Problem Set 2 PAM 6090, Fall 2023
Due Wednesday, October 11th, 11:55pm (Turn in via canvas)

General instructions:

- You can choose your own teams of size 1-4 (not more).
- Please present your answers in a clear, concise fashion. Please submit a single main PDF document with your answers. In your solution packet, include relevant Stata output (e.g., key regression output, key graphs, etc.) and well-annotated Stata .do commands. (Do NOT include pages and pages of "undigested" Stata log files in the main problem set answers.) Make clear reference to regression output and figures in your written answers.
- Please upload the PDF file to canvas. Only one upload per team is needed (but clearly indicate the members of your team in your writeup). In addition to the main answers, please also upload .do and .log files (one .do file; one .log file) as separate files.

Problem 1 – Randomly Assigned X?

This question is about a research paper where multivariate OLS was combined with a compelling research design: Ebonya Washington's 2008 paper ``Female Socialization: How Daughters Affect Their Legislator Father's Voting Behavior." https://www.aeaweb.org/articles?id=10.1257/aer.98.1.311

The paper asks whether having daughters, conditional on overall family size, causes people in Congress to vote more liberally on bills related to women's issues (specifically, more in alignment with the National Organization of Women's preferred votes).

1.1 Download the dataset wash_basic.dta, and replicate the regression reported in the first column of Table 2. Interpret the coefficient on number of girls. (You should be able to almost perfectly match the coefficients in the table, up to rounding-error).

1.2 Succinctly describe the author's identification strategy, or why she believes that the regression results in this Table reflect the true causal impact of daughters. What assumptions are necessary for the coefficients above to consistently estimate the true causal effect?

1.3 Given the stated research design of the paper, what is the most parsimonious specification (the regression with fewest control variables that one could estimate) that should still yield the true causal effect of daughters on voting behavior? Run that regression and comment on how your results affect your evaluation of the validity of the identifying assumptions.

1.4 Conduct some extra analyses to see whether the identifying assumptions made by the paper seem valid.

1.5 If you were a referee (or the editor) for the AER, would you recommend the paper be published or not, or what additional analyses would you request of the author to convince you of the paper's merit?

1.6 What do you think could be going on that explains the difference between results in part 1 and part 3 of this question (and which incorporates evidence you have learned in part 4)?

Problem 2 – Bootstrapping

In a regression setting, does bootstrapping give us the usual OLS-iid estimate of variance, or the Huber-White estimate? We discussed "theory" in class. Let's find out "in practice"!

a)

Construct one set of fake data, using the d.g.p. of problem set 1, problem (1a). That is: x is distributed iid N(0,1). eps_i (the i–th epsilon) is distributed N(0,(1+(x_i)^2)). That is, the variance of epsilon is 1+x^2 for that observation. Then assume that the LHS variable Y is generated according to: y = 2 + 1 * x + epsilon . Make the number of observations (N) in the dataset be 200.

Estimate a regression using the usual estimate of variance, as well as the robust estimate of variance. Then, build a bootstrap to get a bootstrap estimate of the variance of the slope parameter. Which is this closer to?

Choose **at least** 100 bootstrap replications, and experiment with more.

b) Embed your exercise in part (a) within a Monte Carlo study to compare the inference properties of: (i) naïve variance estimate; (ii) the "robust" variance estimate; and (iii) a bootstrap estimate of the variance. (You've already done (i) and (ii) in Problem set 1, possibly with a different true form of heteroscedasticity).

Problem 3 –Nonparametric regression and the bootstrap

For this problem, use the stata dataset "kenya.dta" that you used in Problem set 1. You will only need the "cohort 1" data.

3a) Perform the Fan locally-weighted non-parametric regression on the specifications listed below. (I think the Stata command lpoly, for "local polynomial", is probably a good function to use here. If so, make sure to use the right option to set polynomial order to 1.) Use your own judgment (and eyeballs) to pick a good bandwidth and kernel.

Estimate the relationship on the non-extreme values of the 2000 distribution, only considering the 2000 test score interval of [-1.3, 1.6], for simplicity.

(i) Treatment girls, (nonparametrically) regress the 2001 test (dependent variable) on the 2000 test.

(ii) Comparison girls, (nonparametrically) regress the 2001 test (dependent variable) on the 2000 test.

(iii) Try to replicate the "treatment effect" line from Figure 3, Panel A of the published article. (for now, don't worry about the confidence interval.) That is, take the difference between these two non-parametric plots, and plot them out.

(iv) Try re-estimating (iii) for some smaller and larger bandwidths. Show at least one result that you think is "just barely too small a bandwidth" and one result that you think is "just barely too large a bandwidth". Also show results that you think are most informative for answering the question "how does treatment impact depend on baseline test score"?

(v) How similar or different are the impressions/punchlines for your preferred estimate, compared to your best "replication effort" estimate?

3b) Now let's build the 95% confidence interval, using the bootstrap. This is going to be a bit harder than step 3a. Let's start with your best "replication" bandwidth and kernel specification. To get a confidence interval on this graph, bootstrap the following procedure:

one: draw a sample from the pooled (treatment and control) distribution
two: estimate the fan regression for the treatment group
three: estimate the fan regression for the control group
four: take the difference, and save this.

You will want to build code that runs many bootstrap replications over the four steps above. You will need to make sure that steps two and three are estimated for the same set of points for each bootstrap draw. This will require you to give "lpoly" a common set of points to estimate over, rather than dataset-dependent set of points. (Look up the "at( )" option for lpoly.)

When you save your results in each bootstrap replication, you can save the following variables: (1) the value of points you are estimating on, (2) the estimate for treatment, (3) the estimate for control, (4) the estimate for the difference. (items 2 and 3 are only needed if you also want to compute confidence intervals for the treated and control separately, beyond the difference.)

Then after the bootstrap is done, you can compute the standard error (which will be the bootstrap standard deviation) at each point. Then you can construct the bootstrap confidence interval as the main estimate for that point, +- 1.96 * the estimated bootstrap standard error.

Is the difference between treatment and comparison students significantly different than zero (at over 95 percent confidence) anywhere in the 2000 test score distribution?

3c) What are the implications of these results for our understanding of the relationship between merit awards, child effort, and learning in rural Kenya? Are most test score

gains at the "top" of the distribution, in the "middle", or at the "bottom"? Is there any conclusive evidence of negative externalities for low-achieving students?

4. Prediction Exercise

Let's compare various prediction models used to predict child population counts, at the county-year-age level. Use the data sets population_train.dta to choose and estimate your models and for "in-sample goodness of fit"; and use population_test.dta for "out-of-sample goodness of fit". (These are in the datasets module.)

Start with a "straw man" model that just uses the global sample average as the predicted value. Then compare at least 4 different models (more are okay, up to 8 models total), that will let you examine the following questions:

- You may want to experiment with creating new variables that are transformations of existing variables. (nonlinear transformations, polynomials, interactions, etc.)
- Parsimonious versus "kitchen sink" versus "extreme kitchen sink" (etc.) covariate sets.
    - Remember that you can use interactions, dummy variables for individual values, etc.
- "Basic OLS" vs. LASSO models vs. Post-LASSO models
- (Optional: Random Forest or other "fancier" Machine Learning models)

To assess goodness of fit, show for both in-sample and out-of-sample goodness-of-fit. Compare both the RMSE and the "Median Absolute Percentage Error" (MAPE).

a) Fill in a table like the following:

	RMSE		MAPE	
Model	In-sample	Out-of-sample	In-sample	Out-of-sample
Straw-man model: y-hat = y-bar				
(describe first model)				
(describe second model)				
Etc.				

b) What is your preferred model, and why? If you had to use this to make predictions for an as-yet-unseen dataset, which would you want to use?

c) How do the in-sample and out-of-sample predictions compare? Does this vary by model? What explains the differences (or lack of differences)?

d) Based on your results, what is the strongest case to be made in favor of using Machine Learning as a prediction tool? What is the strongest case to be made in favor of sticking with “traditional” models?