A Method to Find Real Effects in Econometrics

Tony Feng FENGS19@MAILS.TSINGHUA.EDU.CN
Institute for Interdisciplinary Information Sciences
Tsinghua University
Beijing, 100084, China

Abstract

Causality has gained substantial attention from both academics and the industry. However, it is not easy to reveal the real effect of an explanatory variable. The estimated effect is usually impacted by other explanatory variables. In this paper, we investigate the problem of extra explanatory variables and find that the problem can seriously affect the accuracy of the linear model. We further propose a method called matching (Flinn, 2006) to eliminate the impact of extra explanatory variables. Experimental results on the American wage data set suggest that the matching method can be used in reality.

Keywords: Causality, Matching, Econometrics

1. Introduction

In econometrics, causality has always been a difficult problem to study. The lack of proper understanding of cause and effect often leads to many unexpected problems. In econometrics, we usually refer to such problems as endogenous problems. For example, if we use the Ordinary Least Square regression (OLS) (Davidson et al., 2004) to study the relationship between women's education level and their wages, we will get biased answers. We will find that the relationship is not not as strong as expected because women with less education did not usually work as it is shown in Figure 1[1], which means that we have selection bias in our data set. In order to see the side effect of selection bias, we will do an experiment on a toy data set.

Here, we give an accurate definition of our problem. Suppose our object is to figure out the relationship between random variables X and Y. Our bias model is

$$Y = \beta_0' X + \beta' + \epsilon' \tag{1}$$

However, there are some other explanatory variables, $Z_1, Z_2, \cdots, Z_k$, render other differences except educational level between the samples with $X = x_1$ and $X = x_2$. These explanatory variables are also statistically significant for Y. Therefore, the real causal effect should be

$$Y = \beta_0 X + \beta_1 Z_1 + \beta_2 Z_2 + \cdots + \beta_k Z_k + \beta + \epsilon \tag{2}$$

while $\beta_0 \neq \beta_1$. Therefore, β_0 is not the real effect of X on Y.

1. Codes are published at `https://github.com/fengtony686/Intro_2_Matching/blob/master/Code_Of_Figure1.py`.

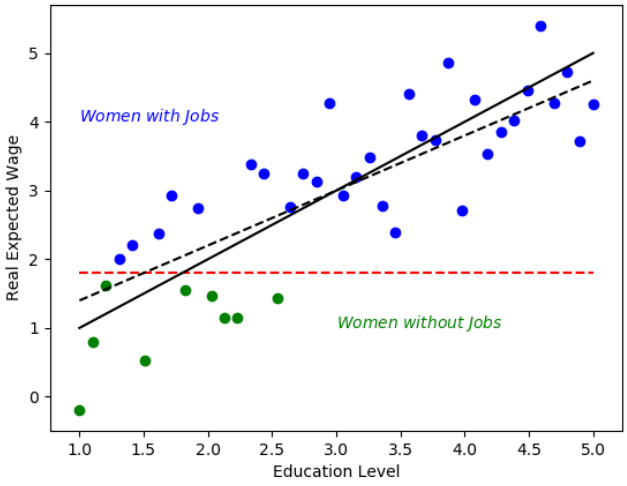

Figure 1: Comparison of results between data sets with selection bias and data sets without selection bias. The solid line is the real causal effect while the dotted line is our results on the data set with selection bias.

R	meanY
0	0.9970268
1	4.0708992

Table 1: The experimental results.

2. Experiment

In this section, we will do an experiment on a toy data set to see the impact of selection bias [2]. We manually set R is a list of 500 normally distributed numbers in $\{0, 1\}$. X is our explanatory variable to Y. We set $Y = 5X + \epsilon_0$. However, our data set is set to have selection bias. We suppose that there are 20% of the subjects acted in the opposite direction due to an explanatory variable Z. That is, 20% of $x \in X$ are different from $r \in R$ while x and r have the same index ($X = ifelse(runif(500) > 0.8, 1-R, R)$). Then we will calculate the mean number of $y \in Y$, which is actually the results of β_0 in $Y = \beta_0 X + \beta_1 + \epsilon$. We get the results in Table 1. Therefore, in the biased OLS model, $\beta_0' \approx 4$ and $\beta_1' \approx 1$. However, the real relationship between X and Y should be $Y = 5X + \epsilon$, which means that the real values of β_0, β_1 should be $0, 1$, correspondingly.

