Problem Set 1: Predicting Income

MECA 4107

**Due Date**: June 26 at 23:59 on Bloque Neo´n June 10, 2022

# 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](#_bookmark0). 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://ignaciomsarmient](https://ignaciomsarmiento.github.io/GEIH2018_sample/)o. [github.io/GEIH2018 sample/](https://ignaciomsarmiento.github.io/GEIH2018_sample/). This website contains data for Bogot´a from the 2018 [GEIH](https://www.dane.gov.co/index.php/estadisticas-por-tema/mercado-laboral/empleo-y-desempleo/geih-historicos).

* 1. **General Instructions**

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

*Income* = *f* (*X*) + *u* (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*) = *Xβ*

* + 1. *Data acquisition*
       1. Scrape the data that is available at the following website [https://ignaciomsarmi](https://ignaciomsarmiento.github.io/GEIH2018_sample/)ento. [github.io/GEIH2018 sample/](https://ignaciomsarmiento.github.io/GEIH2018_sample/).
       2. Are there any restrictions to accessing/scraping these data?

Así es, no fue posible verificar la base de datos de forma fácil y rápida, ya que al inspeccionar la página web, no se identificó el enlace que nos lleva a los datos. Sin embargo, mediante la opción “ver código fuente de la página” fue donde

Era necesario obtener las tablas de cada chunk, sin embargo, no fue sencillo identificarlas ya que las tablas no se encontraban en ese enlace, sino que se obtenían por medio de otro enlace atado al principal. Adicionalmente, la información dela GEIH fue dividida en 10 chunks lo que complicaba aún más obtener la información.

* + - 1. Using pseudocode describe your process of acquiring the data

Inicio del proceso

1. Se limpia el entorno de trabajo en R
2. Se utilizan las librerías necesarias para la totalidad del proceso: pacman, tidyverse, rvest, datasets y data.table.
3. Una vez identificada la url en la que se encontraba cada una de las listas, se procede a crear un loop para cargar la información de cada tabla. La url correspondiente a la primera tabla es la siguiente: <https://ignaciomsarmiento.github.io/GEIH2018_sample/pages/geih_page_1.html>

Para realizar el loop se crea un valor nombrado url que contiene el link anterior sin el numero de la página, así: <https://ignaciomsarmiento.github.io/GEIH2018_sample/pages/geih_page_>

1. Se crea una lista nombrada “tablas” que va a servir para guardar la información de cada tabla.
2. Se ejecuta el loop pegando el valor creado con cada iteración “i” del 1 al 10, ya que la información se encuentra en 10 enlaces. Posteriormente, se carga cada tabla en la lista “tablas” creada.
3. Considerando que la primera columna de las 10 tablas no tiene nombre, se procede a nombrarla para poder unir las tablas y convertirlas a tibble.

Fin del proceso

El proceso anterior se encuentra en el archivo 1\_scraping.R incluido en la carpeta 2. Scripts.

1See <https://www.irs.gov/newsroom/the-tax-gap>

* + 1. *Data Cleaning*. In this problem set, we will focus only on employed individuals older than eighteen (18) working in Bogota´. 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).

* Teoría de capital humano de Becker (1975) y Rosen (1996)

Escolaridad, género, estado civil, sector, horas trabajadas – variables mincerianas

<https://core.ac.uk/download/pdf/76596209.pdf>

experiencia –

* relación positiva de los años de educación y la experiencia con el salario, así como la existencia de diferencias salariales entre hombres y mujeres y el hecho de que la experiencia posee retornos decrecientes, dado que, aunque cada año adicional de esta variable aporta a los salarios cada vez lo hace en menor medida y en cierto punto el aporte es nulo.

1. **Salario y zona regional** Algunos de los autores analizados toman como variable independiente o explicativa la zona donde se realiza la actividad productiva a fin de identificar posibles diferencia salariales, entre estos tenemos a Lema, D., & Casellas, K. (2009)

edad del individuo

Fukusaki, G. Y. (2007) r, tomando como variables independientes los años de escolaridad, la experiencia, el estado civil, el género femenino, el lugar de residencia (rural o urbano)

Kanas, A.; Van Tubergen, F.; Van der Lippen, T. (2009) estudian el papel del capital humano y social en el autoempleo de inmigrantes

<https://repository.ucatolica.edu.co/bitstream/10983/15983/1/TDGII%20CHALAYCARRERA%20FINAL.pdf>

* Con el tiempo las aportaciones científicas demostraron que en la medición del rendimiento de la educación estimado por MCO (ecuación minceriana de ingresos) existían varios sesgos, Griliches (1977), identificó los siguientes: a) Existencia de determinadas variables omitidas en la ecuación de Mincer como la habilidad del individuo. Si el término de perturbación incluye entre otros elementos la habilidad del individuo, y se cumple que las personas con mayor habilidad son las que deciden elegir el mayor nivel educativo, esto provocaría estimaciones inconsistentes, pues la perturbación aleatoria y el nivel educativo (parámetro S) estarían correlacionados y esto daría, como resultado, que se sobreestimase el rendimiento de la educación. b) Medición incorrecta de la cantidad de educación. La ecuación de Mincer (1) implica que existe una única tasa de rendimiento de la educación, sin embargo, tanto la teoría como la evidencia empírica lo contradicen, sugiriendo representar la escolaridad en forma más desagregada y flexible, con el objetivo de recoger de una forma más adecuada los retornos al proceso de inversión en educación (Sapelli, 2009). El no recoger correctamente el nivel educativo estaría provocando una infraestimación en el verdadero rendimiento. c) Tratamiento de la educación como una variable exógena. Diversos estudios han demostrado que la variable educación puede no ser exógena estando afectada por diferentes variables como puede ser los costes directos de la educación, los costes de oportunidad, los ingresos futuros, las imperfecciones del mercado de capitales o incluso la importancia del motivo consumo. El no considerar la endogeneidad de la educación producirá que las estimaciones realizadas por la técnica de MCO proporcione estimadores sesgados e inconsistentes del rendimiento de la educación2 . Si se demuestra que la variable educación es endógena, el método más apropiado para calcular sus rendimientos sería estimarlo por “mínimos cuadrados en dos etapas” (MC2E). En este método la variable que causa la autocorrelación se elimina sustituyéndola por una variable instrumental (ecuación 2)

