfect Doctor

	Observed Data			
	D	Y^0	Y^1	Y^{obs}
	1	?	14	14
	0	6	?	6
	0	4	?	4
	0	5	?	5
	0	6	?	6
	0	6	?	6
	1	?	10	10
	1	?	9	9
Observed Averages		5.4	11	

Will the simple difference-in-means estimator return a valid estimate of the true causal effect?

Is the assignment mechanism in the perfect doctor example random?

Example - Perfect Doctor

	Potential Outcomes			Observed Data		
	Y^0	Y^1	D	Y^0	Y^1	Y^{obs}
	13	14	1	?	14	14
	6	0	0	6	?	6
	4	1	0	4	?	4
	5	2	0	5	?	5
	6	3	0	6	?	6
	6	1	0	6	?	6
	8	10	1	?	10	10
	8	9	1	?	9	9
True Averages	7	5	Observed Averages		5.4	11

$E[Y^1]$ - $E[Y^0]$ = -2

$E[Y^1 \mid D = 1]$-$E[Y^0 \mid D = 0]$ = 5.6

Example

Perfect Doctor

- In the Perfect Doctor example, the treatment each unit receives depends on that unit's potential outcomes (both missing and observed) $\rightarrow$ the assignment mechanism is confounded $\rightarrow$ the treatment effect estimator is biased!
- The assignment depends on both Y_i^1 and Y_i^0 for each unit (of course, in reality we will unlikely have such a perfect doctor).
- The key identifying assumptions in causal inference are on the **assignment mechanism**.

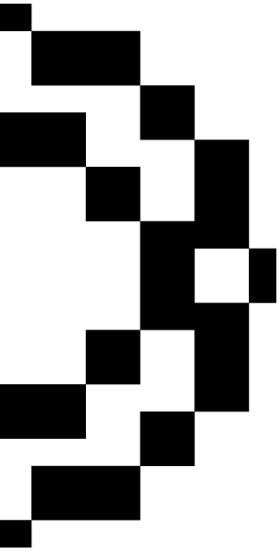
Decomposition of the Simple Differences Outcomes

Decomposition of the Simple Differences Outcomes

$$\underbrace{E[Y_i|D_i=1] - E[Y_i|D_i=0]}_{\text{Estimate of SDO}} = E[Y_i^1|D_i=1] - E[Y_i^0|D_i=0] =$$

$$E[Y_i^1|D_i=1] - E[Y_i^0|D_i=1] + E[Y_i^0|D_i=1] - E[Y_i^0|D_i=0] =$$

$$\underbrace{E[Y_i^1 - Y_i^0|D_i=1]}_{\text{Average Treatment Treated}} + \underbrace{E[Y^0|D=1] - E[Y^0|D=0]}_{\text{Selection bias}}$$



Directed Acyclic Graphs - DAGs

Graphs

- Now we turn from potential outcomes modeling of causal effects to causal graphs
- Very important area, very common to see it in computer science intersections with data science, particularly tech, and often very advanced

Judea Pearl and DAGs

- Judea Pearl and colleagues in Artificial Intelligence at UCLA developed DAG modeling to create a formalized causal inference methodology
- They make causality concepts extremely clear, they provide a map to the estimation strategy, and maybe best of all, they communicate to others what must be true about the data generating process to recover the causal effect

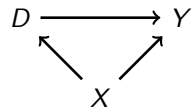
Summarizing Value of DAGs

1. Facilitates the task of designing identification strategy for estimating average causal effects
2. Facilitates the task of testing compatibility of the model with your data
3. Visualizes the identifying assumptions which opens up the model to critical scrutiny

Notations

- Treatment (e.g. intervention, exposure): D
- Outcome (e.g. disease status): Y
- Observed covariates or confounders: X
- Unobserved covariates or confounders: U

Simple DAG

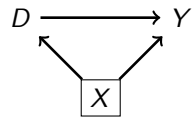


- In this DAG, we have three random variables: X, D and Y.
- There is a direct path from D to Y, which represents a causal effect.
- But there is also a second path from D to Y called the *backdoor path*.
- The backdoor path is given by $D \leftarrow X \rightarrow Y$.
- The backdoor path is not causal, we therefore call X a **confounder**.

Confounding

- Confounding occurs when the treatment and the outcomes have a common cause or parent which creates spurious correlation between D and Y.
- The *correlation* between D and Y no longer reflects the causal effect of D on Y because of selection bias due to a confounder.
- We need then to control for X to close the backdoor path.

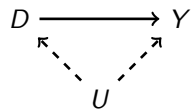
Backdoor Paths



- We can "block" any particular backdoor path by conditioning on variable X so long as it is not a collider (later)
- Once we condition on X, the correlation between D and Y estimates the causal effect of D on Y
- Conditioning means calculating $E[Y \mid D = 1, X] - E[Y \mid D = 0, X]$ for each value of X

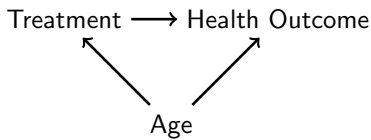
DAG with unobserved variables

- Let's look at a second DAG, which is subtly different from the first.
- Now we have a confounder that is unobserved.
- Since U is unobserved, that backdoor pathway is open.

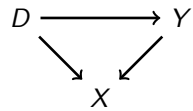


Confounding

- Going back to the Perfect Doctor example, we had confounding in the treatment assignment. What could cause this confounding?
  - Age
  - Gender
  - Genetics
  - etc.



Another simple DAG



- $D \rightarrow Y$ (causal effect of D on Y)
- $D \rightarrow X \leftarrow$ Y (backdoor path 1)

Collider

- Just like last time, there are two ways to get from D to Y.
- But this time, X has two arrows pointing to it, not away from it.
- When two variables cause a third variable along some path, we call that third variable a "collider".
- But so what? Colliders are special because when they appear along a backdoor path, that backdoor path is closed simply because of their presence.

Why is this important?

- DAGs allow us to map the relationship between variables.
- Identify colliders and confounding factors will determine which variables we should (and should not) condition, i.e., control for in the analysis.
- To identify a causal relationship, we want to block (close) all the backdoor paths.

