Universidad de los Andes

Maestría en Economía Aplicada

Angela Paola Morales Guio – 201015503

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 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 de la 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

Nota: 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).

De acuerdo con los aportes de Mincer (1958), Schultz (1960) y Becker (1964) a la Teoría Económica sobre capital humano, las variables que explican el ingreso individual son los años de escolaridad y la experiencia profesional, sin embargo, existen variables omitidas en la ecuación de Mincer y que controlan el ingreso individual como lo son: la habilidad del individuo[[1]](#footnote-1) (ya que aún teniendo el mismo nivel de educación, los individuos no son iguales entre sí, existiendo características distintivas en el mercado laboral), el género ya según el DANE (DANE, 2020) para el año 2019 en Colombia, la brecha salarial general entre hombres y mujeres, según la media, fue de 12,9% de acuerdo con la GEIH, la raza y el tipo de ocupación.

De acuerdo con lo anterior, seleccionamos las siguientes variables:

* Ingreso total:

y\_total\_m income salaried + independents total - nominal monthly

ingtot Ingreso total

* Educación:

• relación positiva de los años de educación y la experiencia con el salario

Endógena, 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.

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

* Experiencia: 6210

Dado que la GEIH no cuenta con información relacionada con los años trabajados en total sino con el tiempo que ha durado en el cargo actual, la literatura ha utilizado como proxy de la experiencia la experiencia potencial. Esta surge de restarle a la edad de la persona los años que ha estudiado y, además, le resta cinco (5) años más en representación de los años de primera infancia en los que no se estudió ni se trabajó.

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.

* Oficio

Ocupación, variable dividida en 7 categorías

* Relab

Tipo de ocupación, variable dividida en 9 categorías

* Variable age
* Variable sex

Descripción =1 male =0 female

* Horas trabajadas:
* estrato1 Estrato de energía para las 13 a.M., y sextil de icv para otras cabeceras y rest

Con las variables presentadas, consideramos el siguiente modelo de regresión lineal:

ln(w𝑖) = 𝛽0 + 𝛽1Esc𝑖 + 𝛽2Exp𝑖 + 𝛽3Race𝑖 + 𝛽4Gen𝑖 + 𝜀𝑖

edad pet

ocupación género

* 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.

En la base de datos de la GEIH para Bogotá en el año 2018 se encontraron las siguientes variables relacionados con el ingreso de los trabajadores:

|  |  |
| --- | --- |
| Variable | Descripción |
| Ingtotob | Ingreso monetario primera actividad, ingreso segunda actividad, ingreso en especie, ingreso monetario des ocupados e inactivos e ingresos provenientes de otras fuentes no laborales |
| Ingtotes | Ingreso total por persona que resulta de sumar cada una de las fuentes de ingresos imputadas a los registros faltantes |
| Ingtot | Ingreso total por persona que resulta de sumar cada una de las fuentes de ingresos tanto observadas como imputadas |
| y\_ingLab\_m | labor income salaried - nominal monthly - all occ. (includes tips and commission |
| y\_salary\_m | salary - nominal monthly - principal occ. (includes tips and commissions) |
| y\_total\_m | income salaried + independents total - nominal monthly |

En primer lugar, descartamos las tres (3) primeras variables (Ingtot, Ingtotes y Ingtotob) puesto que estas incluyen ingresos de personas desocupados e inactivos.

Después nos concentramos en revisar lo que pide el numeral, lo cual corresponde a las ganancias totales de los trabajadores, y puesto que esto incluye tanto los ingresos laborales como los no laborales, seleccionamos la variable y\_total\_m

* 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?

A continuación, se presenta los resultados del modelo de regresión realizado en R con los datos filtrados y seleccionados de la GEIH:

===============================================

Dependent variable:

---------------------------

y\_total\_m

-----------------------------------------------

age 124,116.000\*\*\*

(9,117.985)

age2 -1,331.928\*\*\*

(106.340)

Constant -961,012.000\*\*\*

(181,870.000)

-----------------------------------------------

Observations 14,631

R2 0.014

Adjusted R2 0.014

Residual Std. Error 2,422,881.000 (df = 14628)

F Statistic 106.885\*\*\* (df = 2; 14628)

===============================================

Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

Si bien el modelo arroja que las variables de edad y edad al cuadrado son relevantes para explicar el ingreso con un nivel de significancia del 1% y el modelo demuestra que hay dependencia (F Estadístico) también con un nivel de significancