ARE 2013, UC Berkeley, Fall 2023
Kirill Borusyak

Problem Set 1: Selection on Observables, and the Effects of Maternal Smoking on Infant Health

Due Date: see bCourses. Submitted by email

Instructions:

- You must work in teams of 3 to 5 students. Only one member of the group should submit the answers. You can ask other teams (or other people) for advice; however, you should not copy the code or answers from anyone.
- Late problem sets are not accepted. However, the lowest score among all problem sets will be dropped, no questions asked.
- The solutions must be typed on a computer and submitted in the PDF format. My personal preference is a LyX document stored on a shared disk (e.g. in Dropbox) but Overleaf and other packages are also good for coauthored work.
- Write your answers in full English sentences, referencing the relevant exhibits (i.e., figures and tables). Where it may not be obvious, include the formulas you are applying. Make the exhibits reasonably tidy, although publication quality is not expected.[1] Make sure it is clear either from the exhibit notes or from your answers (or both) what is shown in each row and column (including, e.g., how standard errors are computed). Parts of the same exhibit can be used to answer multiple questions (e.g. with different specifications shown in different table columns or figure panels).
- Avoid indicating statistical significance by stars. Report t-statistics only if you are actually testing hypotheses (e.g., that two coefficients are equal to each other); otherwise, report standard errors of your estimates.[2]
- Each empirical question must be solved using two different statistical programs. Stata, R, and Python are natural choices but it's up to you. Keep in mind that the availability of packages varies across programs. You can use any available commands, except when

[1] For tables, your code may produce the full table or individual coefficients separately that you put together into a table manually — your choice.

[2] Avoiding the stars is becoming a standard policy at leading economic journals. Standard errors and confidence intervals are generally more informative and force the authors and readers to think about economic magnitudes of the effects.

explicitly asked to avoid "canned" commands. If the results are identical (or nearly identical) in the two languages, report only one of them. If you get substantially different answers in the two languages, report both and comment on the issue.

- Include your codes and log-files (in both languages) as appendices, clearly indicating which parts of them correspond to which subquestion(s). I *will* review your code, so make it clear; writing clear code is an important skill for your future work. Try to follow good practices for writing code, such as avoiding repetition (but this will not affect your grade as long as I understand the code).

- For many of the questions, there is no single correct answer. Grading will focus on whether your answer is rigorous, correctly implemented, and well explained — regardless of whether it replicates the numerical answers I got.

- Sometimes you will have to use methods that are not entirely rigorous: e.g., standard errors that are not consistent because correct ones have not been derived in the literature or are too difficult to implement. It's fine but be explicit and show your awareness of the issue.

Question:

This problem set is based on the paper by Almond, Chay, and Lee (2005) and draws on the earlier problem set by Michael Anderson. The goal of this assignment is to examine the causal effect of maternal smoking during pregnancy on infant birthweight. The data for the problem set is an extract of all births from the 1993 National Natality Detail Files for Pennsylvania. Each observation represents an infant-mother match. The data in Stata and CSV formats can be downloaded from the bCourses website. The TXT file gives variable names and labels for those of you who do not use Stata. There should be 43 variables in the data and, after you are finished with the cleaning steps described below, 114,610 observations. The PDF codebook for the data is also posted and will help you understand the relevant variables.

The data here are "real" and quite imperfect, which will help to simulate the unpleasantness of real world data work. Unlike the real world where you will confront this bleak situation largely alone, I will provide you with some hints for working your way through the raw data.

1. The first order of business is to clean the data.

   (a) Fix missing values: In the the data set several variables take on a special value (e.g., 9999) if missing. Tabulate the variables and check the codebook for missing value codes; then replace special values with missing. For about 2/3 of the variables we have already done this; the remaining variables that need to be checked by you are the last ones in the list, from *'cardiac'* to *'wgain'*. (No verbal answer is required here.)

   (b) Recode all indicator variables to take values 0 and 1. A few variables in the dataset can be viewed as unordered categorical variables (with more than two possible values): *stresfip*, *ormoth*, *orfath*, *mrace3*, *birmon*, *weekday*.[3] Recode *mrace3* as a categorical variable. Coarsen *ormoth* and *orfath* into indicator variables. For simplicity, drop *stresfip*, *birmon*, and *weekday*. How would you treat *stresfip*, *birmon*, and *weekday* in the analysis if you kept them?

   (c) Produce an analysis data set that drops any observation with missing values and verify it has 114,610 observations. Do the data appear to be missing completely at random?

