Causal Inference - Notes From Mastering Metrics

February 13, 2016

Preface

The purpose of this document is to take a **first pass** to increase my understanding and foundation of the topic of econometrics and causal inference. As of early February 2016, my understanding in this area is pretty limited. I have identified several references that I can use to to improve my understanding. This includes 'Mastering Metrics', 'Mostly Harmless Econometrics', and the MIT course by Joshua Angrist. This document is a companion sketch that I use during my time of reading Mastering Metrics. It's the very first step, and I'll probably share this with broader audience to get feedbacks.

The Basics

Key Idea: To understand the causal impact of X on Y, it is not enough to make comparisons between those who have X and not (i.e. observational data) due to selection bias. When we say correlation ! = causation, what we really mean is:

$$\underbrace{\text{Difference in Group Mean}}_{\text{correlation study}} = \underbrace{\text{Average Causal Effect}}_{\text{causal inference}} + \text{Selection Bias}$$

The term on the left hand side is what we usually do naively when performing analysis (sort people into DOs and DONTs, and just compare the mean), but we rarely try to do more. In fact, we see that causal effect is actually coupled with selection bias on the right hand side, and what we care about is really only the causal impact term. To formalize the RHS terms, let's use the following notation

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

An example of D could be whether a particular subject has a college degree or not. Note, D (the treatment assignment) can be engineered by an engineer/scientist/experiment, or that the environment simply cause people to "self select" into a treatment group. Let Y to the metric of interest, e.g. Average yearly income, then our question could be around whether those who went to college actually earns a higher wage.

The important thing to realize here is that for each person, there is only one realization – you either went to college or you do not, and only one path is realized, so only one Y is observed. However, in abstract terms, each person (at least in theory) can have two potential outcomes, e.g. which are the wage they would earn if they went to college $Y(D=1)$ v.s. had they not gone to college $Y(D=0)$. The unobserved, alternative outcome that an individual has not realized is called "**counterfactual**". If we knew what the counterfactuals are, causual inference would have been pretty easy :)

Despite the fact that we only get to observe one outcome, we might still want to think about how we would estimate the causal effect if counterfactual were to be available to us (and how it relates to the naive method). Let us define a few more notation:

$$Y(i|D=j) = Y_{ij}$$

which $D \in \{0,1\}$ represent whether a user was treated (1) or not (0) in reality, and $i \in \{0,1\}$ represents the scenarios, 1 had the subject been treated, and 0, had the subject not been treated. For someone who is treated, the counterfactual would be Y_{01}. For someone who is untreated, the counterfactual would be Y_{10}.

Given that we only have the observed outcomes, the naive thing to do is to directly compare them ($Y_{11} - Y_{00}$). With large sample size, we would naturally look at $Avg(Y_{11}) - Avg(Y_{00})$, but this can be problematic, because while we are going after the causal impact of the **treated**, we ended up getting:

$$\begin{aligned} \underbrace{E[Y_{11} - Y_{00}]}_{\text{what we have at hand}} &= E[Y_{11}(-Y_{01} + Y_{01}) - Y_{00}] \\ &= \underbrace{E[Y_{11} - Y_{01}|D=1]}_{\text{the causal impact of treated we are really after}} + E[Y_{01} - Y_{00}] \end{aligned}$$

Notice that, when we simply compare the realized outcome of the treated against the untreated, we see that there is an extra term $E[Y_{01} - Y_{00}]$ lurking, and this can be thought of as the **selection bias**. Why? in the absence of Y_{01}, we are trying use Y_{00} as a proxy for it. However, when this proxy is imperfect, and is often the case (e.g. the subject who have gone to college is biologically smart than those who didn't, so even if they had not gone to college, they might make more), the additional term will mess up our estimation.

We can reason exactly & symmetrically to estimate the causal impact for the untreated group:

$$\begin{aligned} \underbrace{E[Y_{11} - Y_{00}]}_{\text{what we have at hand}} &= E[Y_{11}(-Y_{10} + Y_{10}) - Y_{00}] \\ &= E[Y_{11} - Y_{10}] + \underbrace{E[Y_{10} - Y_{00}|D=0]}_{\text{the causal impact of untreated is what we are really after}} \end{aligned}$$

Here, we would like to use Y_{10} to estimate the causal impact, but all we have is Y_{11} as a proxy. And again, it's entirely reasonable that Y_{11} would be an overestimate of Y_{10} due to intelligence difference. Therefore, the theme of **average casual effect = our poor estimate - selection bias** is a consistent theme.

