Math Camp 2023 – Statistical Methods[*]

Written by Seonmin Will Heo

Department of Economics, UC Santa Barbara

September 8, 2023

Overview

This lecture note is an informal guide to various statistical tools used in applied microeconomics. The main goal of this lecture note is to provide students a brief summary of microeconometric methods that are discussed frequently in various reading groups. The current first-year economics sequence at UCSB provides a good foundation on microeconomic theory, macroeconomic theory, and regression analysis. However, students are not exposed to econometric estimation methods until their second year, and they would not have learned regression discontinuity until the end of their second year if the course is taught in the spring quarter.

Some would argue that the inner workings of these methods are difficult to understand, and they would not be truly comprehensible unless you finish the first-year PhD sequence. However, I beg to differ. While the mathematical derivations of these methods may be difficult to understand, the reasons why we use them are not so difficult to comprehend. And I believe that studying these methods as early as one can comprehend them would allow students to start thinking about research questions even in their first and second years.

Given this lecture note is designed to be covered in two 90-minute sessions, it will lack in substance and detail. I hope that the main takeaway is contextualization. I am sure students will learn most of the methods discussed in this lecture note again in detail, and I hope that the narrative provided in this lecture note will help them understand the methods much faster.

The narrative that worked best for me in understanding a series of statistical methods is the story of "counterfactuals." We start with the best-case scenario where we use the control group as the proxy for the counterfactual of the treated group; that is, in a randomized

[*] The material posted on this note is for personal use only and is not intended for reproduction, distribution, or citation.

experiment, the treated group would have behaved like the untreated group in a counterfactual world where the treated group did not receive treatment. However, in cases where we cannot randomly assign treatment to units, this assumption may not hold. We have to start making different assumptions about the counterfactual of the treated group. In the absence of treatment, would the treated group have behaved like the control group, if their other observed characteristics are similar? If the control group looks very different from the treated group, can we perhaps construct a "synthetic control" that can serve as a counterfactual for the treated group?

The statistical methods discussed in this lecture note are designed by scholars who pursued ways to estimate a causal treatment effect when we could make strong assumptions about the counterfactual outcomes. We will start with the randomized controlled trials and discuss other models such as the instrumental variables estimation, the difference-in-differences estimation, all the way to the synthetic control method.

1 Randomized Controlled Trials

Suppose I want to measure the effect of drinking coffee on my blood pressure. In an ideal world, I could have myself in a cloned universe drink coffee, measure the increase in blood pressure, and compare that with myself who did not drink coffee. But unfortunately, in reality, if I drink coffee, I can only observe myself un-caffeinated before drinking coffee and caffeinated after drinking coffee – I cannot observe myself how I would have been had I not drank coffee. We can never observe the counterfactual outcome, and we have to make certain assumptions about the counterfactual outcome if we want to estimate a causal effect of some treatment. Comparing the values before and after does not identify the treatment effect for obvious reasons, although people make this mistake everyday.

Then what can we do to estimate a causal effect of some treatment? Ideally, we want to estimate the causal treatment effect for every individual, but this seems impossible. Instead, we can try to estimate the treatment effect *on average* by randomly assigning units into the treated group and the control group, like lab rats. Because the act of randomization can eliminate self-selection and therefore the selection bias that could have resulted from it, we can assume that the counterfactual outcome of the treated group would be the outcome of the control group *in expectation.* This intuitive way of estimating a causal effect is called the randomized controlled trials (RCT).[a]

[a] This experimental design is analogous to "A/B testing" or a "lab experiment" where the scientist divides the sample into two groups *randomly*, provides treatment to only one group, and compare the outcomes from the two groups.

1.1 Assumptions

There are strong assumptions to be made.

1. No selection bias
2. SUTVA

Selection bias is the bias introduced when individuals self-select into one of the groups, rendering the groups to be not properly randomized. We'll discuss it again later.

SUTVA stands for Stable Unit Treatment Value Assumption. This assumption implies no interference, no spillover effects, and every unit getting the same dose of treatment.

1.2 The Rubin causal model

In this section, I will use the notations introduced by Rubin (1974), often referred to as the potential outcomes framework. Another commonly used notation (which we won't use here) is from the latent variable models, which are often addressed in classic econometrics textbooks.

Imagine a setting with N individuals $i = 1, \ldots, N$, drawn randomly from a population. We want to estimate the effect of a binary treatment variable D_i on Y_i. D_i takes the value of 1 if the individual i receives treatment, and 0 otherwise.