Blocked backdoor paths

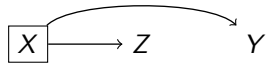
A backdoor path is blocked if and only if:

- It contains a noncollider that has been conditioned on
- Or it contains a collider that has not been conditioned on

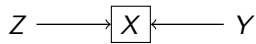
Examples of blocked paths

Examples:

1. Conditioning on a noncollider blocks a path:



2. Conditioning on a collider opens a path (i.e., creates spurious correlations):



3. *Not* conditioning on a collider blocks a path:

Backdoor criterion

Backdoor criterion

Conditioning on X satisfies the backdoor criterion with respect to (D, Y) directed path if:

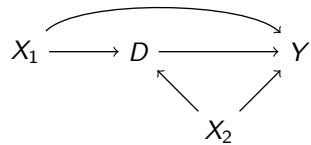
1. All backdoor paths are blocked by X
2. No element of X is a collider

In words: If X satisfies the backdoor criterion with respect to (D, Y), then controlling for or matching on X identifies the causal effect of D on Y.

And again note that a path which has a conditioned-on-collider can still be closed, but only with a noncollider-conditioned-on

What control strategy meets the backdoor criterion?

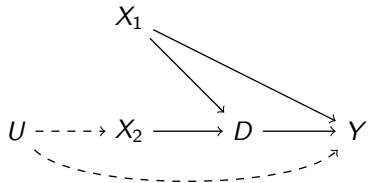
- List all backdoor paths from D to Y. I'll wait.



- What are the necessary and sufficient set of controls which will satisfy the backdoor criterion?

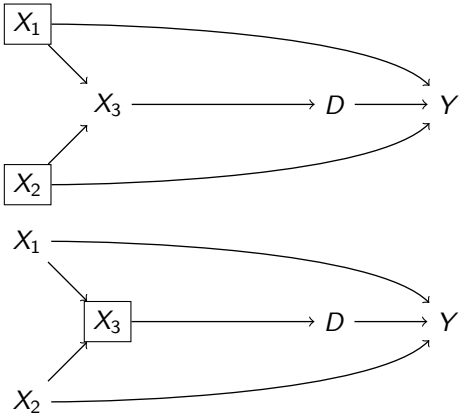
What if you have an unobservable?

- List all the backdoor paths from D to Y.



- What are the necessary and sufficient set of controls which will satisfy the backdoor criterion?
- What about the unobserved variable, U?

Multiple strategies



- Conditioning on the common causes, X_1 and X_2, is sufficient
- . . . but so is conditioning on X_3

Collider Bias

- You can identify average causal effects using a covariate adjustment strategy so long as you do not in the process inadvertently "open" backdoor paths
- Backdoor paths can remain open in covariate adjustment strategies through two ways:
  1. You did not close the path because you did not condition on the confounder
  2. Your conditioning variable opened up a previously closed backdoor path because on that path the variable was a **collider**
- Colliders are "bad controls"; they are often outcomes, but not always and not all outcomes create this problem
- This is the risk of blindly controlling for variables ("kitchen sink regressions")

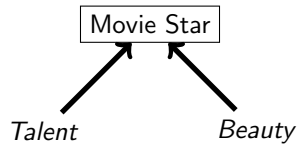
Movie star example of collider bias

Important: Since unconditioned colliders block back-door paths, what exactly does conditioning on a collider do? Let's illustrate with a fun example and some made-up data

- CNN.com headline: Megan Fox voted worst – but sexiest – actress of 2009 (link)
- Are these two things actually negatively correlated in the world?
- Assume talent and beauty are independent, but each causes someone to become a movie star. What's the correlation between talent and beauty for a sample of movie stars compared to the population as a whole (stars and non-stars)?

Movie star DAG

Imagine casting directors pick movie stars based on talent and beauty

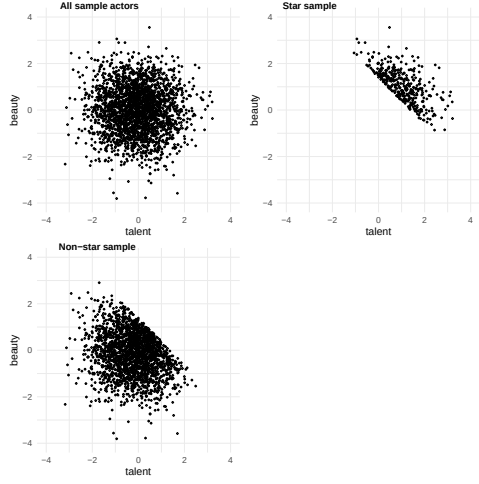


Talent and beauty can become correlated even though they are independent.

Movie star DGP

```
library(tidyverse)

set.seed(3444)

star_is_born <- tibble(
  beauty = rnorm(2500),
  talent = rnorm(2500),
  score = beauty + talent,
  c85 = quantile(score, .85),
  star = ifelse(score>=c85,1,0)
)
```

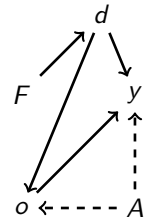
Movie star Example



Occupational sorting and discrimination example of collider bias

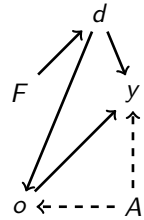
- Let's look at another example: very common for think tanks and journalists to say that the gender gap in earnings disappears once you control for occupation.
- But what if occupation is a collider, which it could be in a model with occupational sorting
- Then controlling for occupation in a wage regression searching for discrimination can lead to all kinds of crazy results *even in a simulation where we explicitly design there to be discrimination*

DAG



F is female, d is discrimination, o is occupation, y is earnings and A is ability. Dashed lines mean the variable cannot be observed. Note, by design, being a female has no effect on earnings or occupation, and has no relationship with ability. So earnings is coming through discrimination, occupation, and ability.