2. Codes are published at `https://github.com/fengtony686/Intro_2_Matching/blob/master/An_Example_Of_Extra_Explanatory_Variables.r`.

The reason of this can be figured out by computing the biased parameters β_0', β' in linear estimation equation (1). Actually, according to equation (1) and (2), we have

$$\beta_0' \approx \frac{y - y'}{x - x'} \tag{3}$$

$$\approx \frac{\beta_0(x - x') + \beta_1(z_1 - z_1') + \beta_2(z_2 - z_2') + \cdots + \beta_k(z_k - z_k')}{x - x'} \tag{4}$$

$$= \beta_0 + \frac{\beta_1(z_1 - z_1') + \beta_2(z_2 - z_2') + \cdots + \beta_k(z_k - z_k')}{x - x'} \tag{5}$$

where $(x, y, z_1, z_2, \cdots, z_k), (x', y', z_1', z_2', \cdots, z_k')$ are two samples in the data set. Therefore, if $\beta_1(z_1 - z_1') + \beta_2(z_2 - z_2') + \cdots + \beta_k(z_k - z_k') \neq 0$, then β_0' is biased. In order to solve this problem and find the real effect of X, we will construct a method in the next section.

3. Matching Method

Suppose we are trying to find the relationship between X and Y. $X = \{0, 1\}$ is the cause and Y is the effect. However, $Z_1, Z_2, \cdots, Z_k$ are all the other explanatory variables for Y. Then our model is to find all the samples which have the same $Z_1, Z_2, \cdots, Z_k$ in the data set. And then we use these samples to estimate the effect of X. The algorithm is as it is shown in Figure 2.

```
CausalEffect(
      // data set of samples
      // each sample has k+2 parameters which are the k+2 variables
    S = {(s.x, s.y, s.z1,s.z2,...s.zk)},
    X,  // cause variable
    Y,  // effect variable
    Z = {Z1,Z2,...,Zk},  // other explanatory variables
)
      // matching process
    non = S.group_by(S.x, S.y, S.z1, S.z2,..., S.zk).summarize(untreated.Y=mean(Y for
X!=1))
    y = S.group_by(S.x, S.y, S.z1, S.z2,..., S.zk).summarize(treated.Y=mean(Y for X!=0))
    join = inner_join(non, y)
return mean(treated.Y), mean(untreated.Y)  // the second output is the real effect of X
on Y
```

Figure 2: Algorithm for the matching method.

Now, we illustrate how the method works. Actually, according to equation (5), if $\beta_1(z_1 - z_1') + \beta_2(z_2 - z_2') + \cdots + \beta_k(z_k - z_k') = 0$, then $\beta_0' = \beta_0$. Our matching method is just simply fix $z_1, z_2, \cdots, z_k$, so $z_i = z_i'$ for $\forall i \in \{1, 2, \cdots, k\}$. Therefore, our estimation of β_0 is the real β_0 and the problem in section 1 has been solved.

union.mean	nonunion.mean
6.687606	6.571178

Table 2: The results of the matching method.

4. Validation

We will use an experiment to show that the matching method can solve some practical problems in this section [3].

Our data set is about the wages of Americans (Yves Croissant, 2020). There are 12 features in the data set as it is shown in figure 3, which are experience, age, blue-collar or not, and so on. The effect variable is the wages of them. What we want to figure out is the relationship between union and wages. Here, if the person was in the union, then $union = 1$, or $union = 0$.

