CUHK Business School

Causal Inference for the Impact of the Medicaid Program

DOTE 6756: Machine Learnong for Business
Currently under the Course Title "Business Intelligence Technologies and Applications"

Due at 11:59PM, on Monday, March 3, 2025

Please complete the following task and submit (a) a brief PDF report articulating your approach and results and (b) a Jupyter Notebook containing your analysis on Blackboard. We also design several sub-questions which may guide your analysis. This project counts **10%** towards the final grade for this course, which means the two projects altogether count **22%** and you have a **2%** extra-credit for the course projects. You are allowed to discuss with anyone about this project, but you should perform the analysis and write the report on your own. Please make the PDF report, without compromising on quality and clarity, as concise as possible.

Background

In 2008, a group of uninsured low-income adults in Oregon was selected by lottery to be given the chance to apply for Medicaid. This lottery provides a unique opportunity to gauge the effects of expanding access to public health insurance on the health care use, financial strain, and health of low-income adults using a randomized controlled design. The Oregon Health Insurance Experiment followed and compared those selected in the lottery (treatments) with those not selected (controls). You may visit this website https://www.nber.org/programs-projects/projects-and-centers/oregon-health-insurance-experiment?page=1&perPage=50 for more information about this experiment and the subsequent research and public policy that emerged based on this experiment.

You job in this project is to estimate the causal effect of being selected by the lottery and enrolling into the medicaid program on emergency department utilization. You may read Taubman et al. (2014) for some background information (available in the Reference folder on GitHub).

Data

Please download the datasets and the relevant documentations from GitHub. The datasets are stored in `.dta` format (the data format of `Stata`; so you can find some code to manipulate the data and do the analysis in `Stata` on GitHub as well). You can load `.dta` data as a Pandas data frame using the function `pd.read_stata("oregonhie_descriptive_vars.dta")`.

Randomization and Treatment Assignment. Oregon selected roughly 30,000 individuals by lottery from a waiting list of about 90,000 for an otherwise closed Medicaid program. The state conducted eight lottery drawings from March through September 2008. Selected individuals won the opportunity – for themselves and any household member – to APPLY for health insurance benefits through a Medicaid program called Oregon Health Plan Standard (OHP Standard). OHP

Standard provides benefits to low-income adults who are not categorically eligible for Oregon's traditional Medicaid program (OHP Plus); to be eligible individuals must be adults ages 19 – 64, not otherwise eligible for Medicaid or other public insurance, Oregon residents, U.S. citizens or legal immigrants, have been without health insurance for six months, have income below the federal poverty level, and have assets below $2,000. The randomly selected individuals chosen by the lottery who completed the application process and met the eligibility criteria were enrolled in OHP Standard. Following some selection rules, the data set contains only these 74,922 individuals. Of these individuals, 29,834 were selected as treatments (i.e. won the lottery and were given the chance to apply for health insurance); **treatment** status is indicated by the variable `treatment` in `oregonhie_descriptive_vars.dta`.

Crucially, the lottery selected individuals, but the opportunity to apply for health insurance was extended to **all household members** of lottery winners: **treatment selection is random only conditional on the number of household members on the waiting list** (this is given by the variable `numhh_list` in `oregonhie_descriptive_vars.dta`. For example, an individual could sign up his or herself as well as a spouse for the lottery, and both have equal probability of being chosen. Thus, this person and his or her spouse are twice as likely to win the opportunity to apply for health insurance as someone who only added their own name to the list, without adding other household members. In short, those in a larger household are more likely to be selected into the **treatment** condition.

Merging Datasets. All datasets contain observations at the individual level. Observations can be linked across different `.dta` files by the unique identifier `person_id`, which appears in all datasets. No other variable appears across multiple datasets.

Data Set Descriptions. Below we describe the 3 datasets concerned in this task:

`oregonhie_descriptive_vars.dta`

This data set contains demographic characteristics that were recorded when individuals signed up for the lottery and lottery selection. You may refer to the code book `oregonhie_descriptivevars_codebook.pdf` for descriptions of the variables in this data set.

`oregonhie_stateprograms_vars.dta`

This data set contains information from the state of Oregon on individuals' participation in the following state programs: Medicaid, the Supplemental Nutrition Assistance Program (SNAP), and Temporary Assistance to Needy Families (TANF). You may refer to the code book `oregonhie_stateprograms_codebook.pdf` for descriptions of the variables in this data set.

`oregonhie_ed_vars.dta`

This data set contains variables derived from administrative data of all visits to twelve hospital emergency departments in the area of **Portland, Oregon**. You may refer to the code book `oregonhie_ed_codebook.pdf` for descriptions of the variables in this data set.

For more detailed descriptions of the entire data set, please read the documents `ohie_startguide.pdf` and `ohie_userguide.pdf`. All the data and their descriptions can be found here: https://www.nber.org/research/data/oregon-health-insurance-experiment-data.

Questions

You job in this task is to estimate **the causal effect of being selected by the lottery and enrolling into the medicaid program on emergency department utilization**. Please address the following questions. You may need to merge different data sets together.

(a) (3 points) **Initial Data Pre-processing and Balance Check**. Because the individuals selected by the lottery have the opportunity to apply for the OHP Standard program, you need to create dummy variables for the number of people in household on lottery list. Why should we use dummy instead of numeric variables for this setting? Because the ED visit data is only available for the Portland area, we will mainly work with the data observations in this area. Please use the OLS approach to check the balance of the treatment and control groups for the individuals in the data sample. Specifically, you need to regress the variable which you want to conduct balance check on (i) the treatment variable, and (ii) the dummy variables for the number of people in household on lottery list. Please conduct balance check for the following variable with the full OHIE data sample (N=74,922):

- Included in the emergency department (ED) sample, i.e., the Portland area, (N=24,646).

Please conduct balance check for the following variable with the ED data sample (N=24,646):

- Year of birth
- Female
- Signed up self for lottery
- Any ED visit, pre-randomization (censored)
- Number of ED visits, pre-randomization (censored)

(b) (3 points) **Causal Effect of Being Selected by Lottery**. Next, for the data sample in the Portland area (N=24,646), please estimate the causal effect of being selected by the lottery on the following outcome:

- Whether an individual was enrolled in any Medicaid program (including the OHP Standard) between the earliest notification date in the sample (10 March 2008) and 30 September 2009.

Please include as appropriate necessary features into your regression model. In particular, do we need to include the dummy variables for the number of people in household on lottery list? Why or why not? Please discuss/justify your choice of features included in the regression model. What is the average treatment effect of being selected by the lottery on being enrolled in any Medicaid program?

(c) (4 points) **Causal Effect of Enrolling into a Medicaid Program on ED Visits.** Estimate the average treatment effect of enrolling into a medicaid program on (i) the probability of any ED visits during the study period and (ii) the (censored) number of ED visits in the study period. Again, you need to include certain features into the regressions to remove the bias and/or reduce the variance of your estimation. Please articulate your identification strategy, present your model specification, and report your estimation results, including the 95% confidence intervals.

Hints

1. This project is essentially a replication of the paper Taubman et al. (2014).
2. If you are more familiar with *R* or `Stata`, feel free to use them to finish your analysis. In this case, you also need to submit your code on Blackboard.

Reference

Taubman, S., H. Allen, B. Wright, K. Baicker, A. Finkelstein. 2014. Medicaid Increases Emergency-Department Use: Evidence from Oregon's Health Insurance Experiment. *Science*. **343**, 263-268.