Causal inference with regression

Piercesare Secchi

Milano, 25 February 2025

Main references for the talk

- Hal Varian (2016), Causal Inference in Economics and Marketing, PNAS, 113 (27) 7310-7315
- Andrew Gelman, Jennifer Hill and Aki Vehtari (2021), Regression and other stories, Cambridge University Press, Cambridge. **See in particular their chapters 18, 19 and 20.**
- Most, if not all, the original ideas explored in this literature are due to Donal Rubin, Guido Imbens, Joshua Angrist. The last two shared the Nobel prize for Economics in 2021.

See the book:

Imbens G, Rubin DL (2015) *Causal Inference in Statistics, Social and Biomedical Sciences: An Introduction*. Cambridge University Press, New York.

Causal inference: intro

From Gelman et al. (2021):

- The problem: what *would happen* to an outcome variable Y, as a result of a treatment (intervention, exposure) Z, given the pre-treatment information X.

- Conceptualization: a comparison between different potential outcomes of what *might have occured* under different scenarios

- In most cases the *comparison* is between a *factual state* - what did happen - and one or more *counterfactual states* - representing what might have happened.

- The comparison could also be among various counterfactuals.

Causal inference: intro

- To simplify, suppose the treatment Z can take only two values 0 or 1.

- E.g. the treatment is a drug: Z=1 means drug received, Z=0 means drug NOT received.

- The goal is to quantify the causal effect of the treatment. In the example: the causal effect of receiving the drug or of not receiving the drug

- A sample of units from a population enter in an experiment (we will later see how): some units receive the treatment, some units do not receive the treatment.

- Consider the two potential outcomes on unit *i*:

$$Y_i^1 = outcome\ if\ the\ unit\ receives\ the\ treatment\ (Z = 1)$$
$$Y_i^0 = outcome\ if\ the\ unit\ does\ NOT\ receive\ the\ treatment\ (Z = 0)$$

- The causal effect of the treatment on unit *i* is:

$$\tau_i = Y_i^1 - Y_i^0$$

Causal inference: the fundamental problem

- For the unit *i* which received the treatment (Z=1), the researcher observes (*factual*) Y_i^1 and the *counterfactual* is Y_i^0;

- For the unit *i* which did NOT receive the treatment (Z=0), the researcher observes (*factual*) Y_i^0 and the *counterfactual* is Y_i^1.

- For the units of the population who were not part of the sample, both outcomes are *counterfactuals*.

- For the unit *i in the experiment*, the *observed outcome* is the ***potential outcome revealed*** by the treatment:

$$Y_i = Z_i \, Y_i^1 + (1- Z_i \,) \, Y_i^0$$

- It is often, if not *always*, impossible to observe both potential outcomes Y_i^0 AND Y_i^1 on the same unit *i* (either in the sample or in the population). *This is the fundamental problem of causal inference*

Average causal effects

- It is often, if not *always*, impossible to observe both potential outcomes Y_i^0 AND Y_i^1 on the same unit *i* (in the sample or in the population).

- Almost all causal strategies make use of *comparison across groups of units*: one or more groups that were exposed to treatment and one or more groups that were not.

- The focus is NOT on the individual causal effect on unit *i*

$$\tau_i = Y_i^1 - Y_i^0$$

  which is inaccessible, but on an **average causal effect $\boldsymbol{\tau}$**

- Sample Average Treatment Effect (SATE), Conditional Average Treatment Effect (CATE) or Population Average Treatment Effect (PATE), are the keywords.

A motiviating example: why we cannot use regression in a naive way

From Varian (2016):

- Suppose you are given some data on **ad spend** and **product sales** in various cities and are asked to predict how sales would respond to a contemplated change in ad spend.

- If Y_i denotes per capita sales in city *i*, and X_i denotes per capita ad expenditure in city *i* , it is tempting to run a simple linear regression of Y on X, and interpret the slope β of the regression line as the increase in sales corresponding to a unit increase in ad expenditure.

- But will **forcing** X→X+1 **cause** Y→Y+ β ?

- Suppose that the sales Y are per capita box office receipts for a movie about *surfing* and X are per capita TV ads for that movie. There are only two cities in the data set: Honolulu, Hawaii and Fargo, North Dakota.

