Problem set 3

PPHA 31102 Statistics for Data Analysis II: Regressions

Jay Ballesteros

February 18, 2026

1. The Oregon Health Insurance Experiment, Revisited

Q1. Question

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?

Q1. Answer

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

Outcome	Bias
happy	0.007311952

In the four out of five outcomes. (`count_visit_dr`,`count_visit_er`, `health_score`, `out_of_pocket_spend`), the bias from not including `numhh_list` is negative, therefore underestimated the true effect of the treatment. In the case of `happy`, the bias is positive, therefore overestimated the true effect of the treatment.

The main take-away of the results, is than in four outcomes, the bias was close to 0, so no significant cause of concern. Nonetheless, the `out_of_pocket_spend` is the only outcome that has a bias that is large enough to change the conclusions about this indicator, as the bias is around -5.84, which is significant. This means that not controlling for `numhh_list` would have led us to underestimate the true effect of the treatment on out-of-pocket spending by a significant amount.

Q1. 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.omit)
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
```

```r
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
```

Q2. Question

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.

Q2. Answer

Group	$\hat{\beta}_1$	S.E.($\hat{\beta}_1$)	p-value
female==0	0.3420393	0.2836245	0.2278876
female==1	0.7926379	0.3118582	0.01105449
age<50	0.6440451	0.2588436	0.01285882
age=>50	0.4097361	0.3933661	0.2976653
race_white==0	0.1638626	0.3303053	0.6198567
race_white==1	0.8066601	0.2760115	0.003480989
health_baseline==0	0.6450163	0.2393665	0.0070589
health_baseline==1	0.4180334	0.4681090	0.3719088

The group with the largest estimated treatment effect is `race_white==1` with a $\hat{\beta}_1$ of 0.8066601, while the group with the smallest estimated treatment effect is `race_white==0` with a $\hat{\beta}_1$ of 0.1638626.

The only groups that show statistically significant treatment effects at the $\alpha = 0.05$, are `female==1`, `age<50`, `race_white==1`, and `health_baseline==0`.

Q2. Code

```
df_q2 <- df %>%
  filter (female == 0)
model <- lm(count_visit_dr ~ treated + numhh_list, data=df_q2, na.action =
↪  na.omit)
coef(summary(model))["treated", 1:4]
##   Estimate Std. Error    t value   Pr(>|t|)
##  0.3420393  0.2836245  1.2059584  0.2278876
```

```
df_q2 <- df %>%
  filter (female == 1)
model <- lm(count_visit_dr ~ treated + numhh_list, data=df_q2, na.action =
↪  na.omit)
coef(summary(model))["treated", 1:4]
##   Estimate Std. Error    t value   Pr(>|t|)
## 0.79263790 0.31185824 2.54166088 0.01105449
```

```r
df_q2 <- df %>%
  filter (age < 50)
model <- lm(count_visit_dr ~ treated + numhh_list, data=df_q2, na.action =
↪  na.omit)
coef(summary(model))["treated", 1:4]
##   Estimate Std. Error    t value   Pr(>|t|)
## 0.64404513 0.25884357 2.48816356 0.01285882
```

```r
df_q2 <- df %>%
  filter (age >= 50)
model <- lm(count_visit_dr ~ treated + numhh_list, data=df_q2, na.action =
↪  na.omit)
coef(summary(model))["treated", 1:4]
##   Estimate Std. Error    t value   Pr(>|t|)
##  0.4097361  0.3933661  1.0416152  0.2976653
```

```r
df_q2 <- df %>%
  filter (race_white == 0)
model <- lm(count_visit_dr ~ treated + numhh_list, data=df_q2, na.action =
↪  na.omit)
coef(summary(model))["treated", 1:4]
##   Estimate Std. Error    t value   Pr(>|t|)
##  0.1638626  0.3303053  0.4960944  0.6198567
```

```r
df_q2 <- df %>%
  filter (race_white == 1)
model <- lm(count_visit_dr ~ treated + numhh_list, data=df_q2, na.action =
↪  na.omit)
coef(summary(model))["treated", 1:4]
##    Estimate  Std. Error     t value    Pr(>|t|)
## 0.806660125 0.276011485 2.922559996 0.003480989
```

```r
df_q2 <- df %>%
  filter (health_baseline == 0)
model <- lm(count_visit_dr ~ treated + numhh_list, data=df_q2, na.action =
↪  na.omit)
coef(summary(model))["treated", 1:4]
##   Estimate Std. Error    t value   Pr(>|t|)
##  0.6450163  0.2393665  2.6946804  0.0070589
```

```r
df_q2 <- df %>%
  filter (health_baseline == 1)
model <- lm(count_visit_dr ~ treated + numhh_list, data=df_q2, na.action =
↪  na.omit)
coef(summary(model))["treated", 1:4]
##   Estimate Std. Error    t value   Pr(>|t|)
```

```
##  0.4180334  0.4681090  0.8930257  0.3719088
```

Q3. Question

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.

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

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

Q3. Answer

	Model (1)	Model (3)	Model (4)
$\hat{\beta}_1$ (treated)	0.5935	0.5745	0.5745
S.E.($\hat{\beta}_1$)	0.2166	0.2163	0.2163
R^2	0.006276	0.009592	0.009592

The coefficient of treated changes slightly from 0.5935 in model (1) to 0.5745 in models (3) and (4). This might indicate that including additional education controls does not have an important effect on the estimated treatment effect. The standard error faces a similar conclusion, as it remains almost unchanged across the three models. For the R^2, we see a slight increase from 0.006276 in model (1) to 0.009592 in models (3) and (4), meaning that education explains a small portion of the variation in the number of doctor office visits, but it does not significantly improve the model's fit.

Q3. Code

