Problem Set 1: Predicting Income

MECA 4107

Due Date: June 26 at 23:59 on Bloque Neón

June 10, 2022

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 GEIH.

1.1 General Instructions

The main objective is to construct a predictive model of individual income

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

Where $Income$ is the income that an individual receives, and X is a matrix that includes potential predictors. In this problem set, we will focus on $f(X) = X\beta$

1. *Data acquisition*
   (a) Scrape the data that is available at the following website https://ignaciomsarmiento.github.io/GEIH2018_sample/.
   (b) Are there any restrictions to accessing/scraping these data?
   (c) Using `pseudocode` describe your process of acquiring the data

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

2. *Data Cleaning.* In this problem set, we will focus only on employed individuals older than eighteen (18) working in Bogotá. In this section, you are going to focus on cleaning and describing the data.

    - The data set include multiple variables that can help explain individual income. Guided by your intuition and economic knowledge, choose the most relevant and perform a descriptive analysis of these variables. For example, you can include variables that measure education and experience, given the implications of the human capital accumulation model (Becker, 1962, 1964; and Mincer (1962, 1975).
    - Note that there are many observations with missing data. I leave it to you to find a way to handle these missing data. In your discussion, describe the steps that you performed cleaning de data, and justify your decisions.
    - At a minimum, you should include a descriptive statistics table, but I expect tables and figures. Take this section as an opportunity to present a compelling narrative to justify and defend 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-earnings profile.* A great deal of evidence in Labor economics suggests that the typical worker's age-earnings 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 the data set, multiple variables describe income. Choose one that you believe is the most representative of the workers' total earnings, justifying your selection.
    - Based on this estimate using OLS the age-earnings profile equation:

$$Income = \beta_1 + \beta_2 Age + \beta_3 Age^2 + u \tag{2}$$

    - How good is this model in sample fit?
    - Plot the predicted age-earnings profile implied by the above equation.
    - What is the "peak age" suggested by the above equation? Use bootstrap to calculate the standard errors and construct the confidence intervals.

4. *The earnings GAP.* Most empirical economic studies are interested in a single low dimensional parameter, but determining that parameter may require estimating additional "nuisance" parameters to estimate this coefficient consistently and avoid omitted variables bias. Policymakers have long been concerned with the gender earnings gap.

    - Estimate the unconditional earnings gap

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

- How should we interpret the β_2 coefficient? How good is this model in sample fit?
- Estimate and plot the predicted age-earnings profile by gender. Do men and women in Bogotá have the same intercept and slopes?
- What is the implied "peak age" by gender?. Use bootstrap to calculate the standard errors and construct the confidence intervals. Do these confidence intervals overlap?
- *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 earnings gap should exist. Estimate a conditional earnings gap that incorporates control variables such as similar worker and job characteristics (X).
  (a) Estimate the conditional earnings gap $log(Income) = \beta_1 + \beta_2 Female + \theta X + u$
  (b) Use FWL to repeat the above estimation, where the interest lies on β_2. Do you obtain the same estimates?
  (c) How should we interpret the β_2 coefficient? How good is this model in sample fit? Is the gap reduced? Is this evidence that the gap is a selection problem and not a "discrimination problem"?

5. *Predicting earnings.* Now we turn to prediction. You built a couple of models in the previous section using your knowledge as an applied economist, the task here is to assess the predictive power of these models.

   (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.)
      i. Estimate a model that only includes a constant. This will be the benchmark.
      ii. Estimate again your previous models
      iii. In the previous sections, the estimated models had different transformations of the dependent variable. At this point, explore other transformations of your independent variables also. For example, you can include polynomial terms of certain controls or interactions of these. Try at least five (5) models that are increasing in complexity.
      iv. Report and compare the average prediction error of all the models that you estimated before. Discuss the model with the lowest average prediction error.
      v. For the model with the lowest average prediction error, compute the leverage statistic for each observation in the test sample. Are there any outliers, i.e., observations with high leverage driving the results? Are these outliers potential people that the DIAN should look into, or are they just the product of a flawed model?

(b) Repeat the previous point but use K-fold cross-validation. Comment on similarities/differences of using this approach.

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

i. Write a loop that does the following:
   - Estimate the regression model using all but the $i-th$ observation.
   - Calculate the prediction error for the $i-th$ observation, i.e. $(y_i - \hat{y}_i)$
   - Calculate the average of the numbers obtained in the previous step to get the average mean square error. This is known as the Leave-One-Out Cross-Validation (LOOCV) statistic.

ii. Compare the results to those obtained in the computation of the leverage statistic

2 Additional Guidelines

I expect the following things from the problem set, omission of any of these guidelines will be penalized.

- You should turn in your document in bloque neón.
- The document should point and include a link to your GitHub Repository.
- You should follow the repository template.
- The repository should include a README file. 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.
- The repository should have at least five (5) substantial contributions from each team member.
- Tables, figures, and writing should be as neat as possible. Label all variables that you include. Any variable, a statistic, etc., included in your figures or tables should be described in the text.
- Your code should be readable and include comments. In coding, like in writing, a good coding style is critical for readable code. I encourage you to follow the tidyverse style guide.