Problem Set 2

Econ 4676: Big Data and Machine Learning for Applied Economics

Due Date: October 15 at 11:00 am

1 Theory Exercises

1. Suppose you have the following spatial model $y = \rho Wy + X\beta + WX\theta + \epsilon$ with $|\rho| < 1$ this is sometimes known as the Spatial Durbin Model
   (a) First consider the following scenario $\beta = \theta = 0$.
      i. Write the Likelihood function. Can you find a closed form for the parameter estimators? Don't forget to be specific on the assumptions you make.
      ii. Suppose instead you use MCO, would you obtain the same estimates?
   (b) Now consider that $\rho = 0$, and let's proceed as before:
      i. Write the Likelihood function. Can you find a closed form for the parameter estimators? Don't forget to be specific on the assumptions you make.
      ii. Suppose instead you use MCO, would you obtain the same estimates?
2. Consider the regression model $y = X\beta + \epsilon$ with $\epsilon \sim N(0, \sigma^2 I)$ furthermore assume that β has a normal prior, i.e. $\beta \sim N(0, \tau^2 I)$.
   (a) Find the posterior distribution.
   (b) Compare it with the ridge formula we saw in class.
   (c) What is the relationship between λ in the ridge model and σ^2 and τ^2?
3. Centered Ridge. Suppose that $\bar{x} = 0$, i.e. the data has been centered. Show that the parameters that minimize $R(\beta, \beta_0) = (y - X\beta - \beta_0\iota)'(y - X\beta - \beta_0\iota) + \lambda\beta'\beta$ are $\beta_0 = \bar{y}$ and $\beta = (X'X + \lambda I)^{-1}X'y$
4. Suppose that we have the following regression model $y = X\beta + \epsilon$, and decide to do the following: Augment the centered matrix X with p additional rows with $\sqrt{\lambda}$, and augment y with zeros. Show that this procedures renders the ridge regression estimates, is there a link to the leverage statistic?
5. Reducing elastic net to lasso. Suppose that you have the following functions $EL(\beta) = (y - X\beta)^2 + \lambda_2\beta^2 + \lambda_1|\beta|$ and $L(\beta) = (\tilde{y} - \tilde{X}\beta)^2 + c\lambda_1|\beta|$ where $c = (1+\lambda_2)^{\frac{-1}{2}}$ show that these two problems are equivalent when $\tilde{y}$ and $\tilde{X}$ are the augmented data versions of the previous exercise.

2 Empirical Problems

The main objective of these sections is to apply the concepts we learned using "real" world data. With these, I also expect that you sharpen your data collection and wrangling skills. Finally, you should pay attention to your writing.

I encourage you to turn each of the following two parts of the problem set in a way that resembles paper. As such, I expect graphs, tables, and writing to be as neat as possible. You can write it in Spanish or English, either language is fine. For students in the Ph.D., it would be a good practice to do it in English.

These parts also involve a lot of coding. Please attach your code with your responses or point to your repo. In coding, like in writing, a good coding style is critical for readable code. I encourage you to follow the tidyverse style guide.

2.1 Getting to know Evanston, IL

This part of the problem set involves a series of spatial data sets on the City of Evanston, IL. The `data` folder contains the shapefiles needed. The first objective is to introduce you to some of the mapping facilities in `R`. The data contains parcels in Evanston from the county assessors' office.[1]. The second objective is to use the tools studied in class to model and predict assessment values using the information included in the assessment data file and combined with infrastructure data.

1. *"Mapping the field"*

   (a) Begin by creating a map that includes census area identifiers (census blocks, census tracts), major infrastructure layers (train line, roads) and Lake Michigan shore line.

   (b) Match the parcel data to the block level file and calculate average assessment values and building area to floor area at the block level.

   (c) Report these results in side by side maps using the map you created in part (a)

   (d) Discuss results

2. *"Out of sight, but not out of mind"*

   (a) Build a table with descriptive statistics for the potential variables you can use to predict assessment values prices. Make sure to include "spatial variables", i.e. distances to major infrastructures.

   (b) Generate a prediction using OLS. Evaluate your prediction using a validation set approach, i.e., split the sample into two samples: a training (70%) and a test (30%) sample. Don't forget to set a seed (in `R`, `set.seed(10101)`, where 10101 is the seed.)

[1]For more info check and variable definitions check https://tinyurl.com/y6y6bhat and https://datacatalog.cookcountyil.gov/stories/s/p2kt-hk36. Note how they use `Gitlab` for version control and ML for prediction

(c) Repeat the previous point but using K-fold cross validation. Comment on similarities/differences of using this approach.

(d) Repeat Ridge strategy to generate predictions.

(e) Model Selection:

   i. Use best subset selection to choose a model. Which is the selected "best model"? (In `R` you can use `regsubset()` from the `leaps` package)

   ii. Now use Lasso. Compare to previous results

(f) Estimate the "best" model found in previous exercises specifying a spatial structure of the data. You can use any of the spatial linear model that you find appropriate. Make sure to discuss how you defined proximity between observations.

Note 1*: Make sure you comment, compare, and discuss your results. Using figures and tables to enhance your argument are highly encouraged. Make sure you write up your results and include it in a pdf.*

Note 2*: If you find that the data set is too large for spatial models you have a couple of options. As a first option you can aggregate at the block level and estimate these models on the more aggregated spatial units. A second option, is to keep a random subsample and operate on the smaller sample. Third, use your AWS account and take advantage of the cloud.*