1 Causality in Data Science

You have studied how OLS controls for the effects of other variables on y. Does this mean that the effect of some regressor x_j on y when estimated by OLS can be given a causal interpretation? Before we address this issue we will first attempt to define what we mean by causality.

We start with a definition of causality borrowed from the experimental sciences. An effect is said to be causal if it is the result of a controlled experiment where everything else is the same and only the "treatment dosage" changes among observations. We can then be sure that the observed effect is a result of the different treatments and from nothing else.

Suppose you want to check if a new drug improves health. We are interested in the causal effect of the drug since we want to be sure that it is the drug affecting health outcomes and not other things that change among patients taking the drug. You run an experiment where you have 100 identical individuals and you give the drug to 50 of them (the treated group). The effect of the drug is the difference between the average health outcome of the 50 individuals in the treated group and the 50 individuals in the control group (those who did not receive the drug). Let y^1 denote the health outcome when the drug is received and let y^0 denote the health outcome when the drug is not received. Then the effect of the drug is estimated by the difference in sample means between the treated and control groups

$$\bar{y}^1 - \bar{y}^0 \qquad \text{Average Treatment Effect (ATE)} \tag{1}$$

This difference estimates the causal effect of the drug because this difference cannot be driven by other things since by construction all the individuals are identical. This naive estimator of the treatment effect is correct because of the experimental design.

1.1 Random Assignment

Finding identical individuals is practically impossible so experiments are usually done in a different way. The key issue in an experiment is to assign the treatment (e.g., the drug) in a random way. This is called a randomized experiment, and we say that the treatment was randomly assigned.

We now have 100 possibly not identical individuals and we assign the drug randomly to 50 of them. Expression (1) still estimates the causal effect of the drug. Why? Random assignment ensures that receiving the drug is not systematically correlated with other variables affecting the health outcome (e.g., age, gender, health history, etc.). So even if the individuals differ between the two groups, random assignment ensures that the average characteristics of the treatment and control groups are the same (same average age, same proportion of males, same average health history, etc.). We say that the other characteristics are balanced. The only systematic difference between the treatment and control groups is the treatment itself and nothing else. This is what is needed to estimate the average causal effect of the treatment. That is, we estimate an "average" treatment effect and this is all that we can aspire to when individuals are not identical.

To be precise, we will never be able to estimate the causal effect for a particular individual because a particular individual is either treated or not treated and therefore we observe only one value of y for this person (i.e., the value when treated or when not treated). Having data on y before and after the experiment does not solve the problem unless we can control for all other things occurring between the two time periods that can affect y.

Ceteris Paribus

When comparing treated to control individuals we want to "hold all other factors (or characteristics) fixed" or "keep all other things equal". In a randomized experiment, this is always the case because the experiment is designed so as to ensure that all other factors are, on average, the same between the treated and control populations. The difference between treated and control average outcomes, equation (1) thus estimates an average causal effect because the average difference is the result of the different treatments and from nothing else. It is therefore easy to estimate the causal effect when treatment assignment is random. Indeed we could define causality as:

A causal effect can be defined as the effect obtained when the treatment is randomly assigned.

1.2 Non-Experimental Data

But if the treatment (e.g., the drug, the amount of education, etc.) is distributed or assigned in a non-random way, estimator (1) usually does not estimate the causal effect. When the assignment is not random, the treatment and control groups may differ in other characteristics (not only in the treatment). If these characteristics affect the outcome variable then the simple comparison of averages also picks up this effect, in addition to the causal effect.

Example 1. Effect of an R&D grant on R&D investment. Suppose large firms receive most of the R&D grants (grants are non randomly assigned) and that large firms also invest more resources in R&D than other firms (say because the "technological size" of the projects are larger). Thus, the difference between the mean R&D investment of the firms that received the R&D grant (the treated firms) and those that did not receive the grant (the control firms) overestimates the causal effect of the grant on R&D investment. Firm size is a characteristic that affects both the extent of R&D expenditures and the probability of receiving an R&D grant.

"Randomized experiments are unusual in data science." For ethical/moral reasons we do not randomly select people to attend school for different number of years. Nor do we give away government money randomly. The data available to us is non experimental and it usually comes from government agencies or private companies.

1.3 Correlation

Before we examine how data science deals with non-experimental data we clarify the difference between correlation and causality. Causality is not the same as correlation. Recall that the correlation between two random variables x_j and y is

$$\rho = \frac{Cov(x_j, y)}{\sqrt{V(x_j)V(y)}} \tag{2}$$

The correlation is the covariance between x_j and y normalized by the standard deviations. Simply finding a correlation between x_j and y is not enough to conclude that a change in x_j causes a change in y. It would be enough if x_j is randomly assigned. But correlation can be the result, for example, of a third factor c: That is, x_j is correlated with c and c is correlated with y. x_j and y might then be correlated, when c is not accounted for, but x_j does not necessarily cause y.

Example 2. Persons with higher intellectual ability (c) study more years (x_j) and also earn higher incomes (y). The data will show a positive correlation between number of years of education and income which does not necessarily reflect a causal effect from education to income. Because education is not randomly assigned, the positive correlation can be the result of other factors, such as **gender, ability, location, income**, etc., and the challenge is to understand whether it reflects, in addition, a causal effect.In this example, the correlation between y and x_j arises because of common factors affecting both y and x_j that are omitted from or not accounted for in the analysis. In short, omitted factors can give rise to a correlation that has nothing to do with causality.

Example 3. (Levitt, 1997) City-level data in the U.S. usually shows a positive correlation between police and crime. The estimated correlation between crime and police officers is 0.86. This positive correlation could be due to city size but even after controlling for population size we get a positive correlation between crime per capita and police officers per capita of 0.37.

This is an example of another problem we face when trying to uncover causal relationships. Both variables may be simultaneously determined and the observed data reflects their equilibrium values. We will later analyze in detail a demand and supply example where price and quantity are simultaneously determined. It is possible that a larger police force does indeed reduce crime but it is also likely that the level of crime affects the number of police officers assigned to a city. Thus, we observe a positive correlation in the data because cities with more crime have larger police forces. In short, simultaneity among the dependent and independent variables is another source of correlation among these variables that is not related to causality. Levitt (American Economic Review, 2002) shows that accounting for this simultaneity implies that the causal effect of police on crime is "negative" despite the positive correlation. Here also correlation does not equal causality.

As these examples show we should never infer causality from correlation, anything is possible. As mentioned before that we are interested in the causal connection between two variables because policy changes should be based on knowledge of causal relationships and not merely correlations. In addition, data science theories (or

any models in science) (implicitly) talk about causal relationships, not correlations, and therefore knowledge of the existence (and/or strength) of a causal relationship can serve to test theories and models. Also, in many instances data science theory may be ambiguous as to the effect of a policy change. For example, do higher taxes increase tax revenue? Do higher R&D grants increase company-financed R&D investments?. If we estimate the causal effect of a policy change we can evaluate the effectiveness of such policy change and make informed recommendations. For this it is crucial to know whether x_j causes y and not merely whether x_j and y are correlated. Unless causality can be established, the estimated correlation has little interest for data scientists.

If an estimated relationship can be given a causal interpretation then we can use the estimated causal effects to answer "what if" questions (what happens to R&D investment if R&D subsidies are increased, all other factors remaining unchanged? i.e. *Ceteris Paribus*) which is crucial for recommending and evaluating policies. One goal of the course is to understand under what conditions it is possible to estimate causal effects with non-experimental data, and how to do it.

2 Partial Effects

In dat science, we are interested in the change in the mean of y due to a change in x_j , holding all other relevant factors constant (partial effect of x_j on $\mathbb{E}(y|j_x)$)

$$\frac{\partial \mathbb{E}(y|x)}{\partial x_j} \tag{3}$$

for continuous x_j. (What if x_j is discrete?)

For example, in the model $\mathbb{E}(y|x_1, x_2) = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2^2$, the partial effect of x_2 on y is $\frac{\partial \mathbb{E}(y|x)}{\partial x_2} = \beta_2 + 2\beta_3 x_2$. One thing that people do not often realize is that, in this case, we are not that much interested in β_2 and β_3 per se because, by themselves, they do not tell us much. Our interest is in $\beta_2 + 2\beta_3 x_2$ and in tracing how this partial effect varies with x_2. If, for example, x_2 measures "size" then we would like to know whether the effect of x_2 on y differs with size. (The impact of x_2^2 in the regression.)

When x_j is discrete, partial effects are computed by comparing $\mathbb{E}(y|x)$ at different settings of x_j, holding all other variables fixed. For example, if x_j is a 0/1 binary (dummy) variable then its partial effect is the change in $E(y|x = x_0)$ when only x_j changes, say, from 0 to 1,

$$\mathbb{E}(y|x_1 = x_{10}, ..., x_j = 1, ..., x_k = x_{k0}) - \mathbb{E}(y|x_1 = x_{10}, ..., x_j = 0, ..., x_k = x_{k0})$$

Example 4. Suppose that $\mathbb{E}(y|x_1, x_2) = \beta_1 x_1 + \beta_2 x_2$, and that x_2 is a dummy variable. Then the partial effect of x_2 is

$$\mathbb{E}(y|x_1, x_2 = 1) - \mathbb{E}(y|x_1, x_2 = 0) = \beta_2$$

In this simple example the partial effect of x_2 does not depend on x_1.

Example 5. If the model now is $\mathbb{E}(y|x_1, x_2) = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$, the partial effect of x_2 is

$$\mathbb{E}(y|x_1, x_2 = 1) - \mathbb{E}(y|x_1, x_2 = 0) = \beta_2 + \beta_3 x_1$$

which dependes on x_1. The partial effect of x_1 is $\frac{\partial \mathbb{E}(y|x_1, x_2 = 1)}{\partial x_1} = \beta_1 + \beta_3$ or $\frac{\partial \mathbb{E}(y|x_1, x_2 = 0)}{\partial x_1} = \beta_1$, depending on the choice of x_2.

Example 6. Suppose x is just the dummy variable D for receiving a drug treatment and we are interested in $\mathbb{E}(y|D)$. We already know that the CEF (Conditional Expectation Function) is linear

$$\mathbb{E}(y|D) = \beta_1 + \beta_2 D$$

and the partial effect of D on y is β_2

$$\mathbb{E}(y|D = 1) - \mathbb{E}(y|D = 0) = \beta_2$$

which is the population version of the difference in means between the treated and control groups, equation (1).

This example shows that β_2 can be equivalently estimated by a simple OLS regression with a single (dummy) regressor (and intercept term). This would also be true if we were interested in $\mathbb{E}(y|D,c)$ and assume that $\mathbb{E}(y|D,c) = \beta_1 + \beta_2 D + c\lambda$

Note.. Sometimes the dependent variable is in (natural) logs, e.g., $\mathbb{E}(ln y|x) = \beta_1 + \beta_2 ln\, x_2 + \beta_3 x_3$ and the partial effect of interest is $\frac{\partial \mathbb{E}(ln\, y|x)}{\partial ln\, x_2}$ which is the elasticity of y with respect to x_2.

2.1 Are Partial Effects Causal?

We can always estimate the partial effect $\frac{\partial \mathbb{E}(y|x)}{\partial x_j}$ by OLS. (How?) But can the (estimated) partial effect always be interpreted as the causal effect of x_j on y? The partial effect of x_j on y can be given a causal interpretation if we control for all other things affecting both y and x_j : This is what the other $x's$ in the model are supposed to do. The difficult part in applied works is to know what the other things that need to be controlled are, and how to measure them! This is what we mean by specification of the data science model: the choice of regressors (and functional form) that will allow us to interpret the estimated partial effects as causal effects.

Data science models usually focus on the relationship of interest, say between y and x_1 but abstract from many other variables (call them x_2) that also affect y and may be correlated with x_1. What is wrong with estimating $\mathbb{E}(y|x_1)$ instead of $\mathbb{E}(y|x_1, x_2)$
Nothing is wrong with this .... it is just not interesting because the estimated effects will likely not be causal unless we can argue that x_1 is not correlated with x_2. This is not always easy to do.

For a causal interpretation we need to have a situation where the only systematic difference between those individuals treated (by x_j) and the non-treated is the treatment itself, and not other variables that also affect outcomes. This is guaranteed to happen in randomized experiments when x_j is randomly assigned to the observations. Recall that a causal effect can be defined as the effect obtained when the treatment is randomly assigned. Random assignment guarantees that the treatment (e.g., schooling) is uncorrelated with anything else. With non-experimental data, however, we need to make sure that there are no other average differences between the subjects. We do this by controlling for other factors. With non experimental data we interpret a partial effect as being causal when we can be sure that the treatment (regressor) was assigned as if in a randomized experiment.

Let us go over the "thought process" involved in specifying a data science model for the effect of R&D subsidies on R&D expenditures.

Example 7. R&D grants and investment in R&D We have non experimental data on y = company-financed R&D investment D = dummy variable for receiving an R&D grant. Let

$$\mathbb{E}(y|D) = \beta_1 + \beta_2 D + \beta_3 f_2 + \beta_4 \cdot \Sigma + \beta_5 \cdot E \ldots$$

be the relationship between R&D investment and R&D grants. Is it a causal relationship? The partial effect of a change in D from $D = 0$ to $D = 1$ is

$$\beta_2 = \mathbb{E}(y|D = 1) - \mathbb{E}(y|D = 0) \tag{4}$$

which can be estimated by the difference in sample means between treated and control groups, equation (1) or by an OLS regression of y on D (and constant term).

For this partial effect to be causal, we need to believe that the R&D grants received are unrelated to other characteristics of the firm that affect R&D investment such as firm size, industry, whether the firm exports or not, etc. If this is true, then β_2 represents a causal effect because the other characteristics affecting y are on average the same between the firms receiving the grant and those that did not receive the grant. It is as if D was assigned randomly to firms.

But if the R&D grant amount received is correlated with firm size, or with industry affiliation or with any other firm characteristic that also affects the dependent variable, then β_2 or its estimate will not be the causal effect of the R&D grant since they will also reflect the effect of other characteristics correlated with D that are also affecting R&D investment:

Suppose that larger firms (in terms of sales or employment) are more likely to receive R&D grants and that larger firms also do more R&D. In this case, an estimate of (4) will result in a (upward) bias estimate of the causal effect of the R&D grant on R&D investment.

Suppose now that instead of focusing on $\mathbb{E}(y|D)$ we would focus on an expectation that conditions also on other firm characteristics:

$$\mathbb{E}(y|D, x_2, ..., x_k)$$

where $x_2, ..., x_k$ are controls (firm size, industry, export status, etc.)

$\mathbb{E}(y|D, x_2, ..., x_k)$ is a multivariate regression function. Note that even if we are interested in the relationship between y and D only, we want to control for other variables that affect both D and y in order to be able to argue that we are estimating the causal effect of D on y. To be clear, even after controlling for $x_2, ..., x_k$ (which is a finite list of factors) we need to assume that D is not systematically related to other unobserved factors (i.e., not among $x_2, ..., x_k$) affecting y; for the partial effect to be causal. We can then run a regression of y on D, $x_2, ..., x_k$ and interpret the estimated partial effect as causal.

The general message here is that, when the regressor of interest is a choice made by a rational agent (e.g., years of education, number of patents, investment in R&D, labor force participation), we need to control for those factors affecting the agent's choice that also affect the dependent variable. Otherwise we will not be able to infer causality from the choice variable to the outcome variable. The problem is that many of these factors are unobserved. Many methodological advances are motivated by the need to deal with unobserved variables in econometric models.

3 Specification of the Data Science Model

The example in the previous section makes clear that our ability to interpret the partial effect as a causal effect depends on the specification of the CEF, i.e., on what other variables we are able to control for. This is the "tricky" part of doing data science.

How many or which xs are "enough"? What does "enough" mean? To answer these questions is to address a core issue in data science work, or the specification of the data science model.

Let us analyze this issue deeper using the standard wage model as an example. For simplicity, let us abstract from demographic characteristics (gender, origin, location, etc.) by assuming that our population is homogeneous in this sense. Let

$$\begin{aligned} y &= \log(\text{wages}) \\ x_1 &= s, \text{exp}, \text{exp}^2 \\ x_2 &= a \end{aligned}$$

where s are years of schooling, exp are years of on-the-job experience, and a is ability. We assume linear CEF's. We have two specifications of the model, with and without ability,

$$\mathbb{E}(y|x_1, x_2) = \mathbb{E}(y|s, exp, a) = \beta_1 + \beta_2 s + \beta_3 exp + \beta_4 exp^2 + \beta_5 a \quad (5)$$

$$\mathbb{E}(y|x_1) = \mathbb{E}(y|s, exp) = \pi_1 + \pi_2 s + \pi_3 exp + \pi_4 exp^2 \quad (6)$$

Note that we have used different notation for the parameters since the CEF is determined by a different joint PDF in each equation. Both specifications can be written in error form with an error that is mean independent of the regressors by construction.

$$y = \beta_1 + \beta_2 s + \beta_3 exp + \beta_4 exp^2 + \beta_5 a + u \qquad \mathbb{E}(u|s, exp, a) = 0 \quad (7)$$

$$y = \pi_1 + \pi_2 s + \pi_3 exp + \pi_4 exp^2 + v \qquad \mathbb{E}(u|s, exp) = 0 \quad (8)$$

Because of the error assumption, in both cases, regressing y against its own regressors gives OLS estimators which are unbiased and consistent estimators of the parameters of that model. The real question we face is this:

$$\text{Is } \beta \text{ or } \pi \text{ of interest?}$$

"Interesting" here means that the partial effect can be given a causal interpretation. Can we assert that controlling for experience only makes the allocation of schooling years random? The "short" model does not control for natural ability while the "long" one does. Is it more believable to interpret β_2, rather than π_2; as a causal effect because the long regression controls for ability, which is deemed very important both in the determination of years of education and wage? (Then, how to measure ability?)