For each individual, there are two potential outcomes, which are the following:

- $Y_i(0)$: potential outcome for i if i is not treated

- $Y_i(1)$: potential outcome for i if i is treated

It is important to note that for any individual in the sample, only one outcome is observed. For the treated individual i, we can only observe $Y_i(1)$. For the untreated individual i, we can only observe $Y_i(0)$. The observed data for each individual, Y_i, would then be the following:

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

Let's start from the ideal parameter of interest. The causal effect of treatment D on outcome Y for individual i we'd ideally want to estimate is

$$\tau_i := Y_i(1) - Y_i(0).$$

Note that this is for each individual i. However, because we can never observe both $Y_i(0)$ and $Y_i(1)$ for any individual i, we would never be able to estimate this treatment effect of interest. Does this mean we can't know anything about the treatment effect?

Let's not give up our hope yet. This was supposed to be a challenging parameter to recover anyways. Let us think for a moment what this means in real life. Suppose you take a pill because you have a fever. Have you ever wondered whether it will actually reduce the fever? What if your fever was about to go away but you took the pill and you're giving all the credit to the pill? Suppose you took the same pill ten years ago when you were much younger and your fever went away. Should you increase your dose if you want the same effect now? What if you're on different medication now? What if you gained a lot more weight in the last ten years? What if you are on a different diet living in a different climate now?

You don't know what exactly is going to happen, but you take the pill anyways. Perhaps it's because we assume that the pill will have the effect of reducing fever *on average*. Sure, the treatment effect may vary slightly even on the same individual, but we're willing to take the pill if it's suppose to reduce fever on average.

That being said, we may be interested in estimating the average treatment effect in some situations. A policy may have heterogeneous effects on individuals, but we would pursue this policy if it has a positive treatment effect on average. Let us call this parameter an average treatment effect (ATE), defined as follows:

$$\tau^{ATE} := \mathbb{E}[Y_i(1) - Y_i(0)] = \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)].$$

Let us go back to our observed data. We observe Y_i. We observe $Y_i(1)$ for the treated and $Y_i(0)$ for the untreated, but we do not observe both for any individual in our data. Let us start from another treatment parameter and work our way up to the parameter of interest, τ^{ATE}. Let

$$\tau^{D} := \mathbb{E}[Y_i(1)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0],$$

which simply indicates a difference in group means. We can rewrite it as

$$\begin{aligned}
\tau^{D} &= \mathbb{E}[Y_i(1)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0] \\
&= \mathbb{E}[Y_i(1)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0] + (\mathbb{E}[Y_i(0)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 1]) \\
&= (\mathbb{E}[Y_i(1)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 1]) + \mathbb{E}[Y_i(0)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0] \\
&= \mathbb{E}[Y_i(1) - Y_i(0)|D_i = 1] + (\mathbb{E}[Y_i(0)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0]),
\end{aligned}$$

where the first term is the average treatment effect for the treated (ATT), and the second term is the selection bias (take some time to check each term).

We now have two questions to be answered:

1. $\mathbb{E}[Y_i(1) - Y_i(0)|D_i = 1]$: Can this term be equal to $\mathbb{E}[Y_i(1) - Y_i(0)]$?

2. $\mathbb{E}[Y_i(0)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0]$: Can this term be equal to zero?

This is where our "randomized" assignment comes into play. The assumption of random assignment can be written as the following:

$$(Y(0), Y(1)) \perp\!\!\!\perp D$$

which means that treatment assignment is independent of potential outcomes. Random assignment also implies that there is no self-selection, making the second term zero. Thus, we have the following:

$$
\begin{aligned}
\tau^D &= \mathbb{E}[Y_i(1)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0] \\
&= \mathbb{E}[Y_i(1) - Y_i(0)|D_i = 1] + (\mathbb{E}[Y_i(0)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0]) \\
&= \mathbb{E}[Y_i(1) - Y_i(0)] \\
&\equiv ATE.
\end{aligned}
$$

In other words, with the assumption of random assignment and SUTVA, the average treatment effect is identified, and we can estimate the ATE by calculating the difference in group means, τ^D.

1.3 How to design a randomized controlled experiment

One way to design an experiment such that the assumptions of random assignment and SUTVA are satisfied is to ensure that there is no spillover within the sample. If we want to estimate the effect of, let's say, cash transfer to individuals, then we have to think carefully about ways in which the treated individuals can influence the decisions of the untreated individuals.

Another way is to estimate this potential spillover effect. We can design the experiment such that the first group (the truly untreated) receives neither the treatment nor the spillovers, the second group does not receive treatment but receives spillovers, and the third group (the treated) receives treatment.

