Problem set 3

PPHA 31102 Statistics for Data Analysis II: Regressions

Jay Ballesteros

February 17,2026

1. The Oregon Health Insurance Experiment, Revisited (16 points)

For this assignment, you can refer to any output that comes from the `lm()` function in R.

Question 1 (3 points)

In Problem Set #1, we found that one of the baseline characteristics, `numhh_list`, was statistically significant from zero at the 5 percent level. Oh no, did randomization fail? It turns out that the researchers expected this. The reason this happened is because treatment was assigned at the household level, and households with more eligible individuals had more chances to win the lottery. Fortunately, we can easily deal with this violation of balance using multivariate regression techniques!

The regression controlling for family size is given as follows:

$$Y = \beta_0 + \beta_1 Treated + \beta_2 numhh\_list + u$$

Recall in problem set #1, you ran the following regression:

$$Y = \beta_0' + \beta_1' Treated + v$$

For each of the five outcomes, calculate the bias from not including `numhh_list` as a control, filling in the table below. Are any of these biases quantitatively large enough to fundamentally change any of your qualitative conclusions about the OHIE?

Answer

Outcome	Bias
count_visit_dr	-0.1978184
count_visit_er	-0.03814492
out_of_pocket_spend	-5.841552
health_score	-0.00206248
happy	0.007311952

Code

```
true_model <- lm(count_visit_dr ~ treated + numhh_list, data=df, na.action = na.omit)
under_model <- lm(count_visit_dr ~ treated, data=df, na.action = na.omit)

bias <- coef(under_model)["treated"] - coef(true_model)["treated"]
bias
```

```
##    treated
## -0.1978184
```

```
true_model <- lm(count_visit_er ~ treated + numhh_list, data=df, na.action = na.omit)
under_model <- lm(count_visit_er ~ treated, data=df, na.action = na.omit)

bias <- coef(under_model)["treated"] - coef(true_model)["treated"]
bias
```

```
##     treated
## -0.03814492
```

```
true_model <- lm(out_of_pocket_spend ~ treated + numhh_list, data=df, na.action = na.omi
under_model <- lm(out_of_pocket_spend ~ treated, data=df, na.action = na.omit)

bias <- coef(under_model)["treated"] - coef(true_model)["treated"]
bias
```

```
##   treated
## -5.841552
```

```
true_model <- lm(health_score ~ treated + numhh_list, data=df, na.action = na.omit)
under_model <- lm(health_score ~ treated, data=df, na.action = na.omit)

bias <- coef(under_model)["treated"] - coef(true_model)["treated"]
bias
```

```
##     treated
## -0.00206248
```

```r
true_model <- lm(happy ~ treated + numhh_list, data=df, na.action = na.omit)
under_model <- lm(happy ~ treated, data=df, na.action = na.omit)

bias <- coef(under_model)["treated"] - coef(true_model)["treated"]
bias
```

```
##     treated
## 0.007311952
```

Question 2 (3 points)

Let's look at which groups increased their doctor office visits the most in response to the treatment. Fill in the table below by running separate regressions of `visit_dr` on `treated`, and controlling for `numhh_list`, i.e. by running model (1) above, for each of the groups listed in the table.

Discuss your findings: which group has the largest estimated treatment effects, which group has the smallest? Be sure to also consider the statistical significance.

Group	$\hat{\beta}_1$	S.E.($\hat{\beta}_1$)
female==0		
female==1		
age<50		
age>50		
race_white==0		
race_white==1		
health_baseline==0		
health_baseline==1		

```r
# Your code here
```

Question 3 (3 points)

Returning to the full data and still focusing on the number of doctor office visits, let's also try controlling for education.

Using information in the variables `hs_degree` and `college_degree`, create a new indicator variable if someone DOES NOT have a high-school degree, call this new variable `NO_hs_degree`.

Try running each of the following regressions. Discuss how your estimated treatment effect, the precision of this estimate (standard error), and R^2 changes from just including `numhh_list`, i.e. model (1) above. If you cannot estimate a coefficient, explain why.

$$count\_visit\_dr = \beta_0 + \beta_1 treated + \beta_2 numhh\_list + \beta_3 hs\_degree + \beta_4 college\_degree + u \quad (3)$$

$$count\_visit\_dr = \beta_0' + \beta_1' treated + \beta_2' numhh\_list + \beta_3' NO\_hs\_degree + \beta_4 hs\_degree + \beta_5 college\_degree + u$$

```
# Your code here
```

Question 4 (2 points)

Now, let's try including all the baseline characteristics as controls to the regression. Rerun the regression of `count_visit_dr` on `treated`, and control for:

- `numhh_list`
- `female`
- `age`
- `race_white`
- `hs_degree`
- `college_degree`
- `health_baseline`

How does your estimated treatment effect, the precision of this estimate (standard error), and R^2 change from the model just including `numhh_list`?

```
# Your code here
```

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Donec a diam lectus. Sed sit amet ipsum mauris. Maecenas congue ligula ac quam viverra nec consectetur ante hendrerit. Donec et mollis dolor. Praesent et diam eget libero egestas mattis sit amet vitae augue. Nam tincidunt congue enim, ut porta lorem lacinia consectetur. Donec ut libero sed arcu vehicula ultricies a non tortor. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aenean ut gravida lorem.

Question 5 (1 point)

Do you think we should include all these other baseline characteristics as controls to ensure that the treatment effects are unbiased, or is it sufficient to just control for `numhh_list`? Explain.

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Donec a diam lectus. Sed sit amet ipsum mauris. Maecenas congue ligula ac quam viverra nec consectetur ante hendrerit. Donec et mollis dolor. Praesent et diam eget libero egestas mattis sit amet vitae augue. Nam tincidunt congue enim, ut porta lorem lacinia consectetur. Donec ut libero sed arcu vehicula ultricies a non tortor. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aenean ut gravida lorem.

Question 6 (1 point)

Do you think we should include all these other baseline characteristics as controls to improve the precision of our estimated treatment effects? Explain.

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Donec a diam lectus. Sed sit amet ipsum mauris. Maecenas congue ligula ac quam viverra nec consectetur ante hendrerit. Donec et mollis dolor. Praesent et diam eget libero egestas mattis sit amet vitae augue. Nam tincidunt congue enim, ut porta lorem lacinia consectetur. Donec ut libero sed arcu vehicula ultricies a non tortor. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aenean ut gravida lorem.

Question 7 (3 points)

Using the model with the full set of controls estimated in Question 4, conduct a hypothesis test that the treatment effect on the number of doctor office visits is equal to the effect of having a high school degree.

Note: In class, we discussed two different ways to conduct this test. You can choose either method.

To receive full credit, you should code this test manually in the manner we discussed in class and not use the `anova()` function or other such functions to perform hypothesis testing.

```
# Your code here
```

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Donec a diam lectus. Sed sit amet ipsum mauris. Maecenas congue ligula ac quam viverra nec consectetur ante hendrerit. Donec et mollis dolor. Praesent et diam eget libero egestas mattis sit amet vitae augue. Nam tincidunt congue enim, ut porta lorem lacinia consectetur. Donec ut libero sed arcu vehicula ultricies a non tortor. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aenean ut gravida lorem.