Occupational sorting and discrimination example of collider bias



Mediation and Backdoor paths

1. $d \rightarrow o \rightarrow y$
2. $d \rightarrow o \leftarrow A \rightarrow y$

Data Generating Process

```r
library(tidyverse)

tb <- tibble(
  female = ifelse(runif(10000)>=0.5,1,0),
  ability = rnorm(10000),
  discrimination = female,
  occupation = 1 + 2*ability + 0*female - 2*discrimination + rnorm(10000),
  wage = 1 - 1*discrimination + 1*occupation + 2*ability + rnorm(10000)
)

lm_1 <- lm(wage ~ female, tb)
lm_2 <- lm(wage ~ female + occupation, tb)
lm_3 <- lm(wage ~ female + occupation + ability, tb)
```

Occupational sorting and discrimination example of collider bias

Table 1: Regressions illustrating collider bias with simulated gender disparity

Covariates:	**Unbiased combined effect**	**Biased**	**Unbiased wage effect only**
Female	-3.074***	0.601***	-0.994***
	(0.000)	(0.000)	(0.000)
Occupation		1.793***	0.991***
		(0.000)	(0.000)
Ability			2.017***
			(0.000)
N	10,000	10,000	10,000
Mean of dependent variable	0.45	0.45	0.45

Occupational sorting and discrimination example of collider bias

- Recall we designed there to be a discrimination coefficient of -1
- If we do not control for occupation, then we get the combined effect of $d \rightarrow o \rightarrow y$ and $d \rightarrow y$
- Because it seems intuitive to control for occupation, notice column 2 - the sign flips!
- We are only able to isolate the direct causal effect by conditioning on ability and occupation, but ability is unobserved

Instrumental Variables

Instrumental Variables

What are instrumental variables?

- A instrumental variables (IV) is a variable, denoted Z, that:
  1. influences treatment assignment;
  2. is independent of unmeasured confounders;
  3. has no direct effect on the outcome, except through its effect on the treatment.

When should we use instrumental variables?

- When we have unobserved confounders;

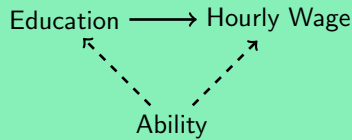
Why should we use instrumental variables

- Allows to estimate the variation in the treatment that is free of the unmeasured confounders;
- This confounder-free variation in the treatment to estimate the causal effect of the treatment.

Instrumental Variables

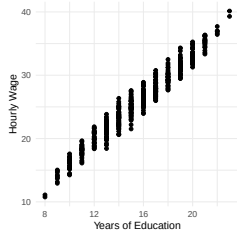
Example

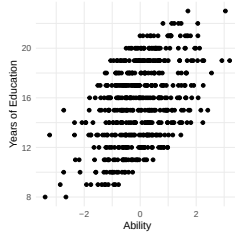
Suppose you want to identify the causal relationship of years of education on wage. One possible drawback could be due to the presence of an omitted unobservable variable, *ability*, that influences simultaneously education and wage.

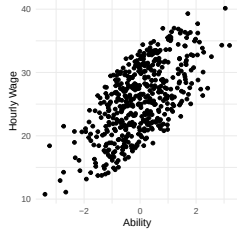


Here, *ability* is a confounder.

Instrumental Variables - Example







Instrumental Variables - Example

By analyzing these plots, we can see a positive relationship between years of education and hourly wage. But also, ability is positively related with years of education and hourly wage. As such, ideally, to uncover this causal effect we would just include ability in the linear regression. However, in practice, we cannot observe ability...

Instrumental Variables

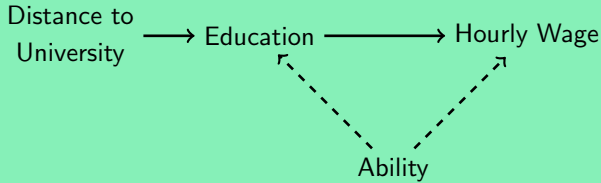
Example

Thus, we resort to instrumental variables! One plausible instrumental variable is distance to university. Why?

1. influences treatment assignment $\rightarrow$ (a) students that live close to university could be more motivated to study because they are in closed contact with the college community, (b) don't have to worry about high expenses to rent rooms.
2. is independent of unmeasured confounders $\rightarrow$ living close to a university it is not related with ability.
3. has no direct effect on the outcome, except through its effect on the treatment. $\rightarrow$ distance to university has an impact on years of education, which in turn impacts the hourly wage.

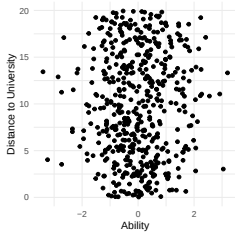
Instrumental Variables

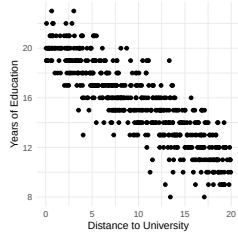
Example

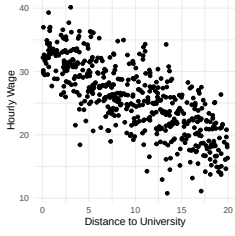


- "Distance to university" is correlated to "hourly wage", through the treatment variable (*education*).
- There is not a causal effect between "distance to university" and "hourly wage".
- Thus, we can know the causal effect by assessing the relationship between "distance to university" $\rightarrow$ "hourly wage" and removing the effect of "distance to university" $\rightarrow$ "education".

Instrumental Variables



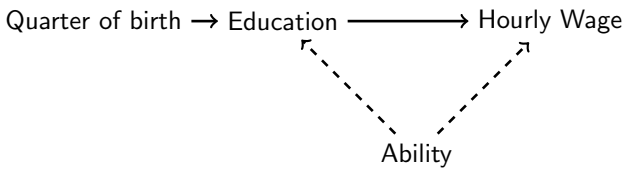




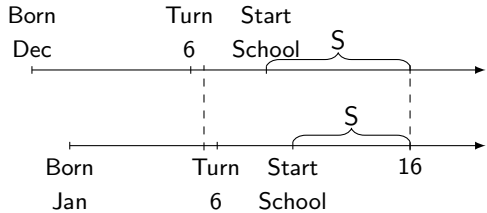
IV examples

The biggest challenge in using IV methods is finding a good IV. Some examples:
The half or quarter of the year of birth

- The date of birth throughout a year is largely randomized by nature, and does not affect later income directly.
- It does directly affects at what age the individual goes to school.
- Due to the compulsory education requirement, it can create one year difference in education.