1.4 Things to consider next

In this abridged lecture note, I do not discuss other important issues in randomized controlled trials. Here are some of the important things to study if interested in learning more about RCT:

1. Selection on observables, stratification, balance tests
2. The case of incomplete compliance, the intention-to-treat (ITT) estimation, and LATE
3. Power calculation

2 Instrumental Variables (IV) Estimation

2.1 Motivation

As we have seen earlier, randomized controlled trials are the gold-standard for examining causal relationships between variables. However, running such experiments can be very expensive, and it can take a very long time to finish. In some cases, running randomized experiments may simply not be possible. Here is a common example. Suppose we want to estimate the effect of smoking on health. How do we design an experiment such that we only examine the causal effect of smoking on health, and not the other way around? The big assumption we make when we run a regression is that the independent variable is not correlated with the error term. If there is an omitted variable such as depression that can affect both smoking and health, then the OLS regression will not estimate the causal effect of smoking on health. What can we do to isolate the causal effect of smoking on health?

Let me describe what was said above using some math. In a simple linear regression model, we have

$$Y = D\beta + U, \tag{1}$$

where Y and U are n-by-1 vectors of an outcome variable and an error term, respectively, and D is a n-by-k matrix where every column is a treatment variable. β is a k-by-1 vector of coefficients. By construct, the model estimates the coefficient vector β that optimizes the following objective function:

$$\widehat{\beta} \ = \ \arg\max_{\beta} \quad (Y - D\beta)'(Y - D\beta),$$

whose first-order condition is $D'(Y - D\widehat{\beta}) = D'\widehat{U} = 0$. Again, this OLS estimation *by design* results in $\widehat{\beta}$ such that $\text{Cov}(X, \widehat{U}) = 0$. Thus, if there is reason to believe that $\text{Cov}(X, \widehat{U}) \neq 0$ (e.g. depression can affect both smoking and health), then the model would not estimate the causal effect of D on Y, but rather, it would be picking up values that would make the error term appear to be uncorrelated with the treatment variables.

2.2 Classical IV model

To overcome the issue of endogenous treatment variables, we use the method of instrumental variables (IV). The goal is to use an instrumental variable ("instrument") that is related to the treatment variable (smoking) but not related to the error term to estimate the causal effect of smoking on health. These two conditions are called the following:

1. *Relevance*: The instrument is related to the independent variable.

2. *Exclusion Restriction*: The instrument can affect the outcome variable only through the independent variable. In other words, the instrument cannot directly affect the outcome.

One good instrument in our example is the tax rate for tobacco products. The tobacco tax rate is related to smoking, and it is reasonable to assume that it is related to health only through its effect on smoking.

The classical IV model assumes no heterogeneity, which makes this model rather restrictive: it implies linearity and constant treatment effects. Yet it is important to study this classical model, because 1) the researchers still use the estimators from this model, and 2) it can be used to justify the needs for a model with heterogeneity and nonlinearity.

2.2.1 Exact Point Identification

Let us first look at the "just identified" case. Suppose we have exactly the same number of instruments Z as the number of treatment variables D. If we pre-multiply the simple linear regression (Equation 1) by Z, we get

$$Z'Y = Z'D\beta + Z'U,$$

where Z is an n-by-k matrix where each column is an instrument. Recall the two assumptions that instruments satisfy:

1. Relevance: $\mathbb{E}[Z'D]$
2. Exclusion restriction: $\mathbb{E}[ZU] = 0$

If we apply the OLS to the result, we obtain

$$\widehat{\beta}_{JIIV} = (Z'D)^{-1}Z'Y.$$

2.2.2 Overidentification

What if we have more instruments than treatment variables ($d_Z > d_D$)? In this fortunate case of having an overidentified model, however, we have to come up with a way to select only d_D out of the d_Z instruments. One simple way would be to just randomly select d_D instruments at the expense of loss in precision. A better way would be to use d_D linear combinations of the instruments. In other words, we could project the columns of D onto the column space of the instruments to minimize the loss, as we would be obtaining the linear combinations of the instruments that are closest to the treatment variables. In linear algebra, this would be pre-multiplying by $Z(Z'Z)^{-1}Z'$:

$$Z(Z'Z)^{-1}Z'Y = Z(Z'Z)^{-1}Z'D\beta + Z(Z'Z)^{-1}Z'U.$$

Then applying the OLS yields

$$\widehat{\beta} = (D'Z(Z'Z)^{-1}Z'D)^{-1}D'Z(Z'Z)^{-1}Z'Y. \tag{2}$$

2.2.3 Two-Stage Least Squares

You may then wonder why we often refer to this estimator as the two-stage least squares estimator. Let $\widehat{D} = Z(Z'Z)^{-1}Z'D$. Then we can rewrite the expression 2 as

$$\begin{aligned}\widehat{\beta} &= (D'Z(Z'Z)^{-1}Z'D)^{-1}D'Z(Z'Z)^{-1}Z'Y \\ &= (D'Z(Z'Z)^{-1}Z'Z(Z'Z)^{-1}Z'D)^{-1}D'Z(Z'Z)^{-1}Z'Y \\ &= \left(\widehat{D}'\widehat{D}\right)^{-1}\widehat{D}'Y.\end{aligned}$$