In Science, we often are interested not just the causal impact on the treated or untreated alone, we are interested in the average causal impact across all groups. Defining μ is the average causal impact, across all groups with a bit of Algebra (see the Northwestern paper), we can show that the average causal impact (combining the treated and the untreated) is:

$$\mu = E[Y_{11} - Y_{00}] - \left\{ \underbrace{E\left[Y_{00} - Y_{01}\right]}_{\text{substitute - counterfactual}} P(D=1) + \underbrace{E\left[Y_{11} - Y_{10}\right]}_{\text{substitute -counterfactual}} P(D=0) \right\}$$

Therefore, the insight here is that we would only get μright when substitutes of the counterfactuals are 0 – this is a very difficult task with observational data.

Are we hopeless then? Not so much, there are ways where we can control or even eliminate selection bias, and we will discuss methods in the following sections to minimize the impact of selection bias. First, let's talk about the golden standard, which is randomized controlled experiment, where assignement is done randomly (so there is no chance for selection bias).

Randomized Controlled Experiment

Key Idea: When we are operating under a randomized controlled experiment (rather than observational data), selection bias disappeared:

$$\begin{aligned} \underbrace{\text{Difference in Group Mean}}_{\text{correlation study}} &= \underbrace{\text{Average Causal Effect}}_{\text{causal inference}} + \text{Selection Bias} \\ &= = \underbrace{\text{Average Causal Effect}}_{\text{causal inference}} + 0 \end{aligned}$$

In the case where treatment assignment is randomized, it means that the treatment D and the outcome Y are independent (the assignment is purely done at random, and no information from Y is taken into account). This is called the **conditional independence assumption (CIA)** and will be true in an truly randomized controlled experiment. Mathematically, this means that $E[Y|D] = E[Y]$, which means that the selection bias term will vanish.

$$\begin{aligned} E[Y_{00} - Y_{01}] &= E[Y(0|D=0)] - E\left[Y(0|D=1)\right] \\ &= E[Y(0) - Y(0)] = 0 \end{aligned}$$

Similarly

$$\begin{aligned} E[Y_{11} - Y_{10}] &= E[Y(1|D=1)] - E\left[Y(1|D=0)\right] \\ &= E[Y(1) - Y(1)] = 0 \end{aligned}$$

As a result, the (substitute - counterfactual) lurking terms will disappear! This is the reason why randomized controlled experiment is the gold statement, because we are gauranteed to eliminate selection bias, achieving Ceteris Paribus. It is also worth noting that:

$$\underbrace{\mu}_{\text{The science question we are going after}} = E[Y_{11} - Y_{00}] = E[Y_{11}] - E[Y_{00}] \approx \underbrace{Avg(Y_{11}) - Avg(Y_{00})}_{\text{What we typically do naively}}$$

comparing the average outcome of those who are treated to the average outcomes of those who are untreated will give us the **right answer**! Unfortunately, we don't always live in a world where randomized controlled experiment is possible, so we must have other tools in our reportorie to battle **selection bias**.

Matching & Regression as Control Matching (Psuedo Ceteris Paribus)

Key idea: Without even knowing regression, a smart approach to make comparison between control & treatment would be to sort users into distinct strata where users are alsmot homogenous in every aspect, except the treatment assignment. In each respective strata, we can then compare the average difference in outcomes (within group difference), and average them to get an estimate of causal inference. Matching helps us to eliminate differences that are unrelated to the treatment of interest, but it's not always easy to do to find good matches for each strata. As a result, we leverage regression techniques to achieve this. Furthermore:

- Regression actively control for variables that might cause selection bias, bring us closer to the truth
- In the case that there are omitted variables, we can still somewhat quantify the "cost" of omission

More concretely, suppose we want to study the impact of type of institutions (private v.s. public universities) on wage. Comparing the wages of those who went to public v.s. private is likely going to be biased because presumbly students with better abilities might tend to apply to top private institution (so self selected into private schools), and also, they are more likely to earn more regardless of the type of the schools they attended. Being a smart researcher, one approach we can take is "matching". For example, in order to control students' abilities, we might want to match students based on similar SAT scores (test taking skills) as well as the schools that they applied and admitted to (academic profiles). If we are lucky enough, we will be able to find enough examples for each strata and do a more rigorous analysis.

Often time though, we might not been able to do this matching, either because it's too manual or we do not have enough sample in each strata. This is where Regression comes in for the rescue (?). The key ingredients in a regression recipe are

- The **dependent** variable: in thise case, student's earnings in life
- The **treatment** (independent) variable: a dummy variable that indicates whether a student is in control or treatment (e.g. public v.s. private)
- A set of **control** (independent) variables: These are variables about the users of which we might want to "control" for before making the comparisons

$$Y_i = \alpha + \gamma I_i \{\text{Treatment?}\} + \sum_{i=j}^{p} \beta_j X_j + \epsilon_i$$

The key here is that γ captures the causal impact of our treatment, conditioning / controlling on all the "control" variables that might vary among samples.

