Problem Set 3

Econ 4676: Big Data and Machine Learning for Applied Economics

Due Date: October 15 at 1:00 pm

The repo link to create your submission is
https://classroom.github.com/g/KThOe0YT

1 Theory Exercises: MLE and Spatial Econometrics

1. *Binary Response Online Updating.* The problem is simple but yet complicated. Online updating is essential because it breaks the storage barrier and helps with computation. Assume that a new observation arrives, and instead of refitting the entire model, you want to update your binary model estimates. Show that the contribution made by the observation i to the likelihood function is

$$l(y,\beta)=\sum_{i=1}^{N}\left(y_i logF(x_i'\beta)+(1-y_i)log(1-F(x_i'\beta))\right) \tag{1}$$

is globally concave with respect to β if the function F is such that $F(-x)=1-F(x)$, and if its first derivative f, and its second derivative f' satisfy the condition:

$$f'(x)F(x)-f^2(x)<0 \tag{2}$$

for all real finite x. Show that this condition is satisfied by the logistic function $\Lambda(x)=\frac{1}{1+e^{-x}}=\frac{e^x}{1+e^x}$

2. In many sub-fields of economics, like finance, it is common that the tails of the noise distribution are much heavier than the standard Gaussian tails. One way to model this is to use a $t-distribution$. Consider the following model: $y_i=\alpha+\beta x_i+\epsilon, i=1,...,N,$, i.e., a model with a constant and a single regressor. But now $\epsilon/\sigma \sim_{iid} t_v$, and $E(X\epsilon)=0$ and v are the degrees of freedom.

   (a) Write down the log-likelihood, using the explicit formula for the density of the $t-distribution$.

   (b) Find the first derivatives of this log-likelihood with respect to the parameters α, β, σ^2, and v.

(c) Assume that σ^2 and v are known. Can you solve for the maximum likelihood estimator of α and β? Do they match the least-square estimators? If not, why and how do they differ?

(d) Using the software of your choice, write a function that takes as arguments some data (y,x) and outputs the MLE estimates from a $t - distribution$.

(e) Simulate some data that follow the model described above, and answer the following questions:

  i. How well does the MLE recover the parameters?

  ii. Does it get better as N grows?

  iii. What about when the variance of X increases?

  iv. How does it compare relative to a naive OLS estimator?

(f) Each student in the class has an account in AWS. You can access it with your `@uniandes.edu.co` user. Set up an AWS instance and install a software of your choice. Attach screenshots of the virtual machine running. Some suggestions to install:

  i. `R` with `RStudio`

  ii. `JupyterLab`

  iii. `Python`

(g) Repeat the simulation in `e` using a parallel or distributed approach in AWS

  i. Did you get the same results as above? If not, why?

  ii. How did you handle the `seed` in this context?

  iii. Was there a computational time gain? Report the differences.

  iv. Indicate who in your team ran the simulations. (I'll check AWS usage)

3. 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 Empirical Problems

The main objective of this section 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 the following section of the problem set in a way that resembles a paper. As such, I expect graphs, tables, and writing to be as neat as possible. You can write it in Spanish or English, and either language is acceptable. 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. Don't forget to upload everything to your repository and follow the template repository.

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. All the relevant data sets are in the `data` folder. The `evanston_parcel_data.csv` file contains parcel level data from Evanston. The data was retrieved from the county assessors' office.[1]. The first objective is to showcase your mapping abilities. The second objective is to use the tools studied in class to model and predict assessment values using all the provided information.

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, etc.), and Lake Michigan shoreline.
   (b) Match the parcel data to the block level file and calculate (i) average assessment values and (ii) building area to floor area at the block level.
   (c) Describe your results. In your description, include a side by side map using the map you created in part (a) that includes the information you generated in (b)
2. In Problem Set 1, we focused on obtaining the best in-sample fit. Here we shift our focus to prediction.
   (a) The data set include multiple variables that can help predict the assessment value of a property. Describe those that you will use in your predictive exercise. I leave it to you to decide which variables to include and their functional form. Don't forget to include the "spatial variables," i.e. distances to major infrastructures, that you created in (1).

[1] For more info 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

(b) We will continue exploring linear models of the form

$$y = X\beta + u \tag{3}$$

Where Y is the assessment price, and X is a matrix with the variables you chose to predict Y. Our approach will be very agnostic, the objective is to minimize the prediction error.

i. Start with a model that only includes a constant. Then estimate more complex models. These models can include more variables, interactions, transformations, etc. I expect that you estimate at least 10 models.

A. Describe and explain how and why you built these models.

B. Report, describe, and compare the prediction error.

C. Explain how you calculated the prediction error.

ii. Discuss the model with the lowest prediction error.

(c) Estimate the model with the lowest prediction error in (b) adding spatial structure to it. You can use any of the spatial models that we talked in class or any other that you deem appropriate. For example, you can run $y = WX\beta + u$.

i. Discuss how you defined proximity between observations.

ii. Did the spatial structure improve your model?

iii. Can you specify a spatial model that improves upon the "best" in part (b)?

(d) *LOOCV.* With your preferred predicted model (the one with the lowest average prediction error) perform the following exercise:

i. Write a loop for the first 1,000 observations ($i = 1$ to $i = 1,000$). The loop should do the following:

- Estimate the regression model using all but the $i - th$ observation.
- Calculate the prediction error for the $i - th$ observation, i.e. $(y_i - \hat{y}_i)$
- Calculate the average of the numbers obtained in the previous step to obtain the average mean square error. This is known as the Leave-One-Out Cross-Validation (LOOCV) statistic.

ii. Compute the leverage statistic for each observation. Show analytically and empirically that the leverage statistic can be used for the computation of the LOOCV statistic.

iii. Use the statistic derived in the previous point to calculate the LOOCV statistic.

Note: *If you find that the data set is too large for spatial models you have a couple of options. As a first option, use a random subsample and operate on the smaller sample. As a second (and recommend) option, use AWS or Azure and take advantage of the cloud.*