IV examples

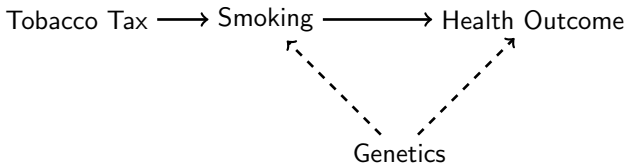


If you're born in the fourth quarter, you hit 16 with more schooling than those born in the first quarter

IV examples

Tobacco tax

- Tobacco tax rate does not directly affect one's health. But it affects the price of tobacco, thus in turn affects how much one smokes.



Two Stage Least Squares (2SLS)

To address estimation, let's consider the first example, with the ability bias:

$$\log(wage_i) = \beta_0 + \beta_1 educ_i + u_i \tag{1}$$

where *wage* is the hourly wage and *educ* years of education.

- We know that ability also influences hourly wage and is unobservable, as such we have $u_i = \varepsilon_i + ability_i$.
- We also know that education is correlated with ability, hence *educ* is a endogenous variable.
- In this case, the standard OLS estimator would be biased.

Two Stage Least Squares (2SLS)

Given an instrumental variable Z...

- What conditions must hold for a valid IV design?
  - $\text{Cov}(educ, Z) \neq 0 \rightarrow educ$ and Z are correlated
  - $\text{Cov}(ability, Z) = Cov(u, Z) = 0 \rightarrow$ "exclusion restriction". This means Z is orthogonal to the factors in u, such as unobserved ability as well as the error, u

Considering the previous example, we know that a plausible instrumental variable is distance to university ($dist\_univ$).

Two Stage Least Squares (2SLS)

- Causal model. Your main research question:

$$\log(wage_i) = \beta_0 + \beta_1 educ_i + u_i \tag{2}$$

- First-stage regression:

$$educ_i = \alpha_0 + \alpha_1 dist\_univ_i + v_i \tag{3}$$

- Second-stage regression. Notice the fitted values, $\widehat{educ}$:

$$\log(wage_i) = \delta_0 + \beta_1 \widehat{educ_i} + e_i \tag{4}$$

Two Stage Least Squares (2SLS)

The fitted value, $\widehat{educ}$, can be seen as the estimated version of *educ* that is uncorrelated with u. Therefore, 2SLS first "purges" *educ* of its correlation with u before doing the OLS regression.

Specification Tests

- **Weak instruments**
  - The null hypothesis is H_0: "All instruments are weak". Thus, rejecting the null hypothesis indicates the instruments used are strong, i.e., it is highly correlated with the endogeneous regressor it concerns.
- **Hausman test for endogeneity**
  - The null hypothesis is H_0: $\text{Cov}(x, u) = 0$. Thus, rejecting the null hypothesis indicates the existence of endogeneity and the need for instrumental variables.
- **Sargan test**
  - This test for the validity of instruments (whether the instruments are corrrelated with the error term) can only be performed for the extra instruments, those that are in excess of the number of endogenous variables.
  - The null hypothesis is H_0: $\text{Cov}(z, u) = 0$. Thus, rejecting the null hypothesis indicates that at least one of the extra instruments is not valid.

Instrumental Variables in R

Example

Now, let's apply this in R! Using data from the U.S. National Longitudinal Survey of Young Men in 1976, consider the following variables:

- *wage*, in cents per hour;
- *education*, years of education;
- *experience*, years of experience;
- *married*, dummy variable indicating whether the individual is married;
- *smsa*66, dummy variable indicating whether the individual lived in the SMSA (standard metropolitan statistical area) in 1976.

Instrumental Variables in R

```
library(ivreg)
library(wooldridge)

data("SchoolingReturns", package = "ivreg")

m_ols <- lm(log(wage) ~ education + experience + I(experience^2)
            + married + smsa66, data = SchoolingReturns)

summary(m_ols)
```

Instrumental Variables in R

```
##
## Call:
## lm(formula = log(wage) ~ education + experience + I(experience^2) +
##     married + smsa66, data = SchoolingReturns)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.84705 -0.24204  0.02312  0.25748  1.35389
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      4.4434180  0.0666673  66.651  < 2e-16 ***
## education        0.0849225  0.0035243  24.096  < 2e-16 ***
## experience       0.0766309  0.0069443  11.035  < 2e-16 ***
## I(experience^2) -0.0021033  0.0003292  -6.389 1.93e-10 ***
## marriedyes       0.1593040  0.0160232   9.942  < 2e-16 ***
## smsa66yes        0.1549223  0.0149435  10.367  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3857 on 2997 degrees of freedom
##   (7 observations deleted due to missingness)
## Multiple R-squared:  0.2456,Adjusted R-squared:  0.2444
## F-statistic: 195.2 on 5 and 2997 DF,  p-value: < 2.2e-16
```

Instrumental Variables in R

```
library(ivreg)

data("SchoolingReturns", package = "ivreg")

m_iv <- ivreg(log(wage) ~ education + experience + I(experience^2) + married + smsa66 |
                nearcollege + experience + I(experience^2) + married + smsa66,
              data = SchoolingReturns)

summary(m_iv)
```

