Problem Set 1

Econ 4676: Big Data and Machine Learning for Applied Economics

Due Date: September 15 at 11:00 am

1 Theory Exercises: Review Stats and OLS

1. Fun wih conditional independence. Suppose that $X \in \{1, ..., K\}$ is a discrete random variable, and let ϵ_1 and ϵ_2 be realizations of two other random variables E_1 and E_2, such that $E_1 = \epsilon_1$ and $E_2 = \epsilon_2$. We want to calculate the flowing probability: $P(X|\epsilon_1, \epsilon_2) = P(X = 1|\epsilon_1, \epsilon_2) \dots P(X = K|\epsilon_1, \epsilon_2)$. (Hint: use Bayes rule.)

   (a) Which of the following sets of numbers are enough for the calculation?

      i. $P(\epsilon_1, \epsilon_2)$, $P(X)$, $P(\epsilon_1|H), P(\epsilon_2|X)$
      ii. $P(\epsilon_1, \epsilon_2)$, $P(X)$, $P(\epsilon_1, \epsilon_2|X)$
      iii. $P(\epsilon_1|X)$, $P(\epsilon_2|X)$, $P(X)$

   (b) Now assume $E_1 \perp E_2|X$ (i.e., E_1 and E_2 are conditionally independent given X). Which of the above 3 sets are enough now?

2. Consider the regression model $y_i = \alpha + \beta x_i + \epsilon, i = 1, ..., N,$, that is 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 $\hat{\beta} = (X'X)^{-1}X'y$?

3. 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.

4. 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

5. 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

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

   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 $R_K^2 > R_{K-1}^2$

6. 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 the classical assumptions hold. Let $\hat{\alpha}_N$ and $\hat{\beta}_N$ be the OLS estimators for α and β respectively. Suppose that this N observations are your train set. You test set consist in one observation and you make a prediction $\hat{y}_{test} = \hat{\alpha}_N + \hat{\beta}_N x_{test}$. Show that $E(\hat{y}_{test} - y_{test}) = 0$ and $Var(\hat{y}_{test} - y_{test}) = \sigma_0 \left(1 + \frac{1}{N} + \frac{(x_{test} - \bar{x})^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}\right)$

7. Consider the regression model $y_i = \beta x_i + \epsilon, i = 1, ..., N,$, that is a model with only one regressor and **no** intercept. Moreover, x_i has been standardized to have mean zero and variance one. All the classical assumptions hold and x_i is fixed (x_i is a non random variable. Consider estimating $\hat{\beta}^r = \frac{\hat{\beta}^{OLS}}{1+\gamma}$ where $\hat{\beta}^{OLS}$ is the OLS estimator and γ is a positive scalar. This is known as the ridge estimator.

   (a) Calculate the bias and the variance of $\hat{\beta}^r$ =,for a fixed γ. Compare it to the bias and variance of the OLS estimator

   (b) calculate the MSE and compare it to the MSE of the OLS estimator. Show that there is a γ for which $MSE(\hat{\beta}^r) < MSE(\hat{\beta})$

   (c) What conclusion can you take with respect of the classical econometric practice of strictly preferring unbiased estimators?

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 Predicting House Prices in Colombia

This part of the problem set involves data on housing prices in Colombia para Barranquilla, Bogotá D.C., Cali, Medellín. 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 called `house_prices_ps1` is available in the `data` folder in the problem set's repo. We will now try to predict the asking price using the other variables in the data set.

1. We will estimate variations of the following model: $y = X\beta + u$ where y is the listed price, X are the available observable characteristics for the houses in the data set.

2. Build a table with descriptive statistics for the potential variables you can use to predict house 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 it

3. *"The Taming of the Shrew"*

   (a) 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.)

   (b) Start with a model that only includes a constant. Calculate and report the average prediction error.

   (c) Estimate more complex models, you can add more variables, interaction, transformations, etc. I expect that you estimate at least 5 models. Report and compare the average prediction error.

   (d) Discuss the model with the lowest average prediction error.

4. Are there some other variables missing, e.g., amenities, that could potentially help the prediction. How could you obtain these variables and add them to this data set? I welcome any external data that you can merge to aid in the prediction and count as a bonus in the grade.

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

   (a) Write a loop for $i = 1$ to $i = n$, where n is the number of observations in the dataset that goes through the following steps:

      i. Estimate the regression model using all but the $i - th$ observation.

      ii. Calculate the prediction error $i - th$ observation $y_i - \hat{y}_i$

      iii. 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.

   (b) 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.

   (c) Use the statistic derived in the previous point to calculate the LOOCV statistic for the models estimated in (3b) and (3c). Discuss your results.

6. *Artisanal MapReduce*

   (a) Estimate your preferred model the full data set 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`, or in `Python` to `sklearn.linear_model.LinearRegression`).

   (b) Partition the data into at least 4 parts, one part for each city (if you want you can partition it further).

   (c) Repeat the `QR` estimation for the 4 partitions/cities.

   (d) Use the MapReduce approach to obtain the results in point (6a)

2.2 Reverend Bayes meets Web Scraping

In this part, we will explore immigrants' salaries in the U.S. We will use data from the H1B Salary Database (`https://h1bdata.info/`). H1B visas are visas that "allow U.S. employers to temporarily employ foreign workers in specialty occupations" (Wikipedia).

1. Explore the H1B Salary Database at https://h1bdata.info/

2. Look for the `robots.txt` file. Are there any restrictions to accessing/scraping these data?

3. Scrape data for New York, Chicago, Los Angeles, Houston, and San Francisco, for the period between 2016 and 2020.

4. Clean and describe the data. Here I expect you to decide what to do with data with missing values, similar names, etc., be clear and honest in your description. There are no wrong answers.

5. Plot and describe the number of applications filed by cities and by years.

6. Rank companies by city and year according to the number of applications filed. Do the top 3 companies in terms of applications change over time? Do a little bit of research about the companies and the kind of work they do. Write very briefly about them.

7. Next, keep only data for New York and the year 2020.

   (a) Estimate the base salary mean, variance, and number of observations for each company.

   (b) Lets assume that these means (ϑ) come from the following model

$$\vartheta_i|\theta_i \sim_{iid} N(\theta_i, \sigma^2/n_i)\, i = 1, \ldots, n \tag{3}$$
$$\theta_i \sim_{iid} N(\mu, \tau^2)\, i = 1, \ldots, n \tag{4}$$

   where i are indexing the companies

   i. Find the posterior distribution of θ_i
   ii. Find the marginal distribution of ϑ_i.
   iii. Now put your *Empirical Bayes* hat and use the marginal distribution to estimate the prior parameters. Plug in to estimate the posterior means.
   iv. Are the *Empirical Bayes* estimates are shrunken towards the overall city mean: yes, no, why?

3 Pedes in terra ad sidera visus (optional for bonus)

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.