If we look closely, $\widehat{D}$ is an $n \times k$ matrix of predicted values from the regression of D on Z (first-stage). The estimator above can then be obtained by regressing Y on $\widehat{D}$ (second-stage). Thus, the estimator is often referred to as the two-stage least squares (TSLS or 2SLS) estimator.

2.3 Modern IV model

The modern view does not assume constant treatment effects for all individuals. It allows for heterogeneity in β. However, as it now matters to which individuals the instrument has an impact, we should be more careful about interpreting the results. More importantly,

Then if the instrument (Z) and the treatment (D) are both binary, then

$$\beta_{IV} = \frac{\text{Cov}(Y,Z)}{\text{Cov}(Y,Z)} = \frac{\mathbb{E}[Y|Z=1] - \mathbb{E}[Y|Z=0]}{\mathbb{E}[D|Z=1] - \mathbb{E}[D|Z=0]}.$$

Imbens and Angrist (1994) provide conditions that would allow for easier interpretation of β_{IV}. First, we need the following assumptions:

1. Exogeneity: $Z \perp\!\!\!\perp (D(0), D(1), Y(0), Y(1))$

2. Relevance: $\text{Cov}(D,Z) \neq 0$

3. Monotonicity: $D(1) \geq D(0)$ or $D(1) \leq D(0) \;\; \forall i$

Under these assumptions, β_{IV} would estimate the local average treatment effect (LATE):

$$\begin{aligned}\beta_{IV} = \frac{\text{Cov}(Y,Z)}{\text{Cov}(Y,Z)} &= \frac{\mathbb{E}[Y|Z=1] - \mathbb{E}[Y|Z=0]}{\mathbb{E}[D|Z=1] - \mathbb{E}[D|Z=0]} \\ &= \mathbb{E}[Y(1) - Y(0)|D(1)=1, D(0)=0] \\ &= LATE,\end{aligned}$$

where LATE is an average treatment effect for the compliers: those who would receive treatment if and only if they are assigned to the treated group.

3 Difference-in-Differences Estimation

3.1 Motivation

Many interesting research questions often require some sort of panel data to be answered properly. How does financial assistance affect college enrollment? Does a policy increase an outcome variable of interest? To address these questions, we would need to track how individuals or firms make their decisions *over time*. Suppose we obtain a panel data; that is, a dataset that includes observations of the same units over multiple time periods. What estimator would you now use to estimate the effect?

A naive approach that resembles the RCT design is to use the before-and-after difference estimator. We simply take the difference between the average outcome of the treated group before and after the treatment:

$$\tau^d = \mathbb{E}[Y_{t_1}(1)–Y_{t_0}(0)|D = 1],$$

where $D = 1$ is the treated group, t_1 is the time period after treatment, and t_0 is the time period before treatment. $Y_{t_1}(1)$ is the outcome of the treated after treatment and $Y_{t_0}(0)$ is the outcome of the treated before treatment.

The identifying assumption here is that the outcome would remain unchanged in the absence of treatment; that is, the outcome would not have changed at all (on average) in the counterfactual scenario where the treated group did not receive treatment. With this assumption, the estimator would yield an average treatment effect for the treated (ATET):

$$\begin{aligned}\tau^d &= \mathbb{E}[Y_{t_1}(1) - Y_{t_0}(0)|D = 1] \\ &= \mathbb{E}[Y_{t_1}(1) - Y_{t_1}(0)|D = 1] + \mathbb{E}[Y_{t_1}(0) - Y_{t_0}(0)|D = 1] \\ &= ATET \quad + \text{counterfactual change of the treated}\end{aligned}$$

The identifying assumption implies that the second term is zero. Note that $Y_{t_1}(0)|D = 1$ that was subtracted from and added to the first term and the second term, respectively, is not observed: it is the counterfactual outcome of the treated after treatment if they were not treated. Thus, this assumption cannot be tested. But do you buy this assumption? Suppose colleges in some states started a financial assistance program. College enrollment in these states could have changed without this program, and we cannot assume that any change in enrollment before and after the program is solely due to the financial assistance program. Because this difference estimator captures both the treatment effect *and* the counterfactual trend, we cannot use this estimator unless we can somehow persuade the audience that the counterfactual trend is very likely to be zero.