Instrumental Variables in R

```
##
## Call:
## ivreg(formula = log(wage) ~ education + experience + I(experience^2) +
##     married + smsa66 | nearcollege + experience + I(experience^2) +
##     married + smsa66, data = SchoolingReturns)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1611 -0.2634  0.0174  0.2894  1.5097
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      2.8252768  0.7871992   3.589 0.000337 ***
## education        0.1834611  0.0478840   3.831 0.000130 ***
## experience       0.1208182  0.0227754   5.305 1.21e-07 ***
## I(experience^2) -0.0022080  0.0003731  -5.918 3.63e-09 ***
## marriedyes       0.1138259  0.0284390   4.002 6.42e-05 ***
## smsa66yes        0.0964324  0.0329224   2.929 0.003425 **
##
## Diagnostic tests:
##                   df1  df2 statistic  p-value
## Weak instruments    1 2997    20.611 5.85e-06 ***
## Wu-Hausman          1 2996     5.384   0.0204 *
## Sargan              0   NA        NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4331 on 2997 degrees of freedom
## Multiple R-Squared: 0.04887,Adjusted R-squared: 0.04729
## Wald test: 65.64 on 5 and 2997 DF,  p-value: < 2.2e-16
```

Difference-in-Differences

What is difference-in-differences (DiD)

- DiD is a very old, relatively straightforward, intuitive research design
- A group of units are assigned some treatment and then compared to a group of units that weren't
- One of the most widely used quasi-experimental methods in economics and increasingly in industry
- Mostly associated with "big shocks" happening in space over time

  "A good way to do econometrics is to look for good natural experiments and use statistical methods that can tidy up the confounding factors that nature has not controlled for us." – Daniel McFadden (Nobel Laureate recipient with Heckman 2002)

What is difference-in-differences (DiD)

Minimum wage on employment - Card and Krugger (1994)

- Goal: Evaluate the effect of minimum wage on employment.
- Units: fast-food restaurants in New Jersey (NJ) and adjacent eastern Pennsylvania (PA).
- Intervention: Raise of the state minimum wage; NJ raised the minimum wage on April 1, 1992, but PA did not.
- Approach: Compare the employment (average per store), observed in both areas, before and after the change.
- Outcome: After the increase in minimum wage, NJ's employment increased.

So, how does this estimator work?

Difference-in-Differences Estimator

Diff-in-Diffs: Without regression

- We start by taking the average value of each group's outcome before and after the treatment

	Treatment Group	**Control Group**
Before	$\overline{y}_B^T$	$\overline{y}_B^C$
After	$\overline{y}_A^T$	$\overline{y}_A^C$

- Then, calculate the difference-in-differences of the averages

$$\left(\overline{y}_A^T - \overline{y}_B^T\right) - \left(\overline{y}_A^C - \overline{y}_B^C\right) \tag{5}$$

Diff-in-Diffs: With regression

- We can get the same result in a regression framework

$$y_{it} = \beta_0 + \beta_1 D_i + \beta_2 Post_t + \beta_3 (D_i \times Post_t) + u_{it} \quad (6)$$

  where $D = 1$ if i is in the treatment group and $Post = 1$ if t is after the treatment.
- We only have two groups (treated/untreated) and two periods (before/after the treatment).

Difference-in-Differences Estimator

In practice, equation (6) implies that we have one regression for the control group:

$$y_{it} = \beta_0 + \beta_2 Post_t + u_{it} \tag{7}$$

and another one for the treated group:

$$y_{it} = (\beta_0 + \beta_1) + (\beta_2 + \beta_3) Post_t + u_{it} \tag{8}$$

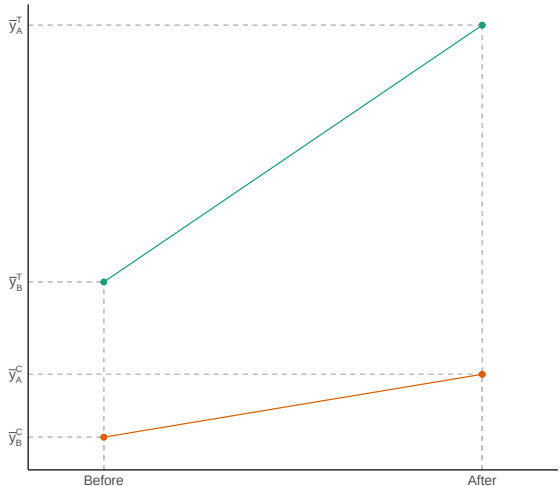
Since, $Post_t$ is a dummy variable, estimating equation (6) through OLS would yield:

- $\overline{y}_{C,B} = \hat{\beta}_0$
- $\overline{y}_{T,B} = \hat{\beta}_0 + \hat{\beta}_1$
- $\overline{y}_{C,A} = \hat{\beta}_0 + \hat{\beta}_2$
- $\overline{y}_{T,A} = \hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3$

Difference-in-Differences Estimator

The regression framework allows to add regressors as controls. In that case, the interpretation is no longer as sample average, but as partial effects after controlling for other factors.

Difference-in-Differences Estimator



Difference-in-Differences Estimator

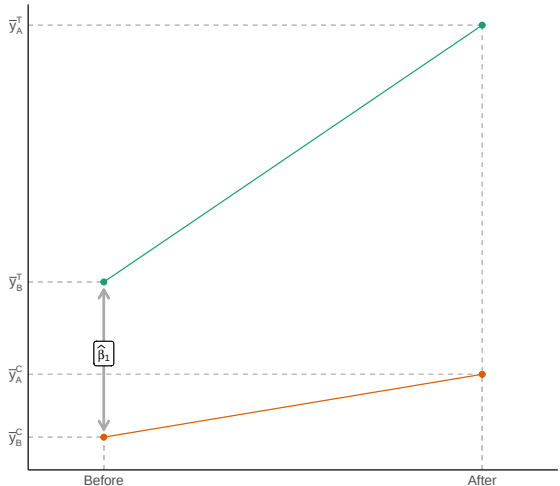
Knowing that we have:

$$\begin{aligned} \overline{y}_{C,B} &= \hat{\beta}_0 \\ \overline{y}_{T,B} &= \hat{\beta}_0 + \hat{\beta}_1 \end{aligned}$$