A motiviating example: why we cannot use regression in a naive way

- Advertiser spent 10 cents per capita on TV advertising in Fargo and observed $1 in sales per capita, while in Honolulu the advertiser spent $1 per capita and observed $10 in sales per capita. Hence the model Y = 10X fits the data perfectly.
- Do you really believe that increasing per capita ad spend in Fargo to $1 would result in box office sales of $10 per capita?
- An omitted confounder: *interest in surfing*. Interest in surfing is high in Honolulu and low in Fargo.
- Marketing executives that determine ad spend presumably know this, and they choose to advertise more where interest is high and less where it is low. So this omitted variable—interest in surfing—affects both Y and X.
- Note: the regression is fine for prediction of sales based on observed ad expenditure, but not for quantifying the causal effect on sales of an increase of ad expenditure.
- **Correlation is not causation…**

Controlled randomized experiments

- Idea: to find out what happens when you change something, it is necessary to change it (Box, Hunter & Hunter (2005), *Statistics for Experimenters*, Wiley)

- In a completely randomized experiment, the probability of being assigned to the treatment is the same for each unit in the sample

- By randomly assigning the units in the sample to treatment (Z=1) or to control (Z=0) the researcher ensures that there are no differences in the (expectation of) the distribution of potential outcomes between groups receiving different treatments.

- I.e. the treatment assignment Z is independent of the potential outcomes (***ignorability condition***)

Controlled randomized experiment: clinical trials

- In a classic clinical trial, one applies a treatment to a sample of subjects and observes some outcome.
- The outcomes for the *treated* subjects (*treatment group*) can be compared to the outcomes for the untreated subjects (*the control group*).
- To determine the *causal effect* of the treatment on the subjects, subjects are allocated *at random* to the treatment or to the control group.
- To warrant the same (two) distributions for potential outcomes Y^0 and the same (two) distributions for potential outcomes Y^1 in the treatment and control groups, **neither the participants nor the researcher** know which treatment the participants are receiving until the clinical trial is over (*double blind experiment*)
- E.g.: subjects in the treatment group receive a drug while subjects in the control group receive a placebo, whose physical appearence, color, etc are identical to the drug.

Controlled randomized experiments: the analysis by means of regression

- When the experimental design yelds no (average) differences between groups in terms of potential outcomes, a simple difference in means of the observed outcomes will yeld an unbiased estimate of the average treatment effect.

- A regression on the treatment indicator Z, adjusting for pre-treatment predictors X, will also be unbiased and has the potential to reduce the variance of the estimator.

- Randomized experiments are considered the *gold standard* of causal inference, and many different designs for randomizations have been studied in the literature: Completely Randomized Experiments, Randomized Block Experiments, Factorial Design… (See DOE theory).

A complex example: the ARMD Trial

- Consider again the Age-Related Macular Degeneration Trial (ARMD) explored at length in the Applied Statistics course of the first semester for illustrating LMM.

- The ARMD data arise from a randomized multi-center clinical trial comparing an experimental Treatment (high dose of alpha-interferon) with a Placebo for patients diagnosed with ARMD.

- The outcome Y is Visual Acuity measured on each patient participating in the trial at baseline and at four post-randomization time points (longitudinal data)

- The analysis can be conducted with different LM (fixed effects) and LMM (fixed and random effects), including among the predictors the variable Z indicating allocation to the treatment or to the control group.

Controlled randomized experiment

- Results of controlled randomized experiments are less likely to be affected by factors that are not related to the treatment (or the intervention) being tested.

- If the goal is to measure the ***impact of the treatment on a population***, one should sample at random the subjects out of the population, before allocation to the treatment or to the control group.

- Controlled experiments are often costly and in many important cases are not feasible. As an example, consider the impact of education on income. An ideal experiment would require randomly selecting the amount of education students acquire, which would be rather difficult and probably unethical. Besides: double-blinding the experiment would be impossible.

Observational studies: when controlled experiments are not feasible

- In an observational study the reseacher draws inferences from a sample selected without the investigator’s control because of ethical concerns or logistical constraints.