3.2 Difference-in-Differences Estimator

This is where the difference-in-differences estimator comes in. This estimator is an attempt to estimate the counterfactual trend for the treated using the trend for the control group. The difference-in-differences estimator can be written as

$$\tau^{DID} = \mathbb{E}[Y_{t_1}(1) - Y_{t_0}(0)|D = 1] - \mathbb{E}[Y_{t_1}(0) - Y_{t_0}(0)|D = 0],$$

where $D = 1$ is the treated group, t_1 is the time period after treatment, and t_0 is the time period before treatment. The critical identifying assumption here is that the observed trend in the outcome of the control group is identical to the counterfactual trend of the treated:

$$\begin{aligned}
\tau^{DID} &= \mathbb{E}[Y_{t_1}(1) - Y_{t_0}(0)|D = 1] - \mathbb{E}[Y_{t_1}(0) - Y_{t_0}(0)|D = 0] \\
&= \mathbb{E}[Y_{t_1}(1) - Y_{t_1}(0)|D = 1] + \mathbb{E}[Y_{t_1}(0) - Y_{t_0}(0)|D = 1] - \mathbb{E}[Y_{t_1}(0) - Y_{t_0}(0)|D = 0] \\
&= ATET + \text{counterfactual trend of the treated} - \text{observed trend of the control group}
\end{aligned}$$

With the identifying assumption, the second term and the third term cancel out. This assumption is referred to as the parallel trends assumption (PTA). The DID estimation with this assumption removes the parallel time trends, capturing only the treatment effect.

3.3 Assumptions

Two main assumptions of the DID estimator are:

1. Parallel Trends Assumption: The treated group and the control group have parallel trends before treatment.

2. No Concurrent Changes: There are no other changes to the treated group at the time of treatment.

To check whether the first assumption (PTA) can be valid, we can check trends of the treated and the control before treatment. If the pre-treatment trends are completely off between the treated and the control, then it becomes difficult to argue that the PTA is valid. If the trends are very similar, then it is likely that the PTA is valid (read the “Notes” section). However, remember that these identifying assumptions are assumptions: we cannot prove whether they hold.

3.4 Difference-in-Differences as a Regression

Let I_{t_1} be an indicator variable that takes 1 if $t = t_1$ and 0 if $t = t_0$. Let D_i be an indicator variable that takes 1 if the individual i is treated and 0 if the individual is untreated. Then the difference-in-differences regression model is

$$Y_{it} = \beta_0 + \beta_1 I_{t_1} + \beta_2 D_i + \delta D_i \cdot I_{t_1} + \varepsilon_t.$$

Let us see how each potential outcome is estimated in this framework and how ATT is estimated. The expectation of each potential outcome is

$$\begin{aligned}
\mathbb{E}[Y_{t_1}(1)|D = 1] &= \beta_0 + \beta_1 + \beta_2 + \delta \\
\mathbb{E}[Y_{t_0}(0)|D = 1] &= \beta_0 + \beta_2 \\
\mathbb{E}[Y_{t_1}(0)|D = 0] &= \beta_0 + \beta_1 \\
\mathbb{E}[Y_{t_0}(0)|D = 0] &= \beta_0.
\end{aligned}$$

Then $\tau^{DID} \equiv \mathbb{E}[Y_{t_1}(1) - Y_{t_0}(0)|D = 1] - \mathbb{E}[Y_{t_1}(0) - Y_{t_0}(0)|D = 0] = \delta$.

3.5 Notes

1. The conventional way of checking parallel trends is doing an "eyeball" test. But beware that an eyeball test is not the most reliable way, as the trends may look similar or different depending on how much the figure is zoomed in.

2. The difference-in-differences estimation is quite straightforward, and the textbook 2×2 DID case makes a lot of sense. However, DID can get complicated very quickly with multiple treatment groups and multiple treatment periods/stages. A general term for a design that handles multiple treatments is called "event studies." I decided not to cover the general design here due to the interest of time, but I will try to include it in the next version if there is a high demand for it.

4 Matching Methods

4.1 Motivation

Recall that the RCT design makes a strong assumption that the outcome variable is independent of the treatment variable: $(Y_1, Y_0) \perp\!\!\!\perp D$. However, if we do not have the luxury of running a field experiment that could effectively divide the sample randomly into the treatment group and the control group, we cannot rely on this assumption. In fact, this assumption is rarely satisfied in the observed data. Suppose we want to estimate the effect of a federally funded job training program on earnings. It is difficult to argue that treatment assignment is random (what are some reasons?).