Then,

$$\overline{y}_{T,B} - \overline{y}_{C,B} = \hat{\beta}_1$$

Difference-in-Differences Estimator



Difference-in-Differences Estimator

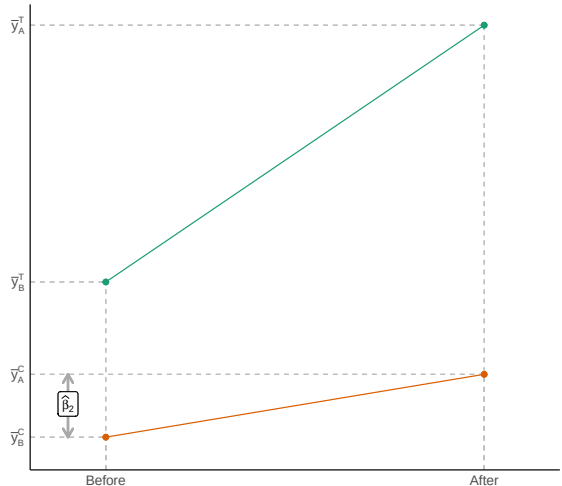
Knowing that we have:

$$\begin{aligned} \overline{y}_{C,B} &= \hat{\beta}_0 \\ \overline{y}_{C,A} &= \hat{\beta}_0 + \hat{\beta}_2 \end{aligned}$$

Then,

$$\overline{y}_{C,A} - \overline{y}_{C,B} = \hat{\beta}_2$$

Difference-in-Differences Estimator



Difference-in-Differences Estimator

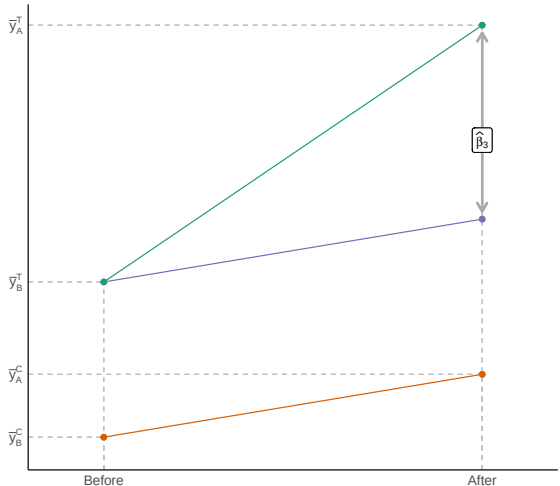
Let's recall the formula to calculate the DiD estimator:

$$\left(\overline{y}_{T,A} - \overline{y}_{T,B}\right) - \left(\overline{y}_{C,A} - \overline{y}_{C,B}\right)$$

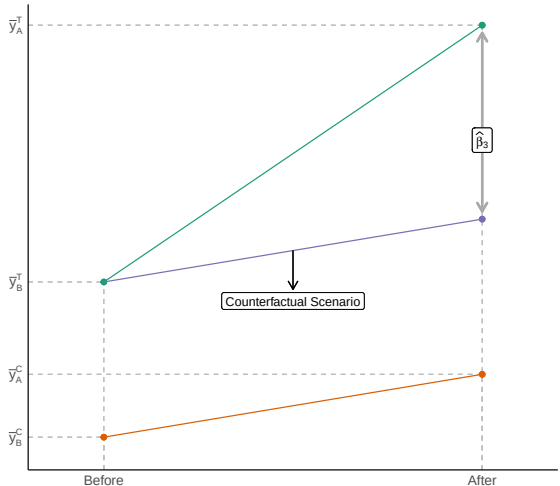
Replacing with the model's estimates:

$$\left(\overline{y}_{T,A} - \overline{y}_{T,B}\right) - \left(\overline{y}_{C,A} - \overline{y}_{C,B}\right) = \left(\hat{\beta}_2 + \hat{\beta}_3\right) - \hat{\beta}_2 = \hat{\beta}_3$$

Difference-in-Differences Estimator



Difference-in-Differences Estimator



Parallel Trends Assumption

The new purple line in the previous plot represents the counterfactual scenario, i.e., how we would expect the treated group to behave in the absence of treatment.
This is based on a key assumption: **the parallel trends assumption**.

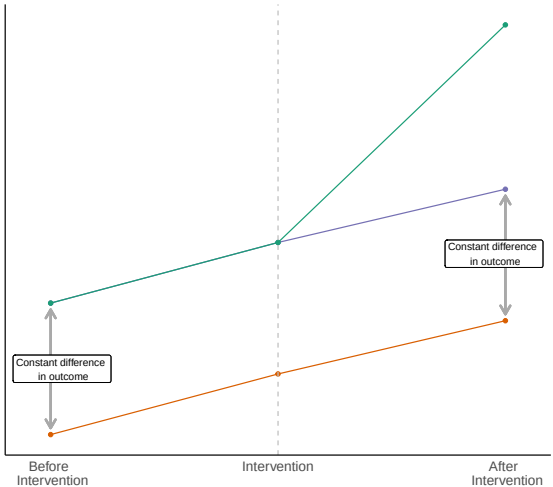
What is the parallel trends assumption

- Parallel trends assumes away the selection bias associated with comparisons
- The assumption is thought to be more plausible than simply assuming simple comparisons held equal

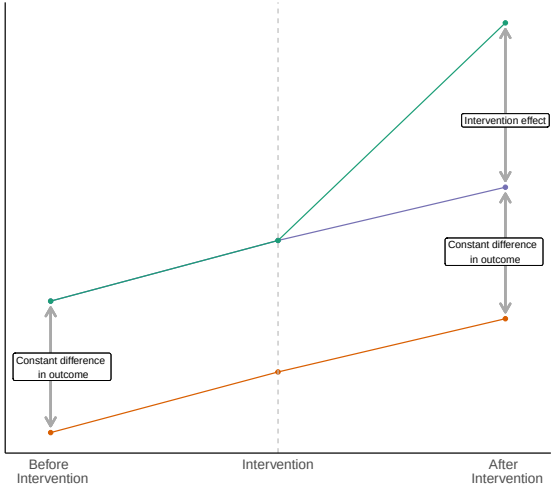
$$E[Y^0|D=0] = E[Y^0|D=1]$$

- But it is still a strong assumption

What is the parallel trends assumption



What is the parallel trends assumption?



