EC 425/525
Spring 2021

Professor Eric Zou
Department of Economics, University of Oregon

Empirical Project 1
The Oregon Health Plan Experiment (OHP)

Due at midnight (Pacific Time) on Thursday, April 29, 2019

In this empirical project, you will analyze experimental data from the Oregon Health Plan Experiment (OHP) which we discussed in class. In particular, you will analyze real data from in-person interviews conducted in the Portland, Oregon, metropolitan area. The interviews were conducted about 25 months after the OHP lottery. The interview included detailed questionnaires on insurance coverage, health care use, health status, inventory of medications. The interview also contains questions that help assess depression and health-related quality of life. Finally, performance of anthropometric and blood-pressure measurements is taken, and dried blood spots are obtained. For more information about OHP and the in-person interviews, see here and here.

Instructions

Please submit your Empirical Project on Canvas. Your submission should be a single PDF file containing three parts:

1. A 4-6 page research summary (double spaced and including references, graphs, and tables)
2. A copy of do-file with your STATA code or an .R script file with your R code
3. A copy of the log file (or screen output) of your STATA or R output

Specific questions to address in your research summary

1. Explain the difference between the variables *treatment* and *ohp_all_ever_survey*. Explain why *treatment* is the treatment variable (D_i), rather than *ohp_all_ever_survey*.

2. Provide evidence that the OHP lottery really did randomly assign individuals to treatment and control groups. Similar to Table 1 in Taubman et al. (2014), please create a nicely formatted table that reports means of 4 to 6 relevant characteristics for individuals in the *control group*.

   Note: Part of this question is to get you to think about which variables should be balanced in a randomized experiment. You need to read carefully through all the variables in the dataset (documentation attached at the end of this file) and decide which 4 to 6 you will summarize.

3. For each of the variables you summarized above, calculate:

   (i) the difference between the mean in the treatment group and the mean in the control group;

(ii) the standard error for the difference in means.

Add these as columns two and three to the table you started in question 2.

4. Is the balance table consistent with individuals having been randomly assigned to treatment group and control groups? Why or why not?

5. Estimate the compliance rate for the OHP experiment. That is, what is the effect of being assigned to the treatment group on the probability of being enrolled in Medicaid?

   Hint: For this question and question 7, you can use the same regression as in question 3, just changing the dependent variable.

6. What is the intent-to-treat (ITT) effect of the OHP experiment on health outcomes? Please create a nicely formatted table that reports ITT estimates on 4 to 6 relevant health outcomes. Again, part of this question is to get you to think about which 4 to 6 variables could be used as health outcome variables.

7. What is the "treatment on the treated" effect (ATET) of the OHP experiment, i.e. the effect among those who applied for Medicaid? Estimate it for every health outcome you chose in question 6 and provide some intuition for the calculation of this estimate.

8. Do you have to worry about attrition bias in analyzing this data? Explain why or why not.

9. Suppose that you are submitting these results to a general interest journal such as *Science* for publication. Write an abstract of 200 or fewer words describing what you have found in your analysis of the OHP data, similar to the abstract in Taubman et al. (2014).

DATA DESCRIPTION, FILE: ohp.dta

The data consist of n = 12,229 individuals involved in OHP. Each individual is assigned a scrambled ID to protect privacy.

Variable Definitions in ohp.dta

Variable	Definition
person_id	Scrambled individual identifier
household_id	Scrambled household identifier
weight_total_inp	Survey weights (inverse probability weighting)
treatment	1 if OHP lottery winner, 0 otherwise
age_inp	Age
bp_sar_inp	Systolic blood pressure, average of three consecutive readings
chl_inp	Total cholesterol (dried blood spot test)
dep_dx_post_lottery	Diagnosed with depression after the lottery
dep_dx_pre_lottery	Diagnosed with depression before the lottery
dia_dx_post_lottery	Diagnosed with diabetes after the lottery
dia_dx_pre_lottery	Diagnosed with diabetes before the lottery
doc_num_mod_inp	Num. of doctor's visits, truncated at 2*99th percentile
edu_inp	Education: highest completed (1 = less than high school; 2 = high school diploma; 3 = post high school, not 4-year college; 4 = 4-year college degree or more)
gender_inp	1 if female
hbp_dx_post_lottery	Diagnosed with hypertension after the lottery
hbp_dx_pre_lottery	Diagnosed with hypertension before the lottery
hispanic_inp	Hispanic/Latino
itvw_english_inp	Interviewed in English
numhh_list	Number of people in household on lottery list
ohp_all_ever_survey	1 if ever enrolled in Medicaid
race_black_inp	Race/Ethnicity is Black
race_nwother_inp	Race/Ethnicity is Non-White Other
race_white_inp	Race/Ethnicity is White
rx_num_mod_inp	Number of prescription medications currently taking

Example R Commands

R command	Description

`*Subset data`
`ohp_cntrl <- subset(ohp, treatment ==0, select =`
`c(xvar1, xvar2, xvar3))`
 	
`*Report summary statistics`
`summary(ohp_cntrl)`
	Subsets the data to the variables *xvar1*, *xvar2*, and *xvar3* for observations with *treatment* equal to 0. Reports summary statistics for this data frame.

`#Install and load sandwich and lmtest packages`
`install.packages("sandwich")`
`install.packages("lmtest")`
`library(sandwich)`
`library(lmtest)`
 	
`#Regression with homoskedasticity-only standard errors`
`mod1 <- lm(yvar~zvar1+wvar, data = ohp)`
`summary(mod1)`
 	
`#Report heteroskedasticity robust standard errors`
`coeftest(mod1, vcov = vcovHC(mod1, type="HC1"))`
	Estimates multivariate regression of *yvar* on an intercept, *zvar*, and *wvar*, with heteroskedasticity-robust standard errors.

`*Method 1`
`mod1 <-lm(yvar ~ zvar+wvar, data = ohp )`
`mod2 <-lm(xvar ~ zvar+wvar , data = ohp )`
 	
`mod1$coef[2]/mod2$coef[2]`
 	
`*Method 2`
`install.packages("AER")`
`library(AER)`
`tot <- ivreg(yvar ~ xvar + wvar`
`|zvar + wvar ,data=ohp)`
 	
`summary(tot, vcov = sandwich,`
`df = Inf, diagnostics = TRUE)`
	These commands show how to estimate the ratio between the coefficient on zvar from two different regressions that have all the same right-hand side variables, and different dependent variables.
	
	The first method refers to the elements of the vector `coef` that contains the relevant coefficients and displays the ratio between them.
	
	The second method uses the `ivreg` command to estimate the ratio in one step.

`ohp$difference <- ohp$xvar1 – ohp$xvar2`	Generates a new variable that equals the difference between *xvar1* and *xvar2*