Simplest Case - only Indicator on Y

Regression for Dummies

An important regression special case is regression only on a dummy regressor. In the context of causal inference we can think of this scenario as that Y is only affected by the treatment, and nothing else. This is obviously naive, but it sheds light for us on why regression can be used here. Let's formalize this by introducing the notations: the conditional expectation of Y, given a dummy variable D, can be written as:

$$\begin{aligned} Y_i &= \alpha + \gamma I_i \{\text{Treatment?}\} + \sum_{i=j}^{p} \beta_j 0 + \epsilon_i \\ E[Y|D=0] &= \alpha \\ E[Y|D=1] &= \alpha + \gamma \\ \gamma &= E[Y|D=1] - E[Y|D=0] \end{aligned}$$

These formulas model that being in the treatment group gives us a uplift of β on Yon average, and none of the X's above played a role in affecting Y. Using this notation, we can write:

$$\begin{aligned} E[Y|D] &= E[Y|D=0] + (E[Y|D=1] - E[Y|D=0])\, Z \\ &= \alpha + \gamma Z \end{aligned}$$

This tells us that $E[Y|Z]$ is a linear function of Z, with intercept α and slope β. Because this relationship is linear, this means the regression will be able to fit $E[Y|Z]$ perfectly **(?)**. More importantly, this implies that the regression estimate γwill be the mean difference of the groups, and can be estimated exactly from $Avg(Y|D=1) - Avg(Y|D=0)$ – This ties randomized controlled experiment, regression, and causal impact together nicely! **(?)**

No Omitted Variable Bias

In real life, many X could affect Y, so using regression to control their influence can help us to get closer to the truth. In fact, the typical model of $Y_i = \alpha + \gamma I_i \{\text{Treatment?}\} + \sum_{i=j}^{p} \beta_j X_j + \epsilon_i$ is probably the more common scenario! If one remembers from statistics 101, the coefficient β_j measures the marginal impact of X_j on Y, while holding other variable constant. If the X_j's and $I\{\text{Treatment?}\}$ are orthogonal, I believe we will also get the right estimate for causual impact.

The mechanics for carrying out the β estimation is standard regression, so I will not cover more here.

Omitted Variable Bias

In many scenarios, even if we controlled all the observable variables, there still could be unobservable variable which cannot be measured. The regression version of the selection bias generated by inadequate controls is called **omitted variables bias (OVB),** and it's one of the most important ideas to keep in mind.

To make things more explicitly, consider the example where we have only one control variable A, and we are trying to understand the impact of "not" including it in the regression as a control. We can write the two scenarios as:

$$\begin{aligned} Y_i &= \alpha^l + \gamma^l I\left\{\text{Treatment?}\right\} + \beta A_i + \epsilon_i^l \\ Y_i &= \alpha^s + \gamma s I\left\{\text{Treatment?}\right\} + \epsilon_i^s \text{ where } \left(\epsilon^s = \beta A_i + \epsilon_i^l\right) \end{aligned}$$

The idea here is for us to understand the difference between the two γs. With some mathematical derivation (which we will show later), we can see that

$$\begin{aligned} \text{Treatment Effect on } Y \text{ in Short} = {} & \text{Treatment Effect on } Y \text{ in Long} + \\ & \text{(Relationship from regressing omitted on included (Treatment + other control))} \times \\ & \text{(Effect of omitted on } Y \text{ in Long)} \end{aligned}$$

Intuitively, we would overestimate the causal impact of treatment if we fail to include other control variables! When I stare at this result, one of the ways I gain intuition is the following:

- Treatment itself could affect Y
- The omitted variable can directly affect Y but ALSO Treatment

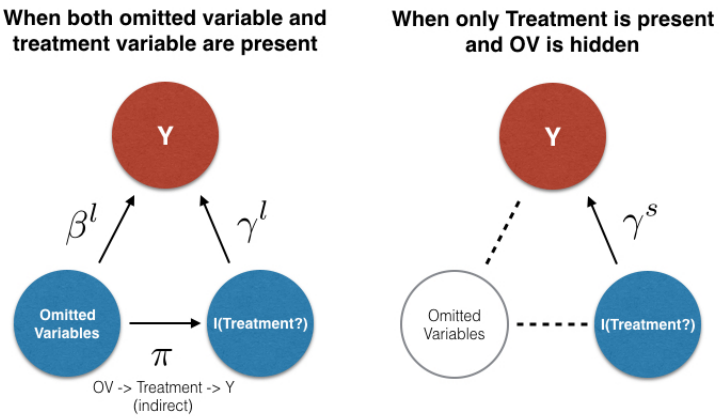

Figure 1:

The rough idea is that had the omitted variable had been present, then we will be able to quantity quite clearly the contribution of both I and A to Y (when both are present). However, in the absence of A, all the credits might be given to I, and we might be overly optimistic of the causal effect of I on Y.