- An observational study often investigates the potential effects of a treatment on units, where the assignment of units to either the treatment or control group is beyond the investigator's control.

- One might be interested in how the treatment affected those units which were actually treated, but were not necessarily randomly chosen for treatment. That is, the focus is on the ***impact of the treatment on the treated*****.**

- If a controlled experiment is not feasible, there are usually two possible ways out:
  - *selection on observables*, the researcher attempts to build a predictive model of who received treatment.
  - *selection on unobservables*, the researcher attempts to find natural experiments that are as good as random and can overcome the confounding variable problem.

Basic identity of causal inference

Let’s consider again potential outcomes:

$$Y^1 = outcome\ if\ the\ unit\ has\ received\ the\ treatment\ (Z = 1)$$
$$Y^0 = outcome\ if\ the\ unit\ has\ NOT\ received\ the\ treatment\ (Z = 0)$$

And let’s indicate with:

$$Y^1_{Z=1}\ (observed)\ the\ outcome\ of\ a\ treated\ unit$$
$$Y^0_{Z=0}\ (observed)\ the\ outcome\ of\ a\ control\ unit$$
$$Y^0_{Z=1}\ (counterfactual)\ the\ outcome\ of\ a\ treated\ unit\ if the\ unit\ had\ not\ been\ treated$$
$$Y^1_{Z=0}\ (counterfactual)\ the\ outcome\ of\ a\ control\ unit\ if\ the\ unit\ had\ been\ treated$$

Basic identity of causal inference

We can consider this basic identity:

$Y^1_{Z=1}$ - $Y^0_{Z=0}$ $=$

= Outcome for treated – Outcome for untreated

= $[Y^1_{Z=1}$ - $Y^0_{Z=1}] + [Y^0_{Z=1}$ - $Y^0_{Z=0}]$

= [Outcome for treated – Outcome for treated if not treated] +

+ [Outcome for treated if not treated – Outcome for untreated]

= **Impact of treatment on treated + selection bias**

The identity shows the key critical issue for causality: comparing the actual outcome [the difference between what happened to the treated and what happened to the untreated] with the counterfactual [the difference between what happened to the treated and what would have happened if they had not been treated] that is the impact of treatment on treated.

Basic identity of causal inference

- As already noticed, we cannot observe what would have happened to the treated if they had not been treated, so we have to estimate that counterfactual some other way

- In a randomized trial, the distribution of the potential outcomes is expected to be the same in the treatment and in the control group. If the distribution of $Y^0_{Z=1}$ is the same as the distribution of $Y^0_{Z=0}$, the expected value of the selection bias is zero!

- Hence the observed difference between the outcome for the treated and the outcome for the untreated is an estimate of the causal impact of treatment on the treated.

Using regression for quantifying the treatment effect on the treated

- Consider an advertiser who wants to know how many extra sales are generated by an increase in ad expenditure.

- The treated subject is here the advertiser! The treatment being increase in ad expenditures.

- Of course, the advertiser might want to take the indirect route which tries to quantify how the consumers respond to an increase of ad expenditure.

- To quantify the impact of ad expenditure on consumers the advertiser might, in principle, run a controlled randomized experiment by selecting a treatment group and a control group among consumers.

- But to quantify the impact of increase in ad expenditure on the advertiser, the advertiser might also take a more direct route, which does not involve an expensive controlled experiment.

Using regression for quantifying the treatment effect on the treated

- The advertiser can increase ad expenditure *for a limited period of time* and then compare the outcome of that experiment to an estimate of the counterfactual—what would have happened during the limited period of time without that increase in ad expenditure.

- To generate the counterfactual, fit a model to the data **before the increase in ad expenditure** and use the model to **predict** what would have happened during the limited time period **after the ad expenditure increase**.

- In building the model used for prediction, one can use all the pre-treatment covariates X deemed to be necessary to explain the variability of the outcome variable Y.

- Be however careful with confounders, that is those variables which affect both (some) predictors X and the outcome variable Y, by including them among the predictors.

- Example: in the Christmas Holiday season both sales and ad expenditures increase. This will confound the impact of an increase of ad expenditure on sales. Hence include a dummy among the predictors to identify the Christmas Holiday season.

