Problem Set 1

Econ 4676: Big Data and Machine Learning for Applied Economics

Due Date: August 24 at 1:00 pm

The repo link to create your submission is
https://classroom.github.com/g/2zlAqMNF

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 $E(\epsilon_i x_i) = 0$

   Solución

   Para esta parte, primero probaremos la ley de esperanzas iteradas (para una variable continua, ya que el caso de una variable discreta es similar). Esta ley nos dice que, dadas dos variables aleatorias Z y W, con función de densidad conjunta $f(Z, W)$, se cumple que

   $$E(Z) = E_W(E(Z|W))$$

   Para ver el porqué, primero notemos que

   $$E_W(E(Z|W)) = \int_{-\infty}^{\infty} E(Z|W)f(W)dW = \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} Zf(Z|W)dZ \right) f(W)dW$$

   donde $f(W) = \int_{-\infty}^{\infty} f(Z, W)dZ$ y $f(Z|W) = \frac{f(Z,W)}{f(W)}$. Por lo tanto, lo anterior puede escribirse así

   $$\int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} Zf(Z|W)f(W)dZ \right) dW = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} Zf(Z, W)dZdW$$

   y, a su vez, tenemos que esta última expresión puede escribirse así

   $$\int_{-\infty}^{\infty} Z \int_{-\infty}^{\infty} f(Z, W)dZdW = \int_{-\infty}^{\infty} Zf(Z)dZ = E(Z)$$

Ahora bien, por la ley de esperanzas iteradas tenemos que

$$E(\epsilon_i) = E_{x_i}(E(\epsilon_i|x_i)) = E_{x_i}(0) = 0$$

Por último, para probar el segundo resultado (que $E(\epsilon_i x_i) = 0$) también podemos usar la ley de esperanzas iteradas:

$$E(\epsilon_i x_i) = E_{x_i}(E(\epsilon_i x_i|x_i)) = E_{x_i}(x_i E(\epsilon_i|x_i)) = E_{x_i}(x_i 0) = 0$$

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

Solución

Del anterior punto tenemos los siguientes momentos poblacionales: $E(\epsilon_i) = 0$ y $E(\epsilon_i x_i) = 0$. Los estimadores de α y β, $\hat{\alpha}$ y $\hat{\beta}$, respectivamente, estarán dados por los análogos muestrales de los momentos poblacionales

$$\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i) = 0$$

$$\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i)x_i = 0$$

(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 $Y = X\beta + \epsilon$ 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

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 a 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.

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

1. The data set include multiple variables that can help explain the price of a property. Choose the most relevant and perform a descriptive analysis of these variables. At a minimum, you should include a descriptive statistics table. 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.

2. 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 you chose to explain the price. I leave it to you to decide which variables to include. Discuss your decisions and results, including a discussion of the fit.

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

4. 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. Changes in the functional form alter the interpretation and the sample fit. In this part of the problem set, you should explore different functional forms and their fit. You can try one by one, or you can estimate Box-Cox forms or anything you deem sensible.

5. 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 *number 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.

6. 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`.)

7. Estimate the preferred model for each of the three cities separately,

   (a) Compare it to the model that consolidates the three cities. Show this in a 4 column table. Discuss

   (b) Is there a way to recover the linear parameter for *number of rooms* of the consolidated sample by combining the parameters obtained in the separated samples for each city?

   (c) Can you come up with a single equation model that shows the coefficients for *number of rooms* for each city? Comment on the size of the coefficients and the standard errors. Are these coefficients the same that you obtained in `(a)`?

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

9. 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, I'll reward you with a bonus on your grade.