   (d) Produce a summary table describing some of the key variables in the final analysis data set. (A useful "Table 1" is one that describes the overall averages of the key variables, and then describes the subsets of people who do and do not receive the treatment, when the treatment is binary.)

2. Our goal will be to analyze the causal effects of maternal smoking during pregnancy on infant birth weight.

   (a) Compute the mean difference in birthweight in grams by smoking status. Is this difference likely to be causal? Provide some evidence for or against.

   (b) You decided to use the "selection on observables" identification strategy which involves assuming that maternal smoking during pregnancy is randomly assigned conditional on some observable determinants of infant birth weight. Now it's time to choose those covariates using informed judgement. Classify the variables in the dataset into different types depending on how they are potentially related to the treatment and the outcome; justify your choices and be explicit about the

[3]Some other variables are ordered categorical variables. We'll treat them as continuous, although that is not the only reasonable choice.

assumptions you are making.[4] Use your classification to decide on the list of covariates you'll keep for covariate adjustment for the rest of the problem set.

3. We will now investigate different methods based for covariate adjustment, starting from regression.

   (a) Use a basic, uninteracted linear regression model to estimate the impact of smoking and report your estimates. Under what circumstances does it identify the average treatment effect (ATE)? (Assemble all of your estimates and standard errors from this and later questions into a table or several tables that would make it easy to compare the methods.)

   (b) Is the estimate in the previous question sensitive to dropping controls one at a time? What do you learn from this exercise?

   (c) For this part only, extend the OLS specification from question 3(a) to control for the covariates using a more flexible functional form. Describe the specification you picked. What are the potential benefits and drawbacks of this approach?

   (d) For this part only, add to the specification of question 3(a) some "bad controls." Check if your estimate changes and discuss the direction of the change.

   (e) Produce the Oaxaca-Blinder estimator for the ATE and ATT. Describe the exact steps you have used. Does your answer differ substantially from the one in 3(a)? Discuss.

4. Next on the list is the propensity score approach.

   (a) Estimate the propensity score using a logit specification without nonlinear terms and interactions. Discuss which covariates appear most predictive of maternal smoking and whether this matches your expectation.

   (b) How good is the overlap of propensity scores between the treated and untreated groups?

   (c) Assess whether this logit specification has been sufficient to balance the covariates.

   (d) Estimate the ATE and ATT via propensity score blocking **or** matching (one

[4] You might find it helpful to review the different types of controls shown on the DAGs in lecture slides B1 but you don't need to limit yourself to those types or use that classification. You are not expected to read medical literature to make a better judgement, as you would in a real study. Don't go into excessive detail, as there are many variables in this dataset.

method is enough).[5] Explain why you picked this one.[6]

(e) Estimate the ATE and ATT via propensity score reweighting. Include the formulas you used.

5. Finally, try doubly-robust methods:

   (a) Estimate the ATE and ATT using one the "mixed methods": regression adjustment combined with propensity score blocking, matching, or reweighting.

   (b) Answer **one** of two questions:

   - Assuming constant effects,[7] estimate the causal effect using post-double-selection LASSO. Use flexible specifications with some polynomials and interaction terms. (Defining these nonlinear terms — "feature engineering" — for the use in LASSO may require some thinking.) How many covariates did you start with? How many were selected by LASSO in the outcome regression? How many in the propensity score regression? How many overlapped? In this application, would you get a very different answer if you didn't include the covariates chosen in the propensity score regression?
   - Estimate the ATE using double/debiased machine learning with the machine learning algorithm of your choice (other than OLS, logit/probit, and lasso). Describe the estimation steps, even if you are using a canned command. (Partial credit if you assume constant effects.)

6. Compare the estimates and standard errors between regression adjustment methods, propensity score methods, and doubly robust methods you have used. Concisely and coherently summarize your results above providing some intuition. Write it like you would the conclusion of a paper. In this summary, describe whether you think your best estimate of the effects of smoking is credibly identified; state why or why not.

[5] This and some later questions ask you to choose one method. The method you pick need not be the same in the two programming languages you are using: e.g., you can try pscore blocking in Stata and matching in R if you want. (And if you want to go beyond the requirement and try more than one method in each language, go ahead by all means.)

[6] Please give an honest answer: e.g., if it is more about implementation convenience than statistical properties, that's fine.

[7] If interested in estimating ATE and ATT with heterogeneous effects using post-double-selection LASSO, see Section 5 of Belloni, Chernozhukov, and Hansen (2013).