BA Final Report

Ni Jiasheng

May 20,2021

1 Introduction

2 DataSets

2.1 Dataset

The dataset I choose from the documentation to analyze the causal effect of medicaid coverage on the emergency usage per person is listed as follows

Descriptive Variables Dataset:

This dataset lists all exogenous variables imposed on the model that will intervene with the prediction outcome of emergency use, later used for the construction of the OLS regression equation and the balance check. Also, it contains treatment information of the lottery assignment, which will be used for instrumental variable for a more randomized grouping mechanism.

State Program Variables Dataset:

This dataset offers the endogenous variable "ohp all ever matchn 30sep2009"(Indicator for Medicaid coverage) whose effect on the emergency usage per person is what the project is targeted. On top of that, it offers huge amounts of pre-randomized variables for me to build a strong causal inference prediction model.

Emergency Department Data Dataset:

This dataset provides the outcome variable "Emergency use", both during the study period and the pre-randomized period, which I can choose to integrate into my prediction model to enhance the prediction.

2.2 Data Preprocessing

For each dataset, I dummize the categorical variable for the OLS regression model to compute the coefficient. Also, I notice that for the ED dataset, there are some missing values in the sample of 24646 individuals. I choose to use the mean value of that column to fill these missing values.

3 Model Interpretation

3.1 Subsampling

According to the report, the final sample data we work on is the individuals from a specified region in Oregon with particular postcode. So before the OLS regression, we should first perform the validity test and balance check, and finally conduct the regression analysis.

3.2 External Validity Check

For a finely designed causal inference test experiment, we have to make sure that the sample we choose can reflect the distribution of the whole population. To examine the external validity, I compare the proportion of different features in the original sample(74922 individuals who are involved in the medicaid lottery program) and that of the final subsample(24646 individuals). If the difference in the proportion is minor enough, then we can say that the subsample explains the population well. Actually in my examinzation, I try on observed pre-treatment features provided in descriptive vars.dta like "birthyear list","female rate","english list", "self lottery","first day lottery","given phone number","pobox list". The table shows the basic statistics of the balance check.

	Difference between sample mean and population mean	Standard Deviation
Birthday list	-0.33728	0.00354
Female rate	0.17073	-0.00040
English list	0.04674	-0.06225
Self Lottery	-0.01071	0.017052
First day lottery	0.00216	0.00305
Given phone number	0.002161	0.004847
Pobox list	0.090141	0.160377

From the table, I find that there isn't a statistically significant difference between full sample(74922) and subsample(24646), so I draw the conclusion that the choice of subsample satisfies the external validity check.

3.3 Internal Validity Check:Balance Check

Next, I perform the balance check of the pre-randomization variables, which include "birthyear list","female rate","english list", "self lottery","first day lottery","given phone number","pobox list", same as the last section.

The method I choose for balance check is to use the OLS Regression. The formula is given below.

$$y_i = \beta_0 + \beta_1 T_{ij}$$

, where y_i is the features to be tested on balance check. T_{ij} is the treatment of Lottery. The null hypothesis for our OLS Regression T-test is that the mean of the variable to be tested on treatment group and control group are the same. If we obtain a significantly large p-value, larger than 0.05, then we can say that the ED sample is balanced on this particular feature.

The table below summarizes the results of the balance check I preform on these features.

	Control Mean	Diff between Treat and Control	P-value(95% confidence)	Comment
Birthday list	1968.3354(0.098)	0.1619(0.157)	0.303	Balanced
Female rate	0.5535(0.004)	-0.0175(0.006)	0.007	Unbalanced,insignificant difference
English list	0.8752(0.003)	-0.0270(0.004)	0.000	Unbalanced,insignificant difference
Self Lottery	0.0713(0.002)	0.0783(0.004)	0.000	Unbalanced, insignificant difference
First day lottery	0.0905(0.002)	0.0059(0.004)	0.122	Balanced
Last day lottery	0.0418(0.002)	-0.0015(0.003)	0.562	Balanced
Given phone number	0.8662(0.003)	0.0093(0.004)	0.035	Unbalanced, insignificant difference
Pobox list	0.0265(0.001)	-0.0003(0.002)	0.879	Balanced

From the table above, we find that for the features "Birthday List","First day lottery", and "Last Day Lottery","Pobox List" are balanced. For the rest of the features, though there is significant evidence that the sample are not balanced, the difference between the treatment group and control group are actually very small. Therefore, we conclude that the sample to some extent satisfy the internal validity.

Additionally, we examine the difference between the treatment and control group using all the pre-treatent combined. The table below reports the pooled F statistics and P values from testing treatment-control balance on sets of variables jointly.

	F-statistic
F-statistic for lottery variables	46.27
F statistic for pre-randomization versions of the outcome variables	1.225
F statistic for lottery list and pre-randomization variables	9.118

As we can see, when we consider the pre-randomization variables inside our balance check, the difference in the treatment and control group are not that significant(F-statistic is smaller than 10).Thus, we can conclude that the treatment and control group are balanced, which will contribute to the robustness of our future analysis.