We know there are a lot of other factors that affect education and also affect wages (e.g., family background, type of education received, occupation, etc.) so that, in principle, we can start with a richer model and then argue that β_2 above is also not causal. This is precisely the main issue that we have to tackle when examining an empirical study. We have to believe that we have controlled for enough factors so as to interpret the partial effects as causal effects. This is what is meant by a correct specification of the model.

The problem of equation (8) is that the data science model says that the conditional expectation of interest is $\mathbb{E}(\ln w|s, exp, a)$ but in fact we estimate $\mathbb{E}(\ln w|s, exp)$. In general, the parameters in $\mathbb{E}(\ln w|s, exp)$ differ from the parameters in $\mathbb{E}(\ln w|s, exp, a)$, so it should not be surprising that regressing y on $(1, s, \exp, \exp^2)$ does not estimate β, the parameter of interest.

The moral of this example is that once we specify the CEF, the error will be mean independent of the regressors and therefore we will estimate the parameters of the CEF correctly. The question is whether those parameters (partial effects) are interesting. And this depends on what you "put" in the CEF. In other words, when regressing y on x we always estimate some partial effects, but are these partial effects causal effect? That is, are we estimating β or π?

It should be clear that we do not need data on all variables affecting y (this would be impossible), but we require that the x's being used in the regression were determined in a way which is uncorrelated with other unobserved factors affecting the dependent variable y (and implicitly embedded in the error term). This depends on which x's we include in the regression. In other words, the interpretation of the partial effects as causal effects relies on how well the model is specified, i.e., on the choice of x's in $\mathbb{E}(y|x)$.

4 Treatment effects

As discussed, in data science (and other social sciences), we are interested in the causal effect of some variable (treatment) x on an outcome variable y. For example, we are interested in the effects of education on earnings, or in the effect of a job training program on labor force participation.

The common problem in the analysis of all these questions is that data scientists usually cannot run laboratory experiments where individuals are randomly assigned to receive the treatment. We cannot randomly assign a given fraction of the population to go to college; people will choose whether or not to go to college by their own decision. They will choose to participate in a training program depending on their costs and returns of the various alternatives. Most likely, each person has his/her own cost-benefit analysis, and outcome must be heavily depending on his/her track records. Thus, simply comparing the mean outcome of treated and non-treated individuals will not reveal the causal effect of the treatment. This is the fundamental challenge one must face when doing empirical work. Data Science tools have been developed to allow us, under additional assumptions, to infer causal effects using non-experimental data.

To examine this issue it is useful to use the "potential outcomes" framework.

4.1 Potential outcomes

Assume for simplicity that there are two possible treatments, 0 (the control group), and 1 (the treatment group). We will assume that every individual in the sample has two potential outcomes:

y_{i1} : outcome if treated
y_{i0} : outcome if not treated

If the treatment is going to college, and the outcome is earnings, we can always think of the two potential outcomes: y_{i1} are earnings of individual i if he goes to college, and y_{i0} are the earnings of individual i if he does not go to college. These potential outcomes are well-defined even before the actual treatment is received.

This framework allow us to define the causal effect of the treatment for individual i as

$$y_{i1} - y_{i0} \tag{1}$$

and the **average treatment effect** (ATE) as

$$ATE = \mathbb{E}(y_{i1} - y_{i0}) \tag{2}$$

Let d_i be the treatment indicator

$$d_i = \begin{cases} 1 & \text{if treated} \\ 0 & \text{if not treated} \end{cases}$$

An alternative causal effect of interest is the **average treatment effect** for the treated,

$$ATE_1 = \mathbb{E}(y_{i1} - y_{i0} | d_i = 1) \tag{3}$$

Some argue that ATE_1 is more relevant for policy purposes because, in general, we will not be interested in what the effect of treatment is for those in the population who will never receive it (we don't care about the causal effect of a training program on multimillionaires!) But notice that ATE is the treatment effect on the total population so that by carefully defining the population of interest ATE can be made as relevant as ATE_1.

4.2 Estimating the causal effect

The problem of causal inference is that we can never observe y_{i1} and y_{i0} together for the same individual. For an individual that receives the treatment the observed outcome is y_{i1} and the counterfactual outcome is y_{i0} and, conversely, for those not treated. We only observe one of the two outcomes for each individual, i.e., we never observe the counterfactual. Thus we will never be able to observe (1). We therefore resort to estimating the expected counterfactual, and this leads us to estimate an average treatment effect such as ATE and ATE_1.

The observed outcome for individual i, y_i, is

$$y_i = y_{i0} + (y_{i1} - y_{i0}) d_i$$

A natural (naive) estimator of the average treatment effect is the difference in means of the outcome between the treated and the untreated

$$ATE_0 = \mathbb{E}(y_i | d_i = 1) - \mathbb{E}(y_i | d_i = 0)$$

which can be estimated very easily by $\bar{y}_1 - \bar{y}_0$, where $\bar{y}_d$ is the mean outcome mean for treated ($d = 1$) and non treated ($d = 0$) individuals.