Adopting the matching method, we divided the samples into groups with similar characteristics except for union. In each group, we compute the mean number of wages for samples that $union = 1$ and $union = 0$, correspondingly. Then for all the average number for samples that $union = 1$, we take their average number and for all the average numbers for samples that $union = 0$, we take their average number. Then the results are in Table 2

	exp	wks	bluecol	ind	south	smsa	married	sex	union	ed	black	lwage
1	3	32	no	0	yes	no	yes	male	no	9	no	5.56068
2	4	43	no	0	yes	no	yes	male	no	9	no	5.72031
3	5	40	no	0	yes	no	yes	male	no	9	no	5.99645
4	6	39	no	0	yes	no	yes	male	no	9	no	5.99645
5	7	42	no	1	yes	no	yes	male	no	9	no	6.06146
6	8	35	no	1	yes	no	yes	male	no	9	no	6.17379
7	9	32	no	1	yes	no	yes	male	no	9	no	6.24417
8	30	34	yes	0	no	no	yes	male	no	11	no	6.16331
9	31	27	yes	0	no	no	yes	male	no	11	no	6.21461
10	32	33	yes	1	no	no	yes	male	yes	11	no	6.26340
11	33	30	yes	1	no	no	yes	male	no	11	no	6.54391
12	34	30	yes	1	no	no	yes	male	no	11	no	6.69703
13	35	37	yes	1	no	no	yes	male	no	11	no	6.79122
14	36	30	yes	1	no	no	yes	male	no	11	no	6.81564

Figure 3: Head of the data set.

Therefore, by using the matching method, we can conclude that the direct effect of the union on wage is the difference between *union.mean* and *nonunion.mean*. From the experimental results, it is easy to see the effect of the method.

5. Conclusion and Related Works

In section 1, we have demonstrated the problem we want to solve and we partly solve this problem in section 3 by using the matching method. Actually, as long as the data set is large enough, we can group the samples by the matching method. Then according to the illustration in section 3, the problem can certainly be solved in theory. However, in reality, there are some problems that make the method not effective enough.

3. Codes are published at `https://github.com/fengtony686/Intro_2_Matching/blob/master/An_Experiment_Of_Matching.r`.

Firstly, if the data set is extremely small, we can not find a group of samples that have similar features except the one we want to study. Then the results of this method will be very inaccurate. However, there are some other methods to solve this problem in this situation. One of the most efficient methods is called Heckman Two Stages method (Pagan, 1986), which is based on a higher level of probability knowledge. Also, the instrumental variable is helpful to solve this problem (Greene, 2003).

Secondly, it is hard to find all the explanatory variables of our effect variable. Therefore, it is not easy to fix all the explanatory variables except the one we want to study. However, we may consider fixing the time of the samples unchanged. For the reason that many variables are changing due to the time, if we fix the time, a lot of explanatory variables will be fixed. For example, if we want to study the relationship between altitude and life spans of people, we can fix the time to study this relationship. Many explanatory variables for life spans of people such as GDP, technology, and environment are all fixed when time is fixed. After these, we use the matching method again to fix all the other explanatory variables. Then we can get the real effect. This method is called difference in difference (Bertrand et al., 2004). Moreover, there are some other methods including fixed effects (Christensen, 2002) method and causal diagrams (Pearl, 2009) to overcome the problem of small data sets.

Nevertheless, the matching method is very important when we want to study the causal relationship between two variables. Simultaneously, it is important for all researchers to take the problem of selection bias severely because it can lead to some serious error as we mentioned.

References

Marianne Bertrand, Esther Duflo, and Sendhil Mullainathan. How much should we trust differences-in-differences estimates? *The Quarterly journal of economics*, 119(1):249–275, 2004.

Ronald Christensen. *Plane answers to complex questions*, volume 3. Springer, 2002.

Russell Davidson, James G MacKinnon, et al. *Econometric theory and methods*, volume 5. Oxford University Press New York, 2004.

Christopher J Flinn. Minimum wage effects on labor market outcomes under search, matching, and endogenous contact rates. *Econometrica*, 74(4):1013–1062, 2006.

William H Greene. *Econometric analysis*. Pearson Education India, 2003.

Adrian Pagan. Two stage and related estimators and their applications. *The Review of Economic Studies*, 53(4):517–538, 1986.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Spencer Graves Yves Croissant. Ecdat: Data sets for econometrics, 2020. URL `https://CRAN.R-project.org/package=Ecdat`.