Problem Set 1: Predicting Income

Due Date: Sunday, November 24

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.

Please turn a pdf document to ignaciomsarmiento@gmail.com.

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) A descriptive analysis of the data. 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. *Predicting wages* . In this section, we will evaluate the predictive power of different specifications.

(a) Split the sample into two: a training (70%) and a testing (30%) sample. (Use 123 as a `random state` to achieve reproducibility.

(b) Report and compare the predictive performance (RMSE) of at least ten (10) specifications that explore different levels of complexity and non-linearities.

(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. (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 to ignaciomsarmiento@gmail.com.
- 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.
- 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.