This estimator will estimate an average causal effect if the treatment is randomly assigned, as this implies that d_i is independent of potential outcomes y_{i1} and y_{i0}. In this case, $E(y_{i1}|d_i = 1)$ does not depend on d; i.e., $\mathbb{E}(y_{i1}|d_i = 1) = \mathbb{E}(y_{i1}|d_i = 0) = E(y_{i1})$, and similarly for y_{i0}. Then

$$\begin{aligned} &\mathbb{E}(y_i|d_i = 1) - \mathbb{E}(y_i|d_i = 0) \\ =&\mathbb{E}(y_{i1}|d_i = 1) - \mathbb{E}(y_{i0}|d_i = 0) \\ =&\mathbb{E}(y_{i1}) - \mathbb{E}(y_{i0}) = ATE \end{aligned}$$

and

$$\begin{aligned}&\mathbb{E}(y_i|d_i=1)-\mathbb{E}(y_i|d_i=0)\\=&\mathbb{E}(y_{i1}|d_i=1)-\mathbb{E}(y_{i0}|d_i=0)\\=&\mathbb{E}(y_{i1}|d_i=1)-\mathbb{E}(y_{i0}|d_i=1)=ATE_1\end{aligned}$$

In other words, with random assignment, the simple difference in means gives an unbiased estimate of both the average treatment effect and the average treatment effect on the treated. There is no difference between ATE and ATE_1 when treatment is randomly assigned.

If d is not randomly assigned this estimator will be biased. Suppose, for example, that more capable individuals are given a scholarship to attend college. The mean wages of those receiving the scholarship will be above the mean wage of those not receiving the scholarship but this difference does not represent (only) the causal effect of the scholarship. To see this, we write

$$\begin{aligned}&\mathbb{E}(y_i|d_i=1)-\mathbb{E}(y_i|d_i=0)\quad\text{observed difference}\\=&\mathbb{E}(y_{i1}|d_i=1)-\mathbb{E}(y_{i0}|d_i=0)\\=&\underbrace{\mathbb{E}(y_{i1}|d_i=1)-\mathbb{E}(y_{i0}|d_i=1)}_{ATE_1}+\underbrace{\mathbb{E}(y_{i0}|d_i=1)-\mathbb{E}(y_{i0}|d_i=0)}_{selection\ bias}\end{aligned}$$

$\mathbb{E}(y_{i1}|d_i=1)-\mathbb{E}(y_{i0}|d_i=1)$ is the causal effect of the scholarship on those who received the scholarship. It represents the mean difference in wages obtained with the scholarship and the wages that would have been obtained without the scholarship, for those individuals who received the scholarship. The selection term picks up the difference in wages before the scholarship is given between those that will receive it and those that will not receive it.

If, as argued, more capable individuals get the scholarship then the selection term is positive and a positive difference in observed mean outcomes does not necessarily imply a positive causal effect of the scholarship. Note that it would be enough to assume that d_i is independent of potential outcomes y_{0i} only to ensure that the simple difference in mean-outcome is an unbiased estimate of ATE_1:

4.3 Difference in differences

Panel data can be useful to infer causality when the treatment is not randomly assigned (the usual situation in data science). The basic assumption is an additive structure for the non-treatment potential outcome

$$\mathbb{E}(y_{it0}|c_i,\lambda_t)=c_i+\lambda_t \tag{4}$$

where t index time.

The outcome of individual i when treatment is absent equals the sum of a time-invariant individual effect and a time-effect common across all i. For simplicity, we abstract from other covariates x_{it} affecting potential outcomes. We next assume

$$\mathbb{E}(y_{it1}|c_i,\lambda_t)=\mathbb{E}(y_{it0})+\delta \tag{5}$$

so that the average causal effect of the treatment is $\delta=\mathbb{E}(y_{it1}-y_{it0})=ATE$, which is constant over t and i. In terms of the observed outcome y_{it} we have

$$y_{it}=y_{it0}+(y_{it1}-y_{it0})d_{it}$$

and the assumptions imply

$$\begin{aligned}&y_{it0}=c_i+\lambda_t+u_{it}\quad\text{where }\mathbb{E}(u_{it}|c_i,\lambda_t)=0\\&y_{it1}=c_i+\lambda_t+\delta+u_{it}\end{aligned}$$

so that we can rewrite the observed outcome as

$$y_{it} = c_i + \lambda_t + \delta d_{it} + u_{it} \qquad (6)$$

We can estimate δ by comparing appropriate differences in observed outcomes between treated and non-treated individuals. Suppose $d_{it-1} = 0$ and $d_{it} = 1$ for an individual i. The mean change in y between $t-1$and t for this treated individual is

$$\begin{aligned}
&\mathbb{E}(y_{it}|c_i, \lambda_t, d_{it} = 1) - \mathbb{E}(y_{it-1}|c_i, \lambda_{t-1}, d_{it-1} = 0) \\
&= c_i + \lambda_t + \delta + \mathbb{E}(u_{it}|c_i, \lambda_t, d_{it} = 1) - (c_i + \lambda_{t-1} + \mathbb{E}(u_{it-1}|c_i, \lambda_{t-1}, d_{it-1} = 0)) \\
&= \delta + \lambda_t - \lambda_{t-1} + \mathbb{E}(u_{it}|c_i, \lambda_{t-1}, d_{it-1} = 1) - \mathbb{E}(u_{it-1}|c_i, \lambda_{t-1}, d_{it-1} = 0)
\end{aligned}$$

where

$$\begin{aligned}
\mathbb{E}(y_{it1} - y_{it0}) &= \mathbb{E}[\mathbb{E}(y_{it1}|c_i, \lambda_t) - \mathbb{E}(y_{it0}|c_i, \lambda_t)] \\
&= \mathbb{E}[c_i + \lambda_t + \delta - c_i - \lambda_t] = \delta
\end{aligned}$$

Suppose now that another individual h was not treated at $t-1$ nor at t. For this control individual, the mean change in outcome between $t-1$ and t is

$$\begin{aligned}
&\mathbb{E}(y_{ht}|c_h, \lambda_t, d_{ht} = 0) - \mathbb{E}(y_{ht-1}|c_h, \lambda_{t-1}, d_{ht-1} = 0) \\
&= \lambda_t - \lambda_{t-1} + \mathbb{E}(u_{ht}|c_h, \lambda_t, d_{ht} = 0) - \mathbb{E}(u_{ht-1}|c_h, \lambda_{t-1}, d_{ht-1} = 0)
\end{aligned}$$

The difference in these differences is therefore (omitting the conditioning on c, λ)

$$\begin{aligned}
&[\mathbb{E}(y_{it}|d_{ht} = 1) - \mathbb{E}(y_{it-1}|d_{it-1} = 0)] - [\mathbb{E}(y_{ht}|d_{ht} = 0) - \mathbb{E}(y_{ht-1}|d_{ht-1} = 0)] \\
&= \delta + \mathbb{E}(u_{it}|d_{it} = 1) - \mathbb{E}(u_{ht}|d_{ht} = 0) + \mathbb{E}(u_{it-1}|d_{ht-1} = 0) - \mathbb{E}(u_{it-1}|d_{it-1} = 0)
\end{aligned}$$

and provided u is mean-independent of d, given $(c, \lambda) - \mathbb{E}(u_{it}|c_i, \lambda_t, d_{it}) = 0$ for any (i, t), the analogous sample means over treated and non-treated individuals will estimate δ, the ATE, consistently.