As an example, let's assume that being smart - you will earn higher wage no matter what ($\beta > 0$) AND you are more likely to apply to elite private school ($\pi > 0$), and going to private school makes you more likely to earn a higher wage ($\gamma > 0$), then in the absence of the smartness variable, we might see that smarter people tend to be more likely to go to private (they self selected into that pool), and these people tend to make more because BOTH they are smart and went to prviate school (the wage got an extra push from not only going to private school but also being smart). On the other hand, less intelligent people might have a smaller change going to private school (they self selected into that group), so not only do they earn less (natural) but also they didn't get the push from going to private school (nurture). In the absence of the intelligence factor, we see that the wage gap is larger than what it would have been had we included intelligence as one of the control variable.

More importantly, the formula above tells us that the extra optimism can be measure by $\pi \times \beta^l$. Intuitively, I think this means that γ^s can be broken down to the correct causal impact γ^l + the part that was missed out / omitted

$(\pi \times \beta^l)$ **(?)**. As a result, the degree in which we are affected by OVB is bounded by:

- The correlation between ommited and included (π)
- The impact of the omitted on Y if everything is taken into account (β^l)

$$\begin{aligned} \gamma^s &= \gamma^l + \pi_1 \times \beta^l \\ OVB &= \gamma^s - \gamma^l = \pi_1 \times \beta^l \end{aligned}$$

So Be careful, we could give too much credit to treatment and overlooked the impact of OV on causal inference.

Regression Sensitivity Analysis

Knowing the relationship from above, this can be a guide in practice. More specifically,

- **For observable variables:** We can calculate OVB directly to see if it is necessary to include them as control (i.e. how effective of a control it is, maybe it doesn't matter because OVB is small / Maybe it matters because OVB could be huge).
- **For un observable variables:** If we know that there are unobservable variables that we cannot measure. If we know that either π or β^l (how I don't know, previous research?), then we can somehow quantify the impact of omitting a particular variable. If they are very small, it's probably safe to ignore them.

Appendix For Math

Regression Anatomy

It can be shown that If we have a multivariate regression $Y_i = \alpha + \gamma I_i \{\text{Treatment?}\} + \sum_{i=j}^{p} \beta_j X_j + \epsilon_i$ then each of the β can be expressed as:

$$\beta_k = \frac{Cov\left(Y,\, \tilde{X_k}\right)}{Var\left(\tilde{X_k}\right)}$$

where $\tilde{X_k}$ is the residual of regression X_k on the other $P-1$ X_i's included in the model. Intuitively, this measures the marginal impact of X_k on Y by first removing the association of X_k with the other variables.

OVB Formula

Why is the above formula important? Remember the short and long models:

$$\begin{aligned} Y_i &= \alpha^l + \gamma^l I \{\text{Treatment?}\} + \beta^l A_i + \epsilon_i^l \\ Y_i &= \alpha^s + \gamma s I \{\text{Treatment?}\} + \epsilon_i^s \text{ where } \left(\epsilon^s = \beta A_i + \epsilon_i^l\right) \\ A_i &= \eta + \pi I \{\text{Treatment?}\} + e_i \end{aligned}$$

$$
\begin{aligned}
\gamma^s &= \frac{Cov\left(Y,\ I\left\{\text{treatment?}\right\}\right)}{Var\left(I\left\{\text{treatment?}\right\}\right)} \\
&= \frac{Cov\left(\alpha^l + \gamma^l I\left\{\text{Treatment?}\right\} + \beta^l A_i + \epsilon_i^l,\ I\left\{\text{Treatment?}\right\}\right)}{Var\left(I\left\{\text{treatment?}\right\}\right)} \\
&= \gamma^l \frac{Cov\left(I\left\{\text{Treatment?}\right\},\ I\left\{\text{Treatment?}\right\}\right)}{Var\left(I\left\{\text{treatment?}\right\}\right)} + \beta^l \frac{Cov\left(A_i,\ I\left\{\text{Treatment?}\right\}\right)}{Var\left(I\left\{\text{treatment?}\right\}\right)} + \frac{Cov\left(\epsilon_i^l\, I\left\{\text{Treatment?}\right\}\right)}{Var\left(I\left\{\text{treatment?}\right\}\right)} \\
&= \gamma^l + \beta^l \pi + 0 \\
\gamma^s &= \gamma^l + \pi\beta^l
\end{aligned}
$$

QED.