ECMA 31330 - Problem Set 3

Bruno Aravena M.

Due date: January 30th 3.30pm

Instructions: For every assignment, you will be asked to submit a document where you write up your answers and insert the figures/ tables requested for the assignment. You will also be required to submit the code where you code your simulations and produce your answers, figures, tables, etc. You can write your code in any programming language you want, but I will be submitting solutions and presenting examples in TA sessions in R.

This means you will submit a .pdf where you answer the questions and include your final figures, and a script of code (e.g. .R) where I can see your work.

1 Growth regressions: Lasso and CV

For this question we will be working again with the data from Sala-i-Martin's paper I Just Ran Two Million Regressions" (AER, May 1997) In particular, you might want to look at the end of the paper where the author describes the data. I also found this resource that might complement the data description.

DATA: You can find the data here

This time, we will deal with the issues in the data in a more guided way to have answers more comparable across problem sets. There is particularly especial or correct about the procedure I propose.

1. To deal with missing data, filter out the rows that do not have the outcome variable (X1). Of the remaining dataset filter out columns where there are more than 10 missing entries. You should have now a dataframe of 117 rows and 41 columns which include 1 outcome variable and 36 potential regressors.

2. Impute the median of each column for each remaining missing value.

3. Considering that we will be working with LASSO, what data processing should you **always** remember to do with your covariates before running a LASSO (or any penalized) regression ? Why ?

4. Implement the step discussed above. Extend it to your outcome variable too for interpretation. Now: what is the interpretation of a $\beta_j = 0.5$ in this case ?

   Let us now focus our attention to cross-validation, folding and out of sample error.

5. Let us begin with a "particular" fold of the data. Split your data where the bottom 80% in the outcome variable is the "training" data set and the top 20% becomes the "test" data set. Run OLS with 3 covariates (X16, X46, X53) in the training data set and compute the out of sample prediction error in the test data set.

6. Run LASSO in this particular training set and choose a model via CV (minimizing out of sample prediction error in the test data set). Report the out of sample prediction error. How many (if any) covariates are chosen?

7. Suppose now we have a more traditional (random) fold of the data which divides the data into training and test again and that we run LASSO in our new (random) training data set and we choose a model based on minimizing out of sample error in the new (random) test data set (Cross-validation). How do you expect that model to perform in the previous (rich countries) test data set ?

8. Create a random split of your data (80/20) as described above. Run LASSO and choose a model through CV. Verify your supposition and report the performance of this model on the previous (rich countries) test data set.

9. Run a 5-fold CV LASSO, where the test set iterates over the whole dataset. Feel free to use a package to streamline this process. Report the performance of this model on the previous (rich countries) test data set.

10. You have produced 4 models and reported their performance for a particular subset. How does prediction error compare across models? Does this make sense? Comment why or why not.

2 Ideas of datasets for final projects

I thought it would be good for everyone to get started thinking about the final project. To that effect, we will create a repository where all students will suggest papers with data (or just datasets even if they do not have a paper associated) that people could consider working with.

As a starting point: think of papers you have encountered and enjoyed reading through your studies and/or datasets that you know of which could be candidates for the final project. On top of that consider the following three crucial conditions:

- Because this is a course on **Economics** & Machine Learning: Paper/data must be of relative economic relevance. There needs to be a case that the paper or the data at least is economics-adjacent.
- Because this is a course on Economics & **Machine Learning**: Make sure the dataset contains at least 15-20 covariates. The more the merrier.
- Because the idea is that you actually work with this data over the next few weeks, **the data must be available**. If you are proposing the replication package of a paper, do verify that the data is indeed contained in the replication package (if not, try another dataset or try reaching out to the authors)

1. **Find at least 2 (4 if you are submitting this problem set in pairs) that have fulfill all the conditions above and complete a Google Form entry for each. After that, list in your .pdf the titles of the papers/datasets associated with this submission**

   The form will ask you about: unit of analysis in the data set (person, country, households, pixel, etc.); number of observations; main outcome variable(s); main independent variable(s); at least 5 covariates that you find interesting; number of covariates; abstract of the paper; (if not from an Economics journal) why do you think this is connected to Economics?

   If you have more papers in mind, I would encourage you to include a few more. Think that you have a chance to inspire your fellow classmates and that you will be receiving a compiled list with over 60 papers which can become starting points for your project.

2. Choose 1 paper. Use your Lasso package to create a CV plot as the one on the left panel of Figure 6.6 (Chapter 6.2). Which 5 variables "survive" the longest?

3. Create 2 - 3 plots to showcase some key relationships (pairs of variables) in the data. No need to get complicated: Properly labeled scatterplot for continuous variables might suffice in some cases. In others adding a third dimension (e.g. through color) or a conditional expectation might be very informative.

4. Comment on the plots. Compared to your prior: are these variables related in the way you expected them? Is the relationship linear?