Problem Set 1: Predicting Income

1 Introduction

In the public sector, accurate reporting of individual income 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 incomes by individuals. An income predicting model could potentially assist in flagging cases of fraud that could lead to the reduction of the gap. Furthermore, an income 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/GEIH2018_sample/. This website contains data for Bogotá from the 2018 *"Medición de Pobreza Monetaria y Desigualdad Report"* that takes information from the GEIH.

1.1 General Instructions

The main objective is to construct a model of individual hourly wages

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

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

The final document, in `.pdf` format, must contain the following sections:

1. *Introduction*. The introduction briefly states the problem and if there are any antecedents. It briefly describes the data and its suitability to address the problem set question. It contains a preview of the results and main takeaways.

2. *Data.*[2] We will use data for Bogotá from the 2018 *"Medición de Pobreza Monetaria y Desigualdad Report"* that takes information from the GEIH.

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

[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.

The data set contains all individuals sampled in Bogota and is available at the following website https://ignaciomsarmiento.github.io/GEIH2018_sample/. To obtain the data, you must scrape the website.

In this problem set, we will focus only on employed individuals older than eighteen (18) years old. Restrict the data to these individuals and perform a descriptive analysis of the variables used in the problem set. Keep in mind that in the data, there are many observations with missing data or 0 wages. I leave it to you to find a way to handle this data.

When writing this section up, you must:

(a) Describe the data briefly, including its purpose, and any other relevant information.

(b) Describe the process of acquiring the data and if there are any restrictions to accessing/scraping these data.

(c) Describe the data cleaning process and

(d) Descriptive the variables included in your analysis. At a minimum, you should include a descriptive statistics table with its interpretation. However, I expect a deep analysis that helps the reader understand the 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.

3. *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 and discussing your results, include:

   - A regression table.
   - 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.)

4. *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.

(b) *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. In this section, estimate the conditional wage gap:

i. First, using FWL

ii. Second, using FWL with bootstrap. Compare the estimates and the standard errors.

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

When presenting and discussing your results, include:

- An estimating equation, explaining the included control variables (beware of *"bad controls"*).
- A regression table, with the estimates side by side of the conditional and unconditional wage gaps, highlighting the coefficient of interest. Controls, should not be included in the table but dutifully noted. [3]
- An interpretation of the "Female" coefficients, a comparison between the models, and the in-sample fit.
- 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," a mix, or none of these issues.

5. *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 achieve reproducibility. In R, for example you can use `set.seed(10101)`, where 10101 is the seed.)

[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.

(b) Report and compare the predictive performance in terms of the RMSE 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 overall performance of the models.

  ii. About the specification with the lowest prediction error.

  iii. For the specification with the lowest prediction error, explore those observations that seem to "miss the mark." To do so, compute the prediction errors in the test sample, and examine its distribution. Are there any observations in the tails of the prediction error distribution? Are these outliers potential people that the DIAN should look into, or are they just the product of a flawed model?

(d) *LOOCV*. For the two models with the lowest predictive error in the previous section, calculate the predictive error using Leave-one-out-cross-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!)

2 Additional Guidelines

- Turn a `.pdf` document in Bloque Neón.
- The document must include a link to your GitHub Repository.
  - The repository must follow the template.
  - The README should help the reader navigate your repository. A good README helps your project stand out from other projects and is the first file a person sees when they come across your repository. Therefore, this file should be detailed enough to focus on your project and how it does it, but not so long that it loses the reader's attention. For example, Project Awesome has a curated list of interesting READMEs.
  - Include brief instructions to fully replicate the work.
  - The main repository branch should show at least five (5) substantial contributions from each team member.
  - The code has to be:
    - Fully reproducible.
    - Readable and include comments. In coding, like in writing, a good coding style is critical. I encourage you to follow the tidyverse style guide.

- Tables, figures, and writing must be as neat as possible. Label all the variables included. If you have something in your figures or tables, I expect they are addressed in the text. Tables must follow the AER format.