Problem set 3

Time to solve some problems!

This problem set mainly builds on this notebook: In-class exercises from lecture 3

1 In-class notebook exercises (50 points)

1.1 Replicate and simulate a real study

1. Print out the full set of regression results from the tops of column 3 and column 4 in the results table. You can just use the ‘summary()‘ method on the regression results to print these. These are the two models already fitted in the notebook.

2. Which one has more control variables? Which one has a higher R-squared? R-squared is the proportion of variance in Y explained by the model. Which one has a smaller standard error on the estimated ATE?

3. Let’s do a simulation to better understand the results. You can start with the data-generating process in the section “Commentary on the standard errors and covariates" of the notebook. In the section "Modeling the simulated data" I printed out two different models fit to the data (Model 1 and Model 2). How does the regression estimate of the treatment effect in the simulated data compare to the estimate for the effect of ‘rem_any‘ in the real experiment results? Look at the estimated coefficient and the standard error. Did we do a reasonably good job of simulating data that resembles the real data?

4. Now we will simulate repeated samples from the data-generating process in the notebook. Set ‘beta_hs‘ = 0.35 for your simulations. Simulate B=1000 data samples and fit Model 1 to all 1000 samples. *See the ‘compare_lpm_prop_test‘ function in the notebook to get an idea of how you can package the results in a convenient way.

5. Across the simulations, what is the standard deviation of the estimated treatment effects for each of the two models? Make a histogram of the estimated treatment effects. (Hint: Use ‘.params['d']‘ on your regression results.)

6. How does your answer in 4 compare to the standard error in your regression output from question 3? *The standard errors in the regression output are estimates of the standard deviation of the sampling distribution of the coefficient estimate!*

7. Across the simulations, what percentage of the confidence intervals include the true ATE in each model? Reference the lecture slides on about the “Confidence interval machine."

Problem set 3

1.2 Shoe technology experiment

1. *Simple difference of means vs. blocking approach:* The baseline "inter-person variability" is set to 5. What are the sampling standard deviations of the estimated ATE (aka standard error) for the non-blocking and blocking design?

2. Blocking reduces the impact of heterogeneity in the data. What is your guess for how the blocking and non-blocking designs will compare when we increase the "inter-person variability"?

3. Make a plot of the standard errors for several different levels of inter-person variability: 0, 1, 5, 10, 20. It can just be a scatter plot or line plot.

4. In terms of the people in our model, generally what does it mean to have inter-person variability = 0? What does it mean when we increase it to a larger number like 20?

5. *Covariate approach:* Now imagine that each person's wear rate is a known function of their weight, strength, and steps per day and we have it (the wear rate) in our data (it is already there!). (Alternative story: We know each person's wear rate by measuring their past shoes.) Use a regression to estimate the ATE and include the wear rate as a covariate in your model. (Hint: Use the function regression_ate.) Simulate the distribution of ATE for a blocked and non-blocked design using the regression estimator. Use "inter-person variability" of 20. How do the standard errors compare now?

6. *Change score approach:* Repeat the previous analysis, but instead use a change score approach. The change score should be the actual wear rate in the experiment minus the historical wear rate of the person. Estimate the average treatment effect on the *change score*, that is, calculate the difference in the average change score between the treatment and control groups. How does the point estimate compare to the other approaches? How do the standard errors compare to the other approaches?

7. *"Fixed-effects", another covariate approach:* That's great but what if we do not have data about each person's wear rate? We can try another strategy. Let's have each person run through 5 pairs of shoes during the experiment. This is easy to change: You just need to add replicates to 'df = pd.concat([df_left, df_right], axis=0)'. Then cut the sample to just N=20 people so we use the same number of shoes in total. Run a simulation with inter-person variability of 20 again. Estimate the ATE using a regression with a dummy variable for each person (use C(i) in your formula). What is the standard error of the ATE estimate now? Practically, what is the tradeoff of this strategy vs. covariate strategy vs. the blocked strategy?

Problem set 3

2 Regression (30 points)

This will build on Part 1 of the notebook from class. Go back to the simulation that you ran in the notebook. We are going to modify it.

1. Imagine that the data from our simulation is a real data set about real people. What does the ATE mean? What does the coefficient on the variable `grad_high_school` mean? Note that `grad_high_school` $= 1$ if the person has completed high school and 0 if they did not.

2. Simulate B=1000 samples with ATE=0.018 and beta_hs=0.35 like you did before. Fit Model 1 (no control variables) and Model 2 (`grad_high_school` as control variable). Use the simulation results to estimate the statistical power of each model for detecting the ATE (vs. a null hypothesis that the ATE=0). (Hint: Use the p-values from the regression results. We will declare a result to be statistically significant if $p < 0.05$.) Print the power of each model. How does including the control variable affect your experiment's statistical power?

3. Repeat the above problem but set beta_hs=0.10. Show the results again.

4. In term of people in our data, what is the difference between beta_hs=0.10 and beta_hs=0.35?

5. Interpret the previous questions together. When does including a control variable increase the power of our experiment? In general, on what types of control variables should we try to collect data?

3 Shoe tech experiment redux (20 points)

This will build on the blocking and shoe tech experiment simulation.

1. Re-run (or re-use) your simulation results for a non-blocking experiment and inter-person variability = 20 and $N = 100$. What is the statistical power of the blocking design and non-blocking design (that is, their power to detect the true ATE in the simulation)?

2. Give your best explanation for why blocking can increase power.

3. Give your best explanation for why regression with covariates can increase power.

4. Give your best explanation for why an analysis with change scores can increase power.

5. For each of the three techniques, give one advantage and one disadvantage.

Bonus (10 points)

1. In the shoe tech experiment, use your simulations to compare the power of blocking vs. (no-blocking+regression using each person's wear-rate rate). Try running a larger number of simulations, say B=5000. Compare an experiment with 5 persons to an experiment with 100 persons.