Note that assuming u_{it} mean-independent of d_{it}, given (c_i, λ_t), is a weaker assumption than assuming y_{i0} is mean-independent of d_{it} which is what characterizes random assignment of the treatment. The assumption amounts to assume that y_{i0} is mean-independent of d_{it} conditional on (c_i, λ_t).

Differences in differences (DD or diffs-in-diffs) works because the within-individual difference eliminates the individual effect while the difference in the within-differences eliminates the time effects between the pre- and post- treatment periods. In Figure 1, we show how the estimator is computed using two individuals (T and C) and two time periods (before and after treatment). Note that both individuals have the same time trend but different level effects. The treatment generates a deviation from trend in the treated individual.

Equation (6) resembles a panel data regression where y is regressed on time dummies and on the dummy variable d; while accounting for the individual effect c_i. We can then estimate δ by panel data estimators like FD (First-Difference). This is actually convenient since these estimators will also compute standard errors of the estimated δ.

An important point to note is that (5) assumes that the time effect λ_t is the same for the treated and non-treated individuals. If this is not the case, the diffs-in-diffs approach breaks down. This is commonly referred as the common trends assumption. In Figure 2 we observe that when the treated individual has no time trend, the DD estimator will be biased. In the case depicted in Figure 2, DD under-estimates the treatment effect. Thus, it is important to check how reasonable this assumption is by checking the trend of y in the treated and control group for periods before the treatment is applied.

This approach can be extended by allowing the non-treatment potential outcome to depend on a vector of covariates x_{it}. These covariates may also control for differences in trend between treated and non-treated individuals. Model (4) is then

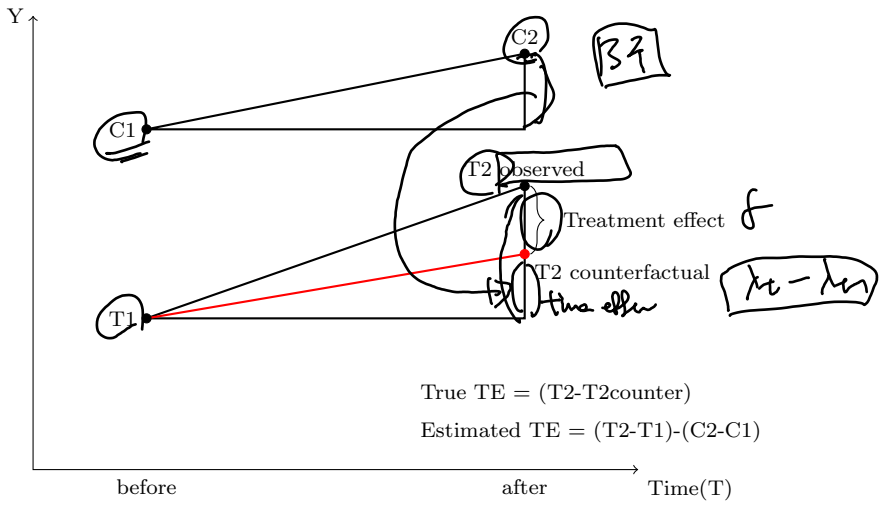


Figure 1: Diffs in diffs with common trends

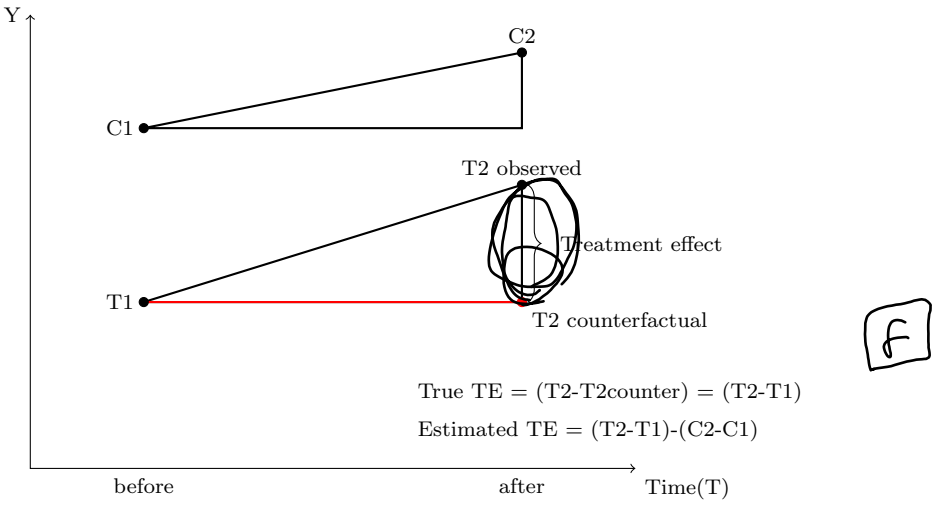


Figure 2: Diffs in diffs with different trends

$$\mathbb{E}(y_{it0}|c_i, \lambda_t) = x_{it}\beta + c_i + \lambda_t$$

leading to

$$y_{it} = x_{it}\beta + \delta d_{it} + c_i + \lambda_t + u_{it}$$

which is the standard panel regression model.