<https://2010.economicsofeducation.com/user/pdfsesiones/095.pdf>

* Sin embargo, los individuos, aun teniendo el mismo nivel de educación formal, no son iguales entre sí, lo cual plantea características distintivas que se destacan en el mercado laboral. Las diferencias pueden encontrarse en el concepto de empleabilidad, definido sucintamente como la “habilidad para obtener o conservar un empleo” (Campos 2003).

La capacidad de empleabilidad de un individuo está en función de la combinación de diferentes elementos, tales como la agilidad mental, los conocimientos, la actitud ante los cambios, la inteligencia emocional o la capacidad de relacionarse. Fundipe (1999) divide las habilidades relacionadas con la empleabilidad en los siguientes cuatro grupos:

<https://journals.openedition.org/revestudsoc/8023>

* La Teoría del Capital Humano debe ser considerada en el marco del análisis de la teoría neoclásica. A partir de los trabajos de Becker (1962,1964) y Mincer (1974), se considera que la productividad de un individuo depende de su dotación de capital humano, ceteris paribus. Predice que, en promedio, a mayores niveles de educación cabe esperar mayores ingresos y que la demanda de educación depende de los costos directos e indirectos y de los ingresos futuros. Schultz (1961) y Denison (1962) demostraron que la educación contribuye directamente al crecimiento económico nacional al mejorar la calidad y la productividad de la fuerza de trabajo.

file:///C:/Users/amorales/Downloads/Dialnet-LaImportanciaDeLaEducacionEnElDesarrollo-3641304.pdf

xxxx

* 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.

xxxxx

* 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 knowl- edge to add value to this section. Do not present it as a “dry” list of ingredients.

Xxxxx

Table estadísticas descriptivas

Histogramas, frecuencia, mapa de calor de correlograma

* + 1. *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* = *β*1 + *β*2*Age* + *β*3*Age*2 + *u* (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.
  + 1. *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*) = *β*1 + *β*2*Female* + *u* (3)

* How should we interpret the *β*2 coefficient? How good is this model in sample fit?

XXXX

* Estimate and plot the predicted age-earnings profile by gender. Do men and women in Bogota´ have the same intercept and slopes?

XXXX

* 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?

XXXX uno para hombre y otro para mujer

* *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*).
  + - 1. Estimate the conditional earnings gap *log*(*Income*) = *β*1 + *β*2*Female* +

*θX* + *u*

*variables control: edad, educación, experiencia – Pto2*

* + - 1. Use FWL to repeat the above estimation, where the interest lies on *β*2. Do you obtain the same estimates?

XXXX..

* + - 1. 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”?

XXXXX

* + 1. *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.
       1. 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.)
          1. Estimate a model that only includes a constant. This will be the benchmark.
          2. Estimate again your previous models
          3. 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.
          4. Report and compare the average prediction error of all the models that you estimated before. Discuss the model with the lowest average prediction error.
          5. For the model with the lowest average prediction error, compute the lever- age statistic for each observation in the test sample. Are there any outliers, i.e., observations with high leverage driving the results? Are these out- liers potential people that the DIAN should look into, or are they just the product of a flawed model?
       2. Repeat the previous point but use K-fold cross-validation. Comment on simi- larities/differences of using this approach.
       3. *LOOCV*. With your preferred predicted model (the one with the lowest average prediction error) perform the following exercise:
          1. 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. (*yi − y*ˆ*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.

* + - * 1. Compare the results to those obtained in the computation of the leverage statistic

# 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 ne´on.
* The document should point and include a link to your GitHub Repository.
* You should follow the [repository template](https://github.com/ignaciomsarmiento/MECA_PS_Repo).
* 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](https://project-awesome.org/) 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](https://style.tidyverse.org/) [style guide](https://style.tidyverse.org/).

########################################

############### DATOS ##################

############# PROBLEM\_SET\_1 ############

# Limpieza de espacio

rm(list = ls())

# Llamado librería

require(pacman)

p\_load(tidyverse,

rvest)

# scrape de datos como tabla desde sitio el web:

url <- "https://ignaciomsarmiento.github.io/GEIH2018\_sample/pages/geih\_page\_"

# Almacenamiento de los datos en tablas

tablas = list()

for (i in 1:10) {

url\_i = paste0(url,i,".html")

tablas[[i]] <- url\_i %>% read\_html() %>% html\_table()

}

# Revisión de tablas

tablas[[1]]

tablas[[5]]

# Renombre primera columna

library(dplyr)

data(package="datasets")

colnames(tablas)[2] <- "item"

# Revisión de tablas

tablas[[2]]