But what if treatment assignment is random conditional on observable characteristics X_i? Continuing with the example above, we know that people self-select into a job training program, and the likelihood of completing the training program and its effect may depend on individuals' education levels and other characteristics. If we can observe these characteristics and believe that treatment assignment is random across trained workers and untrained workers with the same observed characteristics, then we can compare the same "kind" of people with or without job training.

4.2 Assumption

If we believe that treatment assignment is random conditional on observables, i.e.

$$(Y_1, Y_0) \perp\!\!\!\perp D|X,$$

then we can estimate the following conditional treatment effect:

$$\tau(x) = \mathbb{E}[Y_{1i} - Y_{0i}|X = x].$$

In other words, we are matching individuals in the treated group with individuals in the control group based on their characteristics. Formally, this matching estimator relies on two strong assumptions:

1. Conditional Independence Assumption (CIA): $(Y_1, Y_0) \perp\!\!\!\perp D|X$

2. Common Support Assumption (CSA): $0 < Pr(D = 1|X = x) \equiv p(x) < 1 \quad \forall x$

CIA requires that treatment assignment is random once we have conditioned on observed characteristics. This means that there is no selection on unobservables. However, because we cannot observe unobservables, this assumption is not testable.

CSA states that every point $x \in X$ must have both treated and untreated observations. This assumption is straightforward – if we want to match a treated individual with an untreated individual with the same characteristics, there should be an untreated counterpart. This assumption is testable.

4.3 Matching Methods

4.3.1 The Cell Estimator

Let us start with a simple method. We divide the data into multiple cells where each cell would take one value from each observable. For example, suppose we have one observed characteristic, which is the level of education. Let $X = 1$ if the person does not have a high school degree, $X = 2$ if the person has a high school degree, $X = 3$ if the person has a college degree, and $X = 4$ if the person has a higher degree than college. We put the data into each cell of X. Then for every value of X, we calculate the difference in means. We compute a weighted average of these differences. Formally, it would be

$$\widehat{\delta} = \sum_{x \in X} \omega_x \widehat{\delta}_x,$$

where ω_x is the proportion of the sample that takes the value of x. Note that we did not make any assumption about a parametric functional form of the observed characteristics, making this a non-parametric estimator. In other words, we are simply calculating the weighted average of these differences based on their proportions without making an assumption about, say, a linear effect of education on earnings.

4.3.2 Other Exact Matching Methods

You may then wonder, what if we have more observed characteristics to worry about ($\{X_j\}_{j=1}^{J}$ where J is large)? What if X is continuous? Then we may end up with some cells that do not have any untreated worker.

The following methods can be used in such situations:

1. Bandwidth matching
2. Nearest k-neighbor matching
3. Stratified matching

These methods are rather straightforward. First, bandwidth matching is where a treated worker with $X = x$ is matched with untreated workers within bandwidth h, which have $X \in [x - h, x + h]$. Second, nearest k-neighbor matching would involve matching a treated worker with k untreated workers that have the closest values from $X = x$. Third, stratified matching is dividing the entire sample into n mutually exclusive groups.

These methods provide clear instructions on how treated individuals are matched with untreated counterparts. However, they suffer from the dimensionality problem. If we want to condition on several observables, and if they take several values, will the sample be big enough to cover all the combinations of these values? If we want to reduce the number of values by grouping them (categorizing education level into degree rather than number of years), how many groups should I make, and are these groupings justifiable?

Because of this issue, researchers have recently used matching based on propensity scores.

4.4 Propensity Score Matching (PSM)

The CIA states that $(Y_1, Y_0) \perp\!\!\!\perp D|X$. How would CIA look for matching based on propensity scores? Rosenbaum and Rubin (1983) show that

$$(Y_1, Y_0) \perp\!\!\!\perp D|X \quad \longrightarrow (Y_1, Y_0) \perp\!\!\!\perp D|p(x),$$

where $p(x) = Pr(D = 1|X = x)$ is the propensity score. Rosenbaum and Rubin use the law of iterated expectations and the CIA to prove the theorem. Using this theorem, we can match on a single variable instead of conditioning on all the observed characteristics X. This feature removes the dimensionality problem raised earlier. The trick here is to estimate the propensity score $p(x)$ correctly.

There are several ways to estimating the propensity score. We can use nonparametric methods such as cell estimators, or use parametric methods such as a linear probability model, logit, and probit.

4.4.1 Estimating the propensity score $p(x)$

1. Nonparametric Cell Estimators

   This estimation method does not rely on a parametric assumption about the relationship between the propensity score and the observed characteristics. For each value $x \in X$ of the observed characteristics, we estimate the propensity score by calculating the proportion of observations that are treated:

   $$p(x) = \frac{n_{x,D=1}}{n_x}$$

2. Linear Probability Model

   This parametric estimation assumes linear probability in treatment as a function of the observed characteristics X:

   $$Pr(D = 1|X) = X\beta.$$

   This can be estimated via OLS. The coefficients represent the marginal effect of X on the probability of treatment, which makes the model easy to interpret. However, this estimation method suffers from one critical drawback, which is that the predicted probabilities are not bound between $[0, 1]$.