Using regression for quantifying the treatment effect on the treated

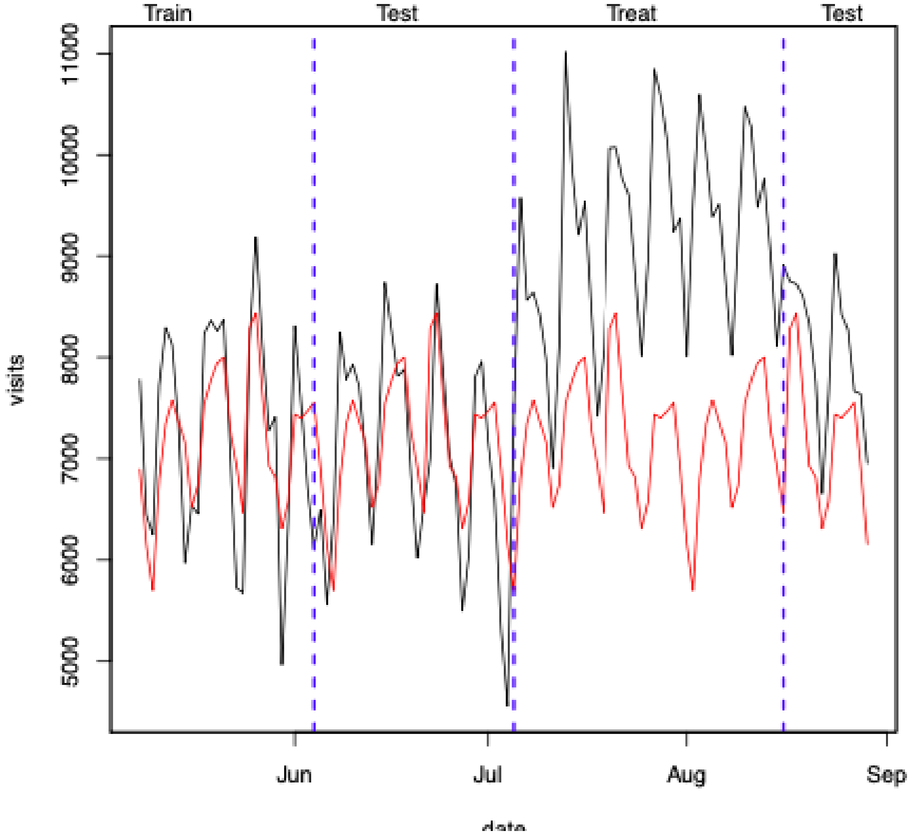


Fig. 1. Hypothetical train-test-treat-compare process. The model is estimated during the training period and its predictive performance is assessed during the test period. The extrapolation of the model during the treat period (red line) serves as a counterfactual. This is compared to the actual outcome (black line) and the difference is the estimated treatment effect. When the treatment is ended the outcome returns to something close to the original level .

Varian (2016)

Using regression for quantifying the treatment effect on the treated

We have used this idea to quantify reputational effects:

Arena, M., Azzone, G., Conte, A., Secchi, P., Vantini, S. (2014), Measuring downsize reputational risk in the oil & gas industry. In: Paganoni, A., Secchi, P. (eds.) Advances in Complex Data Modeling and Computational Methods in Statistics. Springer, Milan

See also:

F. Ieva, P. Secchi, S. Vantini (2015). Big data: the next challenge for statistics, *Lett Mat Int*, 3, 111-120.

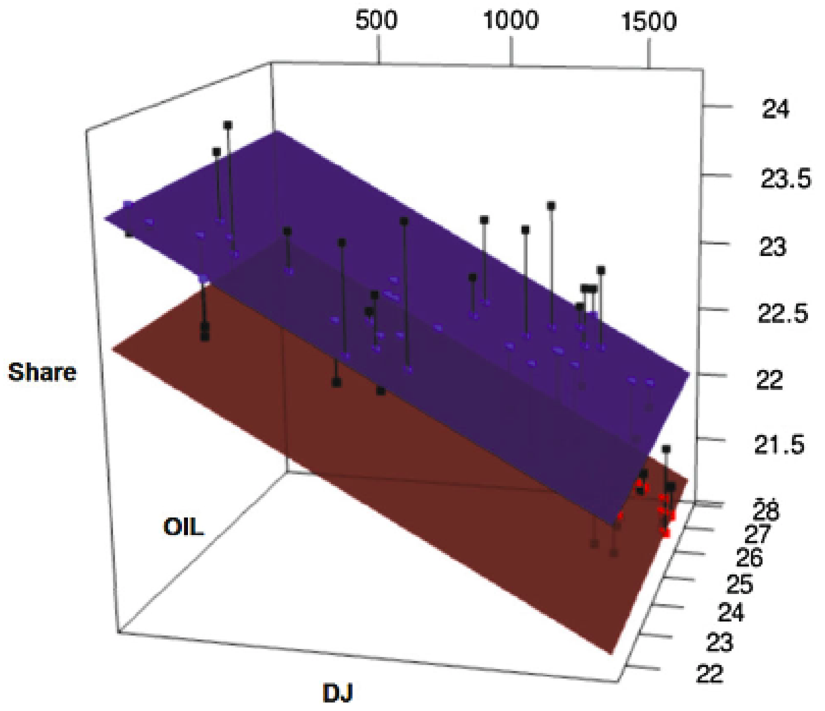


Fig. 9 Graphic representation of the model for estimating the reputational impact of a particular event, based on share price, the price of a barrel of oil and the market index (DJ). The *blue plane* indicates the model estimated before the event, and the *red* one the model estimated the days following the event. The vertical translation between the two planes corresponds to the estimated absolute loss of value caused by the reputational event (colour figure online)

Regression discontinuity

- In some cases, the subjects receiving the treatment are those who pass a given threshold

- Example: we want to evaluate the effect of a tutoring program on Calculus on first-year students of Engineering at a certain University. A controlled experiment is not feasible, also because the University administrators want to prioritize access to the program for the students who are struggling the most academically. So the tutoring service is open only to students whose grade in the mid-term Calculus exam is below 17/30.

- In this cases, observations close to, but just below, the threshold should be similar to those close to, but just above, the threshold. Comparing subjects close to the threshold, but on different sides, becomes therefore appealing.

Regression discontinuity

- In the example: the students with a mid-term grade just below 17/30 are the treatment group, those with a grade just above 17/30 are the control group.

- Note: given the uncertainty inherent in the measurement of the mid-term grade, assignment to treatment or to control is (almost) random. I.e. the investigator expects the potential outcomes in the two groups to have the same distributions (**ignorability condition**).

- Indeed the experimenter could add a small amount of random error to the score to be thresholded, to make the experiment even more similar to a randomized controlled experiment

Regression discontinuity: an example taken from Carpenter C, Dobkin C (2011) and reported in Varian (2016)

Carpenter C, Dobkin C (2011) The minimum legal drinking age and public health. *Journal of Economic Perspectives* 25(2):133–156.

The paper examines the impact of the minimum legal drinking age on mortality in US.

There is a major jump in motor vehicle accidents at the age of 21. Someone who is 20.5 years old isn’t that different from someone who is 21 years old, on average, but 21 year olds have much higher death rates from automobile accidents, suggesting that the minimum drinking age causes this effect.

Use a regression model with a dummy for age above or below the threshold 21, to test if the effect is significant (ANCOVA)

