Problem Set 1: Predicting Income

Ciencia de Datos para la toma de decisiones en Economía

Due Date: Friday, September 9 at 4 pm

1 Introduction

In the public sector, accurate reporting of individual wages is critical for computing taxes. However, tax fraud of all kinds has always been a significant issue. According to the Internal Revenue Service (IRS), about 83.6% of taxes are paid voluntarily and on time in the US.[1]. One of the causes of this gap is the under-reporting of wages by individuals. A wage predicting model could potentially assist in flagging cases of fraud that could lead to the reduction of the gap. Furthermore, a wage prediction model can help identify vulnerable individuals and families that may need further assistance.

The objective of the problem set is to apply the concepts we learned using "real" world data. For that, we are going to scrape from the following website: https://ignaciomsarmiento.github.io/Barranquilla/. This website contains data for Barranquilla from the 2018 GEIH.

Please turn a pdf document to i.sarmiento@uniandes.edu.co.

1.1 General Instructions

The main objective is to construct a model of individual wages

$$w = f(X) + u \tag{1}$$

where w is the wage that an individual receives, and X is a matrix that includes potential explanatory variables/predictors. In this problem set, we will focus on $f(X) = X\beta$.

1. *Data.* We will use data for Barranquilla from the 2018 GEIH

    (a) Data acquisition

        i. Scrape the data that is available at the following website https://ignaciomsarmiento.github.io/Barranquilla/.

        ii. Describe the data set briefly, the process of acquiring the data, and if there any restrictions to accessing/scraping these data.

[1] See https://www.irs.gov/newsroom/the-tax-gap.

(b) Data Cleaning and Description.[2] The data set contains all individuals sampled in Barranquilla. However, in this problem set, we will focus only on employed individuals older than eighteen (18) years old receiving a wage. Restrict the data to these individuals and perform a descriptive analysis of the variables used in the problem set. Keep in mind that the GEIH:

- Contains multiple wage measures. Choose those that best suit your analysis.
- There are many observations with missing data or 0 wage. I leave it to you to find a way to handle these data.

At a minimum, you should include a descriptive statistics table. Still, I expect a deep analysis that helps the reader understand your data, its variation, and the justification for your data choices. Use your professional knowledge to add value to this section. Do not present it as a "dry" list of ingredients.

2. *Age-wage profile.* A great deal of evidence in Labor economics suggests that the typical worker's age-wage profile has a predictable path: *"Wages tend to be low when the worker is young; they rise as the worker ages, peaking at about age 50; and the wage rate tends to remain stable or decline slightly after age 50".*

In this subsection we are going to estimate the *Age-wage profile* profile for the individuals in this sample:

$$log(w) = \beta_1 + \beta_2 Age + \beta_3 Age^2 + u \tag{2}$$

When presenting your results, also include:

- A discussion of the variable chosen as a measure of wage.
- An interpretation of the coefficients and it's significance.
- A discussion of the model's in sample fit.
- A plot of the estimated age-earnings profile implied by the above equation. Including a discussion of the "peak age" with it's respective confidence intervals. (Note: Use bootstrap to construct the confidence intervals.)

3. *The gender earnings GAP.* Policymakers have long been concerned with the gender wage gap, and is going to be our focus in this subsection.

(a) Begin by estimating and discussing the unconditional wage gap:

$$log(w) = \beta_1 + \beta_2 Female + u \tag{3}$$

where $Female$ is an indicator that takes one if the individual in the sample is identified as female.

[2] This section is located here so the reader can understand your work, but probably it should be the last section you write. Why? Because you are going to make data choices in the estimated models. And all variables included in these models should be described here.

(b) Next, estimate, plot, and discuss the predicted age-wage profile and the implied "peak ages" with the respective confidence intervals by gender.

(c) *Equal Pay for Equal Work?* A common slogan is "equal pay for equal work".One way to interpret this is that for employees with similar worker and job characteristics, no gender wage gap should exist. Estimate a conditional earnings gap incorporating control variables such as similar worker and job characteristics. Compare the estimes to one that uses bootstrap to obtain the standard errors.

When presenting and discussing your results, don't forget to address the following issues:

- A discussion of the variable chosen as a measure of wage, if it is the same or different from the previous point.
- An interpretation of the "Female" coefficients, a comparison between the models, and the in-sample fit.[3]
- A discussion about the implied peak ages and their statistical similarity/difference.
- A thoughtful discussion about the unconditional and conditional wage gap, seeking to answer if the changes in the coefficient are evidence of a selection problem, a "discrimination problem," or none of these issues.

4. *Predicting earnings.* In the previous sections, you estimated some specifications with inference in mind. In this subsection, we will evaluate the predictive power of these specifications.

   (a) Split the sample into two: a training (70%) and a testing (30%) sample. (Don't forget to set a seed to achive reproducibility. In R, for example you can use `set.seed(10101)`, where 10101 is the seed.)

   (b) Report and compare the predictive performance of all the previous specifications with at least five (5) additional specifications that explore non-linearities and complexity.

   (c) In your discussion of the results, comment:

      i. About the performance metric you have chosen and your rationale for choosing it.

      ii. About the specification with the lowest prediction error.

      iii. For the specification, explore those observations that seem to "miss the mark." To do so, compute the influence statistic for each observation in the test sample, and examine its distribution. Are observations in the tails of the influence statistic distribution? Are these outliers potential people that the DIAN should look into, or are they just the product of a flawed model?

[3] Tip: Look how applied papers construct their results tables. These papers usually present comparable results in the same table with coefficients side by side, which helps the reader follow the discussion.

(d) *LOOCV*.For the two models with the lowest predictive error in the previous section, calculate the predictive error using Leave-one-outcross-validation (LOOCV). Compare the results of the test error with those obtained with the validation set approach and explore the potential links with the influence statistic. (Note: when attempting this subsection, the calculations can take a long time depending on your coding skills, plan accordingly!)