What is the parallel trends assumption?

We do a visual inspection of the plots and/or test the difference in means in the periods before the intervention. If the parallel trends assumptions holds, then the differences in means between the control and treated group, before the intervention are zero.

Job Injury

Context

- In 1980, Kentucky raised its cap on weekly earnings that were covered by worker's compensation.
- We want to know if this new policy caused workers to spend more time unemployed.
- If benefits are not generous enough, then workers could sue companies for on-the-job injuries, while overly generous benefits could cause moral hazard issues and induce workers to be more reckless on the job, or to claim that off-the-job injuries were incurred while at work.

Job Injury - Variables

The main variables are:

- *durat*, duration of unemployment benefits, measured in weeks.
- *afchnge*, dummy variable indicating whether the observation happened before or after the policy change.
- *highearn*, indicator variable marking if the observation is a low (0) or high (1) earner. This is our group (or treatment/control) variable.

The control variables are:

- *age*, age of the individual
- *male*, dummy variable indicating whether the individual is male.
- *married*, dummy variable indicating whether the individual is married.
- *hosp*, dummy variable indicating whether the individual was hospitalized.

Job Injury - Model

The model is the following:

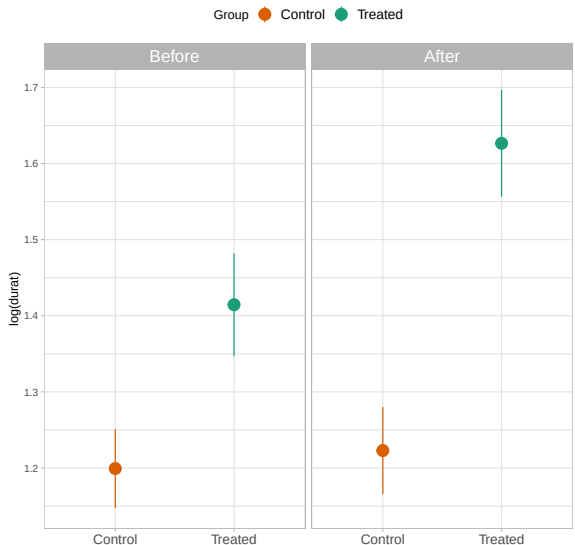
$$\log(durat_{it}) = \beta_0 + \beta_1 highearn_{it} + \beta_2 afchnge_t + \beta_3 (highearn_{it} \times afchnge_t) + \beta_4 age_{it} + \beta_5 male_i + \beta_6 married_{it} + \beta_7 hosp_{it} + u_{it} \quad (9)$$

Job Injury - Exploratory Analysis

We can start by assessing the averages for each group (control/treated), before and after the intervention:

```
library(wooldridge)
library(ggplot2)
ggplot(injury,
       aes(x = factor(highearn),
           y = log(durat),
           color = as.factor(highearn))) +
  stat_summary(geom = "pointrange", size = 1,
               fun.data = "mean_se", fun.args = list(mult = 1.96)) +
  scale_x_discrete(breaks=c(0,1), labels=c('Control', 'Treated')) +
  facet_wrap(.~afchnge,
             labeller = labeller(afchnge = c('0'='Before', '1'='After'))) +
  scale_color_manual(values = c('#d95f02', '#1b9e77'),
                     labels=c('Control', 'Treated')) +
  labs(x = "", y = "log(durat)", color = 'Group')+
  theme_light() +
  theme(legend.position = "top")
```

Job Injury - Exploratory Analysis



Job Injury - Estimate model

Next, we will estimate the model:

```
did_estim <- lm(log(durat) ~ highearn + afchnge +
                highearn * afchnge + male + married + age + hosp,
                data = injury)
```

Job Injury - Estimate model

```
##
## Call:
## lm(formula = log(durat) ~ highearn + afchnge + highearn * afchnge +
##     male + married + age + hosp, data = injury)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9100 -0.8257  0.0846  0.7451  4.4303
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.667737   0.057168  11.680  < 2e-16 ***
## highearn          0.117468   0.043597   2.694  0.00707 **
## afchnge           0.051957   0.037166   1.398  0.16217
## male             -0.072089   0.037881  -1.903  0.05707 .
## married           0.033814   0.033759   1.002  0.31656
## age               0.008904   0.001221   7.292  3.4e-13 ***
## hosp              1.084683   0.032921  32.948  < 2e-16 ***
## highearn:afchnge  0.168276   0.059219   2.842  0.00450 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.195 on 6836 degrees of freedom
##   (306 observations deleted due to missingness)
## Multiple R-squared:  0.1597,Adjusted R-squared:  0.1589
## F-statistic: 185.7 on 7 and 6836 DF,  p-value: < 2.2e-16
```

Job Injury - Parallel trends

Now, the important question is: is this valid?

Since we don't have information for more time periods before the intervention, we cannot assess the parallel trends assumption.

Medicaid Expansion

Context

- We will examine the effects of Medicaid expansions (under the Affordable Care Act) on insurance coverage.
- We want to know if this new policy caused an increase in the share of low-income childless adults with health insurance in the state.

Medicaid Expansion - Variables

The main variables are:

- *dins*, share of low-income childless adults with health insurance in the state.
- *after*, dummy variable indicating whether the observation happened before or after the policy change.
- *treat*, indicator variable marking if the observation is a low (0) or high (1) earner. This is our group (or treatment/control) variable.

In this setting, we will have state-level data, where a group of states was never treated, and another group had the medicare expansion in 2016.