```
df <- df %>%
  mutate(NO_hs_degree = ifelse(hs_degree == 0 & college_degree == 0, 1, 0))
```

```
model_1 <- lm(count_visit_dr ~ treated + numhh_list, data=df, na.action =
↪  na.omit)
summary(model_1)
##
```

```
## Call:
## lm(formula = count_visit_dr ~ treated + numhh_list, data = df,
##     na.action = na.omit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -6.760  -5.760  -3.760  -0.028 137.834
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.3039     0.3368  24.658  < 2e-16 ***
## treated       0.5935     0.2166   2.740  0.00614 **
## numhh_list   -2.1378     0.2495  -8.567  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.86 on 12155 degrees of freedom
##   (71 observations deleted due to missingness)
## Multiple R-squared:  0.006276,   Adjusted R-squared:  0.006113
## F-statistic: 38.38 on 2 and 12155 DF,  p-value: < 2.2e-16
```

```
model_3 <- lm(count_visit_dr ~ treated + numhh_list + hs_degree + college_degree,
↪  data=df, na.action = na.omit)
summary(model_3)
##
## Call:
## lm(formula = count_visit_dr ~ treated + numhh_list + hs_degree +
##     college_degree, data = df, na.action = na.omit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -8.014  -5.568  -3.568   0.158 138.432
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      6.9842     0.4178  16.715  < 2e-16 ***
## treated          0.5745     0.2163   2.657   0.0079 **
## numhh_list      -1.9910     0.2509  -7.936 2.28e-15 ***
## hs_degree        1.2746     0.2724   4.679 2.92e-06 ***
## college_degree   2.4467     0.3978   6.151 7.95e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.84 on 12153 degrees of freedom
##   (71 observations deleted due to missingness)
## Multiple R-squared:  0.009592,   Adjusted R-squared:  0.009266
## F-statistic: 29.43 on 4 and 12153 DF,  p-value: < 2.2e-16
```

```r
model_4 <- lm(count_visit_dr ~ treated + numhh_list + NO_hs_degree + hs_degree +
↪  college_degree, data=df, na.action = na.omit)
summary(model_4)
##
## Call:
## lm(formula = count_visit_dr ~ treated + numhh_list + NO_hs_degree +
##     hs_degree + college_degree, data = df, na.action = na.omit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -8.014  -5.568  -3.568   0.158 138.432
##
## Coefficients: (1 not defined because of singularities)
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      9.4309     0.4505  20.934  < 2e-16 ***
## treated          0.5745     0.2163   2.657 0.007905 **
## numhh_list      -1.9910     0.2509  -7.936 2.28e-15 ***
## NO_hs_degree    -2.4467     0.3978  -6.151 7.95e-10 ***
## hs_degree       -1.1721     0.3438  -3.409 0.000653 ***
## college_degree      NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.84 on 12153 degrees of freedom
##   (71 observations deleted due to missingness)
## Multiple R-squared:  0.009592,   Adjusted R-squared:  0.009266
## F-statistic: 29.43 on 4 and 12153 DF,  p-value: < 2.2e-16
```

Q4. Question

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`?

Q4. Answer

The estimated treatment effect in this model with all controls is 0.585611, which is slightly lower than the estimated treatment effect of the model that only includes `numhh_list` (difference of 0.007889). We have a similar history with the standard error: when comparing both both models we get a difference of only 0.003. Finally, the R^2 of this model is 0.03257, which is higher significantly higher than the R^2 of 0.006276 in the model with only `numhh_list`, indicating that including all baseline characteristics as controls improves the fit of the model and explains more of the variation in the `count_visit_dr` variable.

Q4. Code

```
model_5 <- lm(count_visit_dr ~ treated + numhh_list + female + age + race_white
↪  + hs_degree + college_degree + health_baseline, data=df, na.action = na.omit)
summary(model_5)
##
## Call:
## lm(formula = count_visit_dr ~ treated + numhh_list + female +
##     age + race_white + hs_degree + college_degree + health_baseline,
##     data = df, na.action = na.omit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.226  -5.147  -3.178   0.368 139.813
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.408752   0.597213   5.708 1.17e-08 ***
## treated         0.585611   0.213681   2.741  0.00614 **
## numhh_list     -1.611988   0.249436  -6.463 1.07e-10 ***
## female          2.715487   0.214414  12.665  < 2e-16 ***
## age             0.009106   0.009785   0.931  0.35211
```

```
## race_white       1.233884   0.237740   5.190 2.14e-07 ***
## hs_degree        0.919203   0.276283   3.327  0.00088 ***
## college_degree   2.000781   0.400539   4.995 5.96e-07 ***
## health_baseline  2.319483   0.256797   9.032  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.68 on 12116 degrees of freedom
##   (104 observations deleted due to missingness)
## Multiple R-squared:  0.03257,    Adjusted R-squared:  0.03193
## F-statistic: 50.99 on 8 and 12116 DF,  p-value: < 2.2e-16
```

Q5. Question

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.

Q5. Answer

For unbiasenes, no. It is sufficient to just control for `numhh_list`. Since this is an RCT, treatment is randomly assigned and therefore uncorrelated with all other baseline characteristics, so there is no omitted variable bias from excluding them.

Q6. Question

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

Q6. Answer

Yes, including additional baseline characteristics can help improve precision. We saw in Q4 that R^2 increased from 0.6% to 3.3%, indicating these controls do explain some variation in doctor visits. However, in practice, the standard error only decreased slightly (from 0.2166 to 0.2137), so the precision gains for this case are minimal.

Q7. Question

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.

Q7.1.

The null hypothesis is for this question is:

$$H_0 : \beta_{treated} = \beta_{hs\_degree}$$

For the alternative hypothesis:

$$H_A : \beta_{treated} \neq \beta_{hs\_degree}$$

.

Q7.2.