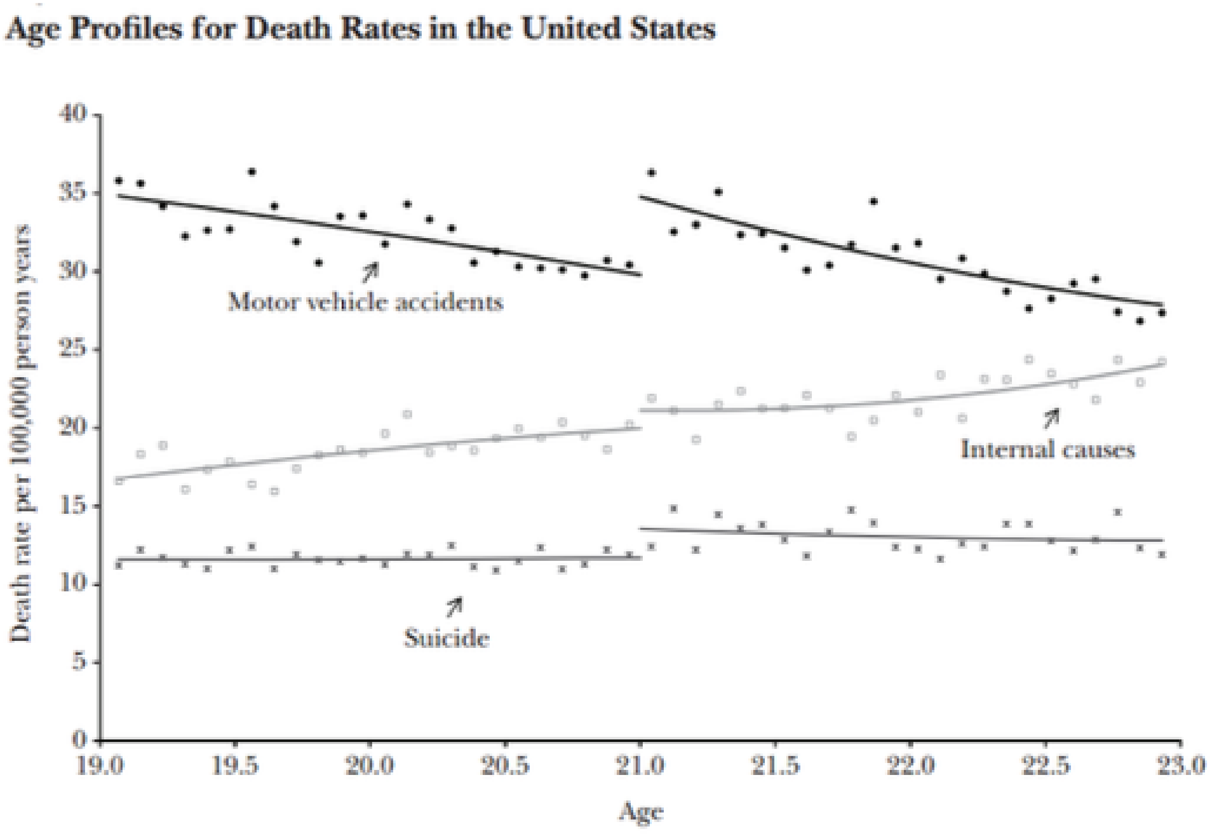


Fig. 2. Death rates by age by type of death.

Longitudinal data: difference-in-difference models

- In some cases data are longitudinal, that is each unit is observed over time.

- As an example, Varian (2016) considers data on advertising expenditures across DMAs (Designated Marketing Areas). Prior to a campaign, there is zero ad spending, while during the campaign, there is ad spending at some level **in certain DMAs but not in others**. The outcome Y is sales.

- The Treatment is ad expenditure (Z=1) or no ad expenditure (Z=0), but now we observe the outcome over time, that is before (B) or after (A) the time of Treatment. Therefore:

$Y^1(B)$ and $Y^1(A)$ are the potential outcomes of a treated DMA (Z=1), before and after the time of Treatment, respectively;

$Y^0(B)$ and $Y^0(A)$ are the potential outcomes of a DMA not treated (Z=0) before and after the time of Treatment, respectively.

Longitudinal data: difference-in-difference models

- We are interested in estimating the impact of the treatment on the treated, that is :

$$Y^1_{Z=1}(A) \; - \; Y^0_{Z=1}(A)$$

the difference between the actual outcome for the treated and the (unobserved) counterfactual – what the outcome would have been if the unit had not been treated.

- The model assumes that:

$$Y^0_{Z=1}(A) \text{ - } Y^0_{Z=1}(B) = Y^0_{Z=0}(A) \text{ - } Y^0_{Z=0}(B)$$

i.e. that the difference in potential outcome for no-treatment, before and after the time of treatment, is the same in the treatment group and in the control group.