3. Logit and Probit

   To overcome the issue of out-of-bound predicted probabilities, we can model the conditional probability such that the predicted values are bound between 0 and 1.

   - Logit: $Pr(D = 1|X) = \dfrac{e^{X\beta}}{1 + e^{X\beta}}$

- Probit: $Pr(D=1|X) = \Phi(X\beta)$

These functions on the right-hand side are CDFs, which are bound between 0 and 1 for the entire domain. These models assume that the likelihood increases with the higher value of X.

These models differ mainly in these "link functions" – the functions that link between the linear predictor and the mean of the distribution function. Logit models assume that the error term follows a logistic distribution, while probit models assume a normal distribution for the error term. Logit models are often used for their easier interpretation of the coefficients (cf. the odds ratio).

4.4.2 Estimating the treatment effect

Now that we estimated the propensity score, we move on to estimating the treatment effect. There are again several ways of estimating the treatment effect.

1. Including $p(x)$ as a covariate

   A rather straightforward way of estimating the treatment effect is to include the propensity score as a covariate:

   $$Y_i = \alpha + \tau D_i + \beta p(X_i) + u_i,$$

   where D_i is the treatment status for unit i and X_i is the vector of observed characteristics for unit i.

2. Blocking on $p(x)$

   We divide the propensity score $p(x)$ to make K number of blocks. Suppose we make 5 blocks. Then each block k will have a width of $1/5 = 0.20$ based on $p(x)$. We then compute the treatment effect $\widehat{\tau}_k$ within each block. The final matching estimator is then

   $$\widehat{\tau}_b = \sum_{k=1}^{K} \widehat{\tau}_k \cdot \frac{N_k}{N},$$

   where N_k is the number of treated and control observations whose $p(x)$ fall within the block k.

   We can obtain the treatment effects $\widehat{\tau}_k$ by running the following regression:

   $$Y_{ik} = \alpha + \tau_k D_{ik} + X_i\beta + u_{ik}.$$

   We can then calculate $\widehat{\tau}_b$ using the formula above.

3. Weighting by $p(x)$

This method involves weighting units to obtain the ATE in a rather simple way. Consider giving the weight $1/p(x)$ for the treated group and the weight $1/(1-p(x))$ for the control group. Then we can show that

$$\begin{aligned}\mathbb{E}\left[\frac{DY}{p(X)}\right] &= \mathbb{E}[Y(1)] \qquad \text{and} \\ \mathbb{E}\left[\frac{(1-D)Y}{1-p(X)}\right] &= \mathbb{E}[Y(0)].\end{aligned}$$

Then we can combine them to obtain the ATE:

$$\begin{aligned}ATE &\equiv \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] \\ &= \mathbb{E}\left[\frac{DY}{p(X)}\right] - \mathbb{E}\left[\frac{(1-D)Y}{1-p(X)}\right]\end{aligned}$$

We can estimate ATE by its sample analog with $\widehat{p}(X)$:

$$\widehat{\tau}_M^{PS} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{D_iY_i}{\widehat{p}(X_i)} - \frac{(1-D_i)Y_i}{1-\widehat{p}(X_i)}\right).$$

The intuition behind this estimator is that the estimated $\widehat{p}(X)$ indicates the proportion of units that are treated, and weighting each group by $1/p(x)$ and $1/(1-p(x))$ make all the units be represented equally across two groups.

This weighting can be applied by running the weighted OLS in the following regression

$$Y_i = \alpha + \tau D_i + X_i\beta + u_i$$

with the weight

$$\omega_i = \sqrt{\frac{DY}{p(X)} - \frac{(1-D)Y}{1-p(X)}}.$$

4.5 Summary

We just learned various matching methods in a SparkNotes fashion, starting with the key assumptions for matching (CIA, CSA), exact matching methods, all the way to propensity score matching. In many cases, we need to condition on many covariates to argue for conditional independence. But a large number of variables inevitably leads to the high dimensionality problem, where there are too many cells both the treated and control units need to fill in. When the exact matching methods suffer from this critical issue, propensity score matching offers a novel solution where we estimate the propensity score using all the observed characteristics and match on this one variable instead of having to match on all the covariates. The support condition is now more likely to be satisfied.

