Problem Set 1

Econ 4676: Big Data and Machine Learning for Applied Economics

Due Date: April 24 at 11:00 am

1 Theory Exercises: Econometrics Review

1. Consider the regression model $y_i = \alpha + \beta x_i + \epsilon, i = 1, ..., N,$, a model with a constant and a single regressor. Assume that $E(\epsilon_i|x_i) = 0 \ \forall i$.

   (a) Show that $E(\epsilon_i|x_i) = 0$ implies $E(\epsilon_i) = 0$ and $E(\epsilon_i x_i) = 0$

   (b) Use the two previous implications to derive the Method of Moments estimator

   (c) Can you accommodate the terms in the previous point to put the estimator in the famous formula $\hat{\beta} = (X'X)^{-1}X'y$?

2. Prove the following properties of R^2:

   (a) The OLS estimator maximizes R^2

   (b) $0 \leq R^2 \leq 1$

   (c) For the two-variable model $Y_i = \alpha + \beta x_i + u_i$, show $r^2 = R^2$, where r is the sample correlation coefficient between Y and X.

3. Consider the linear regression $y = \beta_1 \iota + X_2 \beta_2 + u$ where ι is an n-vector of 1s, and X_2 is an $n \times (k-1)$ matrix of observations on the remaining variables. Show, using the FWL Theorem, that the OLS estimators of β_1 and β_2 can be written as

$$\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} n & \iota' X_2 \\ 0 & X_2' M_\iota X_2 \end{pmatrix}^{-1} \begin{pmatrix} \iota' y \\ X_2' M_\iota y \end{pmatrix} \tag{1}$$

   where M_ι is the matrix that takes deviation from the sample mean

4. Given the model $Y = X\beta_0 + \epsilon$ where X is $n \times k$. Let also $\hat{\beta}$ denote the OLS estimator and R_k^2 denote the R^2 (centered), where the subscript k means a model with k explanatory variables.

(a) Show that

$$R_k^2 = \sum_{k=1}^{K} \hat{\beta}_k \frac{\sum_{i=1}^{n}(X_{ik} - \bar{X}_k)Y_i}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \tag{2}$$

where $\hat{\beta}_k$ is the $k-th$ element of $\hat{\beta}$, X_{ik} is the $i-th$ element of the $k-th$ explanatory variable, Y_i is the $i-th$ element of Y, $\bar{X}_k = \sum_{i=1}^{n} X_{ik}/n$, and $\bar{Y} = \sum_i^n Y_i/n$

(b) Suppose that you delete an explanatory variable from the model (so that the model has $K-1$ explanatory variables) and obtain R_{K-1}^2, show that $R_K^2 > R_{K-1}^2$

5. Suppose you want to minimize the following function

$$f(\beta_1, \beta_2) = \frac{1}{2}(\beta_1^2 - \beta_2)^2 + \frac{1}{2}(\beta_1 - 1)^2 \tag{3}$$

(a) Compute the gradients $\frac{\partial f}{\partial \beta_1}$ and $\frac{\partial f}{\partial \beta_2}$

(b) Write the following function

i. Give initial values β_1 and β_2

ii. Until $f(\beta_1^i, \beta_2^i)$ *"does not change much do"*

- $\beta_1^{i+1} = \beta_1^i - \eta \frac{\partial f}{\partial \beta_1}$
- $\beta_2^{i+1} = \beta_2^i - \eta \frac{\partial f}{\partial \beta_2}$
- compute $|f^{i+1} - f^i|$
- if $|f^{i+1} - f^i| < tol$ stop, otherwise continue
- $i \leftarrow i + 1$

iii. here you need to define the step size η and what *"does not change much do"*.

A. Pick a "small" step and a "big" step.

B. Set a high tolerance rate (tol) and a small tolerance rate to define *"does not change much do"*.

iv. Graphically illustrate these results

6. 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 then the same model as exercise 1: $y_i = \alpha + \beta x_i + \epsilon, i = 1, ..., N,$, 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 recovers the parameters?

ii. Does it get better as N grows?

iii. What about when the variance of X increases?

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

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. Don't forget to upload everything to your repository and follow the template.

2.1 Exploring the Housing Market in Colombia

This part of the problem set involves data on housing prices in Colombia. The data was provided by https://www.properati.com.co. It contains information on listing prices as well as features of the properties on sale. The data set is called `co_properties.csv` and you can download it from here.

We will explore the data set and try to explain and predict the asking price based on OLS and the available data.

1. In this problem set we will focus only on houses and apartments that are on sale in Bogotá D.C., Cali, and Medellín. I care only about these type of operation, properties, and cities. You are welcome to use all the data set, but results should be relevant for these subgroups.

2. The data set include multiple variables that can help explain the price of a property. Build a descriptive analysis of these variables. At a minimum, you should include a descriptive statistics table for the potential variables you can use to explain prices. Note that there are many observations with missing data. I leave it to you to find a way to handle these missing data. Don't forget to discuss your decisions and the data. Take this section as an opportunity to present a compelling narrative to justify or defend your data choices. Do not present it as a "dry" list of ingredients.

3. Estimate a linear model using OLS of the form

$$Y = X\beta + u \tag{4}$$

   Where Y is the asking price and X is a matrix with the variables that you chose to explain the price. I leave to you to decide which variables to include. Please discuss your decisions and your results, including a discussion of the fit.

4. Compute the leverage statistic for each observation. Are there any "outliers", i.e. observations with high leverage driving the results.

5. One difficulty with linear models is that the interpretation of the estimated parameters is intimately connected with the units of measurement of the included variables. However, it is often convenient to present estimates of semi-elasticities housing prices, or even elasticities. This changes the functional form and the sample fit. In this part of the problem set, I want you to explore the different functional forms and their fit. You can try one by one, or you can estimate Box-Cox forms or anything you deem sensible.

6. Once you've chosen your "preferred functional" form for the equation, you can also transform your independent variables by adding polynomials and interactions. At this point, explore different transformations of your independent variables. There are two purposes here (1) explore heterogeneity in the sample (2) improve the in-sample fit. You should keep this in mind when discussing your results. One of those models should include the linear and the square term of rooms, compute the number of rooms that maximizes the expected price. Don't forget to calculate also the standard errors. You should discuss the relevance of this result, taking into account the number of rooms observed in the sample.

7. Estimate the preferred model for each of the three cities separately, compare it to the model that consolidates the three cities. Is there a way that you can recover the parameter of rooms of the consolidated sample by combining the parameters obtained in the separated samples.

8. Can you come up with a single model that reproduces the estimates of the single regression in one. Comment on the standard errors.

9. Estimate your preferred model using the `QR` decomposition. Compare these results to the traditional `out of the box` estimation methods results (for example, in `R` would be comparing it to `lm`, `Stata` would be to `reg`.

10. How well does your preferred model at predicting asking prices in Barranquilla? Comment in terms of prediction error. If it performs well, argument why, and if not, explain.

11. Are there some other variables missing, e.g., amenities, that could potentially help? How could you obtain these variables and add them to this data set? Here I expect a thoughtful discussion, but if you want to add data you are welcomed and I'll reward you with a bonus in your grade.

3 Pedes in terra ad sidera visus

Each student in the class has an account in AWS educate that you can access with your `@uniandes.edu.co` account.

1. Set up an `EC2` instance

2. Install the software of your choice, could be any of these

   (a) `R` with `RStudio`

   (b) `JupyterLab`

   (c) `Python`

   Attach screen shots of the virtual machine running.