- The assumption implies that:

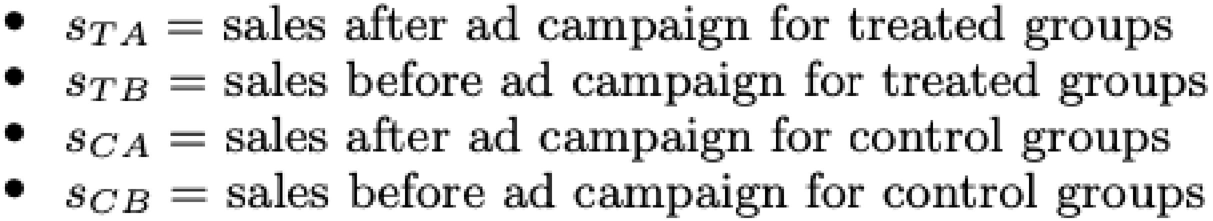
$$Y^1_{Z=1}(A) \; - \; Y^0_{Z=1}(A) = \; [Y^1_{Z=1}(A) - Y^0_{Z=1}(B)] \text{ - } [Y^0_{Z=0}(A) \text{ - } Y^0_{Z=0}(B)]$$

i.e. a difference in differences.

Difference in difference: Varian (2016) example

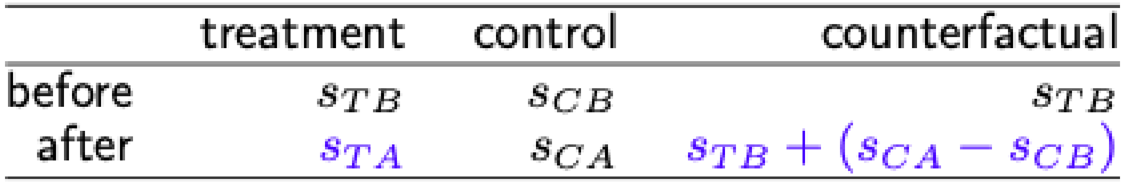
- Going back to the Varian (2016) example, suppose that some DMAs were exposed to an ad (treated), and some were not.

- Let



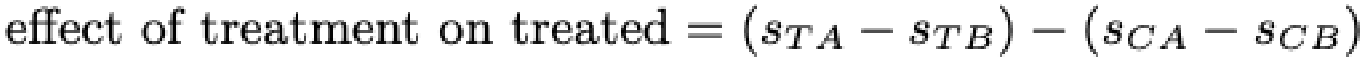
- s_{TA} = sales after ad campaign for treated groups
- s_{TB} = sales before ad campaign for treated groups
- s_{CA} = sales after ad campaign for control groups
- s_{CB} = sales before ad campaign for control groups

- Then



	treatment	control	counterfactual
before	s_{TB}	s_{CB}	s_{TB}
after	s_{TA}	s_{CA}	$s_{TB} + (s_{CA} - s_{CB})$

and



$$\text{effect of treatment on treated} = (s_{TA} - s_{TB}) - (s_{CA} - s_{CB})$$

- Regression models are allowed, adding predictors such as weather, news events, and other exogenous factors of this sort which impact sales in addition to the ad expenditure.

Additional topics: natural experiments

- In some lucky cases, «nature» provides treatment and control groups which are as good as if they were generated by randomization.

- Varian (2016) illustrates The Super Bowl example.

- It is well known that the home cities of the teams that are playing the Super Bowl have an audience about 10-15% larger than cities not associated with the teams playing.

- It is also well known that companies that advertise during the Super Bowl have to purchase their ads months before it is known which teams will, actually, be playing.

- The combination of these two facts implies that two essentially randomly chosen cities will experience a 10-15% increase in ad impressions for the movie titles shown during the Super Bowl. If the ads are effective, we might expect to see an increase in interest in those movies in the treated cities, as compared to what the interest would have been in the absence of a treatment.

Additional topics: natural experiments

- The outcome, interest, is measured in two ways:

a) the number of queries on the movie title for all the movies;
b) the opening weekend revenue, which could be obtained only for a subset of the movie titles.

- Data for the cities whose teams are not playing are used to estimate the boost in query volume after being exposed to the ad as compared to before.