4 OLS Regression

4.1 Effect of lottery selection on ED usage

The formula I choose for the OLS Regression as implied by the report is:

$$y_{ih} = \beta_0 + \beta_1 Lottery_h + \beta_2 * X_{ih} + \beta_3 * V_{ih} + \epsilon_{ih} \quad (1)$$

where Lottery takes value in 0,1 indicating treatment group or control group of the medicaid experiment. i indicates individuals, h indicates household.

X_i is the number of people in a household.

V_{ih} is the variables used to enhance the prediction power of the OLS Regression model.

What I get for the difference between the ED use outcome of treatment group and control group judged by Lottery distribution(0 or 1) is -0.0115 with standard deviation at 0.031 and p-value at 0.716, which shows that the treatment group and control group is well-balanced. In other words, there is no correlationship between the instrumental variable and the outcome ED use, which is exactly what we want for a well-randomized experimental design.

However, this result only represents the intent-to-treatment ATE, or local average treatment effect since we are assuming that people who are qualified for medicaid by the lottery selection will definitely enroll for the medicaid program. In other words, we are assuming the sample individuals to be compliers, which is different from the real-time decision-making process. Therefore, this analysis could bring about some volatility.

4.2 First stage IV testing

To figure out the causal effect of MEDICAID on the ED usage per person, I use the instrumental variables method for a more accurate interpretation. For the target of our project, the effect of Medicaid Coverage(Measured by insurance coverage) on ED usage per person. I first test the choice of IV variables, which is Lottery selection.

The assumption I take so as to fullfill the validity of IV are strong first stage assumption which I examined below and exclusion restriction assumption so that the IV is useful for our regression results.

The first stage testing relies on the following equation:

$$MEDICAID_{ih} = \delta_0 + \delta_1 * LOTTERY_h + \delta_2 * X_{ih} + \delta_3 * V_{ih} + \epsilon_{ih} \quad (2)$$

X_{ih} and V_{ih} are defined in equation (1).

For X_{ih}, I choose the number of people in a household, ranging from 1 to 3.

For V_i, I choose all the pre-randomzied variables from "State Program Dataset", which includes the pre-randomized outcome of ED usage. I think this could bring improvements on the credibility of the prediction model.

Finally, from the OLS Regression, I get that the F-statistic for the IV variable "LOTTERY" is around 56.42, which is larger than 10, which implies that "LOTTERY" is a good choice of instrumental variables for predicting the effect of MEDICAID coverage on ED usage per person.

4.3 Effect of Medicaid Coverage on ED usage through IV

The casual effect is based on the following equation:

$$y_{ih} = \pi_0 + \pi_1 * MEDICAID_{ih} + \pi_2 * X_{ih} + \pi_3 * V_{ih} + \epsilon_{ih} \quad (3)$$

V_{ih} are defined in equation (2). The causal effect of MEDICAID coverage on emergency use is interpreted by the coefficient π_1. And for the X_{ih}, I include all the descriptive variables including ”BirthYear”,”female rate”,”English List”, ”self lottery list”,”first day lottery”,”given phone number”,”pobox list” for better prediction.

Finally, from the OLS Regression, the casual effect of the MEDICAID Coverage on ED use per person is around 0.3302 with the standard deviation at around 0.1067 and p-value around 0.0020, which perfectly shows that MEDICAID coverage has positive effect on the ED use per person.

5 Additional results

In additional to the casual effect of the MEDICAID on ED use per person for all the visits. We always regress the casual effect of MEDICATD on the ED use per person on some subsamples, which are those who visit the emergency department in the pre-treatment period(from January 1, 2007 to March 9, 2008) zero times, once, twice, and five times. If all the casual effect are positive, then there is relatively strong evidence that MEDICAID can have positive effect on ED use per person. The formula is defined in Equation (1) and the results are shown in the table below.

	Casual Effect	P-value
No visits	1.26	0.000
One visit	2.8095	0.000
Two visits	6.8516	0.000
Five+ visits	12.687	0.000

These results prove the positive casual effect of MEDICAID program on ED use per person in a way that Simpson's Paradox doesn't seem likely to be happen.

6 Results and conclusion

6.1 Results

From the analysis above, there is strong evidence that medicaid coverage(measured by insurance coverage) can have positive influence on the ED usage per person.

6.2 Reflection

From this project, if we want to figure the relationship between two factors, how to perform an effective experimental design is very important.

We may start from external validity and internal validity test, using different pre-randomization features and controlling them in the OLS regression model. On top of that, we may apply instrumental variables under two assumptions difficult to verify just through data to capture the closest form of the casual effect of MEDICAID coverage on the ED usage per person. Directly performing an OLS Regression on the endogenous, exogenous variables and outcome variable are not feasible, which would result in over or under estimation of the outcome, resulting in Simpson's Paradox.

However, due to lots of other un-observable confounding features interfering with our final results, we may only deduce a partially referential casual effects from the limited data. But it has always been much better than inferring the casual effect from pure guess.