However, this creates another problem of estimating $p(x)$. If there are only a few covariates with discrete values, we can rely on nonparametric methods to estimate the propensity score (but then, the support condition is also likely to be met in exact matching methods...). But with some continuous variables, we may have to rely on some functional forms for estimation.

There is even a bigger concern. Economists care a lot of self-selection, and as long as we cannot properly control for treatment assignment (Z) and take-up (D), the audience will have a hard time believing the conditional independence assumption, a strong assumption that is often *not* satisfied in reality. Take college enrollment, for example. No matter how many characteristics we observe about students (e.g. SAT scores, parents' education level, parents' income, and neighborhood they live in), we may still think that there is something fundamentally different about students who are enrolled in college from those who are not. However, because these matching estimators require a strong assumption about the unobservables – choices given observables X are assumed to be random. Thus, these matching estimators that are based on "selection on observables" models are less widely used in economics.

5 Synthetic Control Method

5.1 Motivation

Let us revisit the DID estimation for a moment. The key identifying assumption we made was the parallel trends assumption, where we set the observed trend of the control group equal to the counterfactual trend of the treated. But there are cases where the argument/assumption that the control and the treated would have the parallel trends is not convincing. Making the matter worse, imagine we only have one treated unit where all the control units do not satisfy parallel trends. One such case is a policy implementation in one state in the United States. If there was a tax change in California, can we find another state whose observed trend is parallel to the counterfactual trend in California? Each state is quite unique in their own ways, and it is difficult to argue that a neighboring state can be used as the control for the DID estimation.

We can stop here and move onto another research idea, but some economists had a good idea: if there is no single candidate that can serve as a good control, then what if we artificially construct a weighted average of control units that "matches" parallel trends?

Abadie, Diamond, and Hainmueller (2010) examine the effect of California's anti-tobacco law (Proposition 99) on cigarette sales. There are no other states that have the same parallel trends as California before the law was passed, so the authors construct a synthetic California that is the weighted average of the control states.

5.2 Constructing a synthetic control

How do we find the weights for the control units to construct a synthetic control? I will first describe the procedure of constructing a synthetic control and then use an example to walk through the process again.

Let Z_1 be pre-treatment outcomes and covariates for the treated unit and Z_0 be pre-treatment outcomes and covariates for the control group. Suppose we have k variables and J control units, then Z_1 is a $k \times 1$ matrix and Z_0 is a $k \times J$ matrix. The weights are obtained by the following:

$$\omega_j^* \equiv \arg\min_{\omega \in \mathbb{R}^J} (Z_1 - Z_0 W)' V (Z_1 - Z_0 W) \quad s.t. \quad \omega_j \geq 0 \; \forall j \quad \text{and} \quad \sum_{j=1}^{J} \omega_j = 1,$$

where W is $J \times 1$ vector of weights for each control unit and V is a weighting matrix. Then the estimate for the treatment effect in the post-treatment period is $Y_1 - Y_0 W*$ where $W*$ is a $J \times 1$ vector of obtained weights ω_J^*.

Let us use the signature example for the synthetic control method to describe the process. The outcome of interest is per-capita cigarette sales, and the treatment (Proposition 99) occurs in 1988. Then to find the weights, we need to first decide what variables to use to construct a synthetic control. We can use lagged outcome variables (per-capita cigarette sales in 1975, 1980, and 1988). We can also include other characteristics for each state that would help us construct a plausible synthetic control, such as cigarette price, logged per-capita income, percentage of population aged 15-24, and per-capita beer consumption, all averaged over 1980-1988. Then Z_1 will have all these seven variables ($k = 7$). The authors also remove the states that passed a similar law and consider the remaining 38 states ($J = 38$).

5.3 Notes

1. The key assumption here is that the treatment and the synthetic control have the same parallel trends. Another assumption we implicitly make is that the counterfactual outcome of the treated unit can be constructed as a linear combination of the control units. Other implicit assumptions are similar to those in DID (no other concurrent changes to the treated unit, only the treated unit receives treatment).

2. The pre-treatment trend of the synthetic control is almost identical to the pre-treatment trend of the treated unit *by construct*, since the objective function includes lagged outcome variables. One way to check whether this assumption could be valid is to form a synthetic control using a subset of data that includes only half of the pre-treatment periods, use the obtained weights to construct a synthetic control, and then see whether the parallel trends hold between the synthetic control and the treated unit for the removed pre-treatment periods.

3. The synthetic control method is used when we have one treated unit (a "case study") and several control units from which we can form a synthetic control. What if we want to apply the same framework to a case with multiple treated units? In addition, can we try to construct a synthetic control that has the parallel trend and not necessarily the same level? Google "synthetic difference-in-differences method" for more information.