Medicaid Expansion - Model

The model is the following:

$$dins_{it} = \beta_0 + \beta_1 treat_{it} + \beta_2 after_t + \beta_3 (treat_{it} \times after_t) + u_{it} \quad (10)$$

Medicaid Expansion - Data

We will start by retrieving the data:

```
library(tidyverse)

df <- haven::read_dta("https://raw.githubusercontent.com/Mixtape-Sessions/Advanced-DID/main/Exercises/Data/ehec_data.dta")
df_new <- df %>%
          #filter so first year in the sample is 2013
          filter(year >=2013) %>%
          #filter to include control group and treated group (in 2016)
          filter(is.na(yexp2) | yexp2 == 2016) %>%
          #Create treat and after variables
          mutate(treat = ifelse(is.na(yexp2), 0, 1),
                 after = ifelse(year<2016, 0, 1)) %>%
          select(stfips, year, dins, treat, after)
```

Medicaid Expansion - Estimate model

Next, we will estimate the model:

```
library(fixest)
model <- feols(dins ~ treat + after + treat*after,
               cluster=~stfips, data = df_new)
summary(model)
```

Medicaid Expansion - Estimate model

```
## OLS estimation, Dep. Var.: dins
## Observations: 126
## Fixed-effects: stfips: 18
## Standard-errors: Clustered (stfips)
##              Estimate Std. Error t value   Pr(>|t|)
## after        0.042478   0.003387 12.5413 5.1089e-10 ***
## treat:after 0.069530   0.005124 13.5697 1.5017e-10 ***
## ... 1 variable was removed because of collinearity (treat)
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 0.026436     Adj. R2: 0.741147
##                   Within R2: 0.51232
```

Medicaid Expansion - Estimate model

The results indicate that the policy caused an increase of 0.07 percentage points increase in the share of low-income childless adults with health insurance.

Medicaid Expansion - Parallel trends

Now, the important question is: is this valid? Let's check the parallel trends assumption:

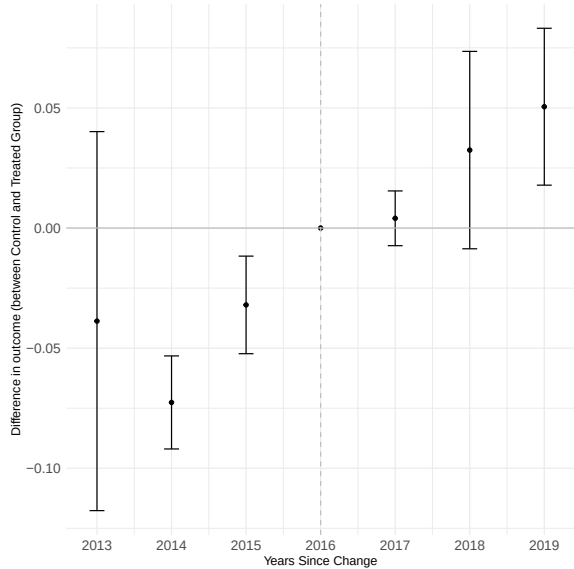
```
library(fixest)
#Estimate event-study model
model <- feols(
  dins ~ i(treat, i.year, ref = 0, ref2 = 2016) + i(year) + i(treat) | stfips,
  data = df_new, cluster = ~stfips,
)

# Extract event-study coefficients
coefs_event <- broom::tidy(model) %>%
  # Select only event-study terms
  filter(str_detect(term, "treat::1:year")) %>%
  # Create year column
  mutate(
    year = as.numeric(str_extract(term, "treat::1:year::(.*)", group = 1))
  ) %>%
  mutate(
    conf.low = estimate - 1.96 * std.error,
    conf.high = estimate + 1.96 * std.error
  ) %>%
# Add 0 for omitted year 2015
  add_row(
    term = "treat::1:year::2016",
    estimate = 0,
    std.error = 0,
    year = 2016,
    conf.low = NA,
    conf.high = NA
  )
```

Medicaid Expansion - Parallel trends

Now, the important question is: is this valid? Let's check the parallel trends assumption:

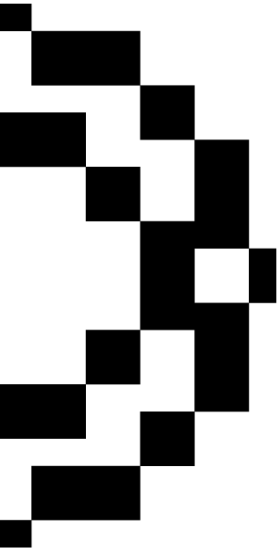
```
ggplot(coefs_event, aes(x = year, y = estimate)) +
  geom_point() +
  geom_errorbar(aes(ymin = conf.low, ymax = conf.high), width = 0.2) +
  geom_vline(xintercept = 2016, linetype = "dashed", color = "grey") +
  geom_hline(yintercept = 0, color = "grey") +
  scale_x_continuous(breaks = 2013:2019) +
  labs(x = "Years Since Change",
       y = "Difference in outcome (between Control and Treated Group)") +
  theme_minimal()
```

Medicaid Expansion - Parallel trends



Medicaid Expansion - Parallel trends

The parallel trends assumption does not hold. The conclusions are not valid.



Advanced topics

Remarks on parallel trends assumption

There are some issues with testing for pre-trends:

1. Even if pre-trends are exactly parallel, this will not guarantee that the post-treatment parallel trend assumption is satisfied.
2. The pre-trends test has low power, which could lead to failing to reject H_0 when it is false, i.e., stating that the parallel trend assumption holds, when it does not.
3. Conditioning the analysis on "passing" a pre-trends test induces a selection bias, known as pre-trend bias.

Remarks on parallel trends assumption

Recommendations:

1. Understand the context of your study: What is the expected behavior of the outcome variable? What are possible confounders?
2. Check for plausible conditional parallel trends assumption, by including covariates in the model. This requires other estimators such as Regression Adjustment, Inverse Probability Weighting and Doubly Robust estimators (outside the scope of this course).
3. Considering the context-specific knowledge, perform robustness checks and sensitive analysis (see Rambachan and Roth (2022)).

Staggered Designs

So far, we assumed that the treatment occured at the same time. How about if the units are treated with diferential timing?

Some new methods:

Estimator	**R** packages
Goodman-Bacon (2021)	bacondecomp
Callaway & Sant'Anna (2020)	did
De Chaisemartin & D'Haultfoeuille (2020)	DIDmultiplegtDYN
Borusyak, Jaravel & Spiess (2021)	didimputation
Sun & Abraham (2020)	sunab

Table 2: Estimators and corresponding packages