- This is used to estimate the counterfactual: what the boost would have been without the 10-15% additional ad impressions in those cities associated with the home teams.

Additional topics: data restructuring

From Gelman et al. (2021):

- *Matching* refers to any procedure that *restructures* the original data sample to generate an *analysis sample* which looks like it was created from a randomized experiment. These methods are attracting increasing interest due to the availability of big dataset

- We would like the matched groups (treatment and control) to exhibit balance and overlap with respect to the pre-treatment variables considered as confounders.

- The goal is to create ***by restructuring and not by randomization*** two groups, identified by Z=1 or Z=0, such that the potential outcomes Y^0 and Y^1 are expected to have the same distributions in the two groups (**ignorability**).

Additional topics: data restructuring

- Restructuring may remove some observations or may upweight the contribution of others

- If the goal is to estimate the effect of the treatment on treated, for each treatment unit we might find a control unit that we deem most similar to the treatment unit according to some metric, and retain this unit as a match. Unmatched control units are discarded.

- **Propensity score matching** is a common strategy to perform data restructuring.

Propensity score matching: 5 steps procedure

1. *Identification of the confounders*. These are the covariates that predict both the treatment and the outcome. We need to condition on the confounders to satisfy ignorability. *Use your engineering prior knoweldge, to identify confounders among the available covariates!*

2. *Estimate the propensity score*. For instance, if the goal is to estimate the effect of treatment on the treated units, fit a logistic model to predict the probability of belonging to the treatment group using as predictors the confounders identified at step 1.

3. *Matching to restructure data*. Using the propensity score as a summary, for each treated unit find a match from among the untreated units, by choosing one with the closest propensity score (Alternatives: with replacement or without replacement)

Propensity score matching: 5 steps procedure

4. *Diagnostic for balance and overlap*. Compare the distributions of each potential confounder across treatment and matched control group. Use tests, or simply plot absolute difference in means… Check overlap by plotting the histogram of the propensity scores before matching, i.e. for the original data, and after matching, i.e. for the restructured data

Repeat 2 to 4, for instance changing the propensity score model, until adequate balance and overlap is reached.

5. *Fit a regression model including Z* and and all potential confounders as predictor using the restructured data

Instrumental variables

When the ignorability assumption for the treatment assignment seems to be weakly tenable, there might be another variable W, an instrument, that might be assumed to be randomly assigned to the units of the sample.

If the instrument W is predictive of the treatment Z, then one could use it to identify the causal effect of Z, if W only affects Z but not the outcome Y (**exclusion restriction**)

For more details, see Chapter 21 of Gelman et al. (2021)

Instrumental variables: an informal example

Example: Y = Health status (target), Z=smoking habit (treatment), W=encouragment to quit smoking (instrument).

If W (randomly assigned) is positively correlated with Y, this can happen only through its effects on smoking, supporting the conjecture that smoking causes poor health. Hence if we were to fit the model

$$Y = \beta_0 + \beta_1 Z + \beta_2 W + \ error$$

we would conclude (test) that β_2 is zero.

We can thus consider the model:

$$Y = \beta_0 + \beta_1 Z + error \qquad (1)$$

$$Z = \gamma_0 + \gamma_1 W + error \qquad (2)$$

from which we obtain

$$Y = \beta_0 + \beta_1(\gamma_0 + \gamma_1 W) + error = \beta_0 + \beta_1\gamma_0 + \beta_1\gamma_1 W + \mathrm{error} \qquad (3)$$

- The estimate of $\beta_1\gamma_1$, obtained by fitting model (3), captures the causal effect of W on Y (W is randomly assigned)
- The estimate of γ_1, obtained by fitting model (2), captures the causal effect of W on Z (W is randomly assigned)
- By taking the ratio of the two estimates, one get an estimate of the causal effect of Z on Y.

What I didn’t talk about – Causal Graphs and Causal Calculus

- *Causal Graphs and Causal Calculus aka do-Calculus.*

See the works by Judea Pearl and in particular

- J. Pearl, M. Glymour, N.P. Jewell (2016), *Causal Inference in Statistics*, Wiley

- Pearl, J., & Mackenzie, D. (2019). *The book of why*. Penguin Books.