4.4 Policy changes

In many applications the treatment is a policy change which is applied to all individuals in a given group and we use group-level data to estimate the casual effect of the treatment. For example, a tax change in state (group) s is applied to all individuals in state s. In this case, DD is fixed effect estimation applied to aggregate (state level) data. Potential outcomes Equations (4) and (5) are then

$$\mathbb{E}(y_{st0}|c_s, \lambda_t) = c_s + \lambda_t$$
$$\mathbb{E}(y_{st1}|c_s, \lambda_t) = \mathbb{E}(y_{st0}|c_s, \lambda_t) + \delta$$

where c_s is an individual effect for state s which can be thought of as the mean of c_i for individuals in state s, $\mathbb{E}(c_i|i \in s) = c_s$. We then have

$$y_{st0} = c_s + \lambda_t + u_{st} \qquad \mathbb{E}(u_{st}|c_s, \lambda_t) = 0$$
$$y_{st1} = c_s + \lambda_t + \delta + u_{st}$$

leading to the observed outcome

$$y_{st} = \delta d_{st} + \lambda_t + c_s + u_{st} \tag{7}$$

In this type of aggregate analysis we have 2 periods: $t = 1$ is before the policy change and $t = 2$ is after the policy change (we could easily incorporate more periods to the analysis). It is important to have a group that is treated at $t = 2$ and another group that is not treated at $t = 2$. No group is treated at $t = 1$: Thus, following the logic of DD we get

$$[\mathbb{E}(y_{s2}|d_{s2} = 1) - \mathbb{E}(y_{s1}|d_{s2} = 0)] - [\mathbb{E}(y_{s2}|d_{st} = 0) - \mathbb{E}(y_{st-1}|d_{st-1} = 0)] = \delta$$

provided u_{st} is mean-independent of d_{st} for all (s, t). We can easily estimate these conditional means from sample data to obtain an estimate of δ.

Card and Krueger (1994) examine the effect of a raise in the minimum wage on employment. New Jersey raised its minimum wage on April 1, 1992 from 4.25 to 5.05 dollars per hour. They collected data on wages in fast-food restaurants for February and for November 1992 for New Jersey and Pennsylvania. Pennsylvania kept the minimum age at 4.25. Their data are presented in the following table (from Angrist and Pischke,2009)

Variable	PA (i)	NJ (ii)	Difference, NJ-PA (iii)
1. FTE employment before, all available observations	23.33 (1.35)	20.44 (0.51)	−2.89 (1.44)
2. FTE employment after, all available observations	21.17 (0.94)	21.03 (0.52)	−0.14 (1.07)
3. Change in mean FTE employment	−2.16 (1.25)	0.59 (0.54)	2.76 (1.36)

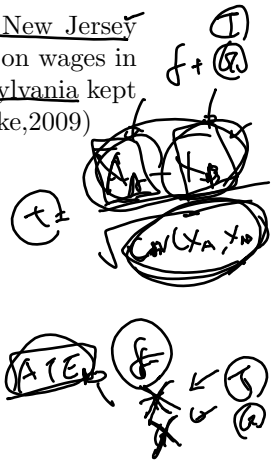


Figure 3: Average employment per store before and after the rise in New Jersey minimum wage
Note: Adopted from Card and Krueger (1994), Table3. Standard errors shown in paranthese. The sample consists of all stores with available data on employment. FTE (full-time-equvalent) employment counts each part-time worker as half a full-time worker. Employment at six closed stores is set to zero. Employment at four temporarily closed is treated as missing.

The surprising result is that the DD estimator indicated that employment per store increased by 2.76 FTE in New Jersey as a result of the raise in the minimum wage! This does not make much sense economically. The problem here might be in the common trends assumption. Maybe employment in Pennsylvania was trending downward whereas in New Jersey there was no trend?

If this was the case then there is no need for the second differencing and actually subtracting the employment growth in Pennsylvania from that in New Jersey will make New Jersey look as if its employment was decreasing by less than it should have been, and we would therefore infer that the minimum wage had a positive effect on wages. The actual data are represented in Figure 4.

Indeed, employment in NJ between February and April appears flat whereas that in Pennsylvania decreased a little bit. A better example is given in Angrist and Pischke. The graph below shows grade repetition rates in German states which experienced a short school year in 1967 and 1968 due to a policy change (24 weeks instead of 37 weeks), and in Bavaria which did not experience a change in the length of school year.

In this example, the common trend assumption prior to 1967 seems reasonable. We observe a deviation from trend in the treated states which rapidly disappears after the end of the intervention. Note that (7) resembles a panel regression model with state and time effects and a dummy variable regressor which varies at the state and time level. Define

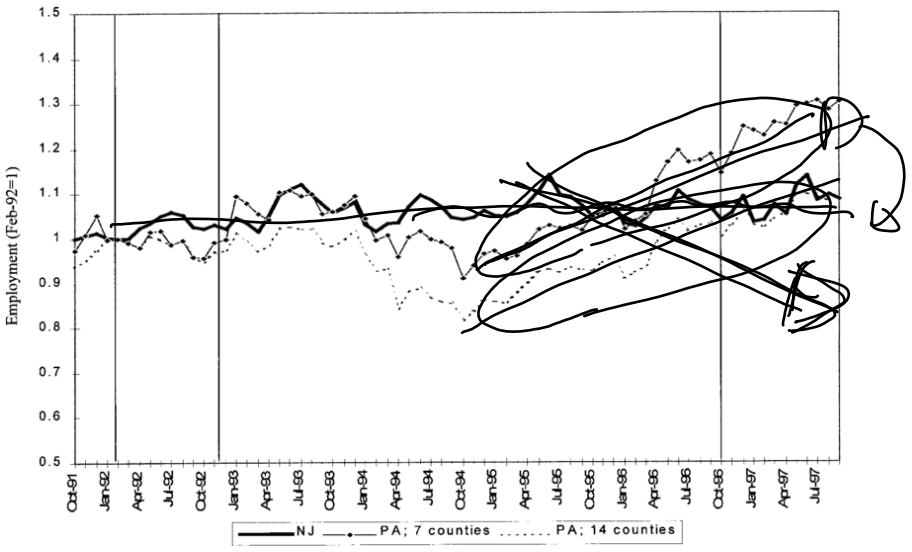


Figure 4: Employment rates from 1991 to 1997 in NJ and PA. PA 7 counties for urban and PA 14 counties for urban and rural area.

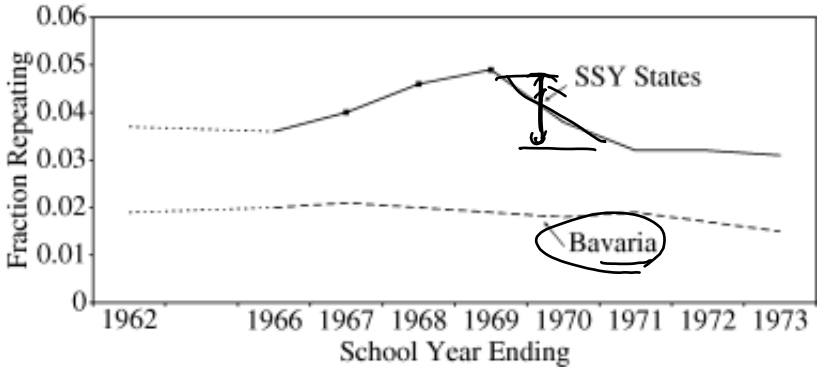


Figure 5: Average rates of grade repetition in second grade for treatment and control schools (from Pischke 2007).

$$T_s = \begin{cases} 1 & \text{if states } s \text{ is treated} \\ 0 & \text{if states } s \text{ is not treated} \end{cases}$$

$$D_t = \begin{cases} 1 & \text{for periods at or after the treatment} \\ 0 & \text{for periods before the treatment} \end{cases}$$

Then

$$d_{st} = T_s \times D_t$$

When there are two periods (7) reduces to

$$y_{st} = D_t + c_s + \delta(T_s \times D_t) + u_{st} \tag{8}$$

i.e., there is an interaction term picking up the post-policy change in outcome for the treated states (groups), which is assumed to be the same for all treated groups.
If there are only two groups, treated and control group, then the regression takes the more familiar from (adding covariates x_{st})

$$y_{st} = x_{st}\beta + D_t + T_s + \delta(T_s \times D_t) + u_{st}$$

This is the most common way in which the causal effect of a policy change in estimated: use group-level data to run an OLS regression with fixed group and time effects as well as an interaction term. The coefficient on the interaction term is the ATE.

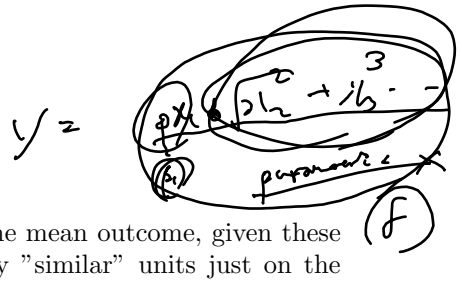
5 Propensity Score Matching

Define probability of treatment given the covariates. This is called the propensity score:

$$p(x) = Prob(d = 1|x)$$

Assume

$$(a)\ \mathbb{E}(y_0|x,d) = \mathbb{E}(y_0|x)$$
$$(b)\ \mathbb{E}(y_1|x,d) = \mathbb{E}(y_1|x)$$
$$(c)\ \ 0 < p(x) < 1$$



Assumptions (a) and (b) say that we can identify enough covariates such that the mean outcome, given these covariates, does not depend on whether treatment occurs. i.e., we can identify "similar" units just on the boundary of treatment. Assumption (c) ensures we do not have units that are certain of treatment. These assumptions are labelled the strong ignorability of treatment (conditional on x).

Then one can prove

$$ATE = \frac{\mathbb{E}(D - p(x))y}{p(x)(1 - p(x))}$$
$$ATE_1 = \frac{\mathbb{E}(D - p(x))y}{Prob(D = 1)(1 - p(x))}$$

ATE and ATE_1 are non-parametrically identified once we know the propensity score function. We need to estimate this function to make this approach operational. Parametric approaches are to use Probit or Logit analysis.

Let $\hat{p}(x) = F(x; \hat{\gamma})$ be the predicted propensity score. Then consistent estimators for the treatment effects are

$$ATE = N^{-1} \sum_{i=1}^{N} \frac{(D_i - \hat{p}(x_i))y_i}{\hat{p}(x_i)(1 - \hat{p}(x_i))}$$
$$ATE_1 = (N^{-1} \sum_{i=1}^{N} D_i) N^{-1} \sum_{i=1}^{N} \frac{(D_i - \hat{p}(x_i))y_i}{(1 - \hat{p}(x_i))}$$

where we note that $(N^{-1} \sum_{i=1}^{N} D_i)$ is a consistent estimator of $Prob(D = 1)$. A popular approach is to identify the ATE from simple linear regression:

$$OLS : y_i \ on \ 1, D_i, \hat{p}(x)$$

The coefficient on D is a consistent estimate of the ATE.

Interpretation: $\hat{p}(s)$ functions as a control variable that contains all the relevant information in the covariates that is relevant to whether treatment occurs. Controlling for this information, treatment dummy does not suffer from endogeneity bias so we can estimate its effect consistently.

Using weaker assumptions that $\mathbb{E}(y_0|p(x))$ and $\mathbb{E}(y_1|p(x))$ are linear in $p(x)$ rather than not being functions of x at all, as before, we can estimate ATE from

$$OLS : y_i \ on \ 1, D_i, \hat{p}(x_i), D_i(\hat{p}(x_i) - \hat{\mu}_p)$$

where $\hat{\mu}$ is the sample average of $\hat{p}(x) : N^{-1} \sum_{i=1}^{N} \hat{p}(x_i)$. Again, the coefficient on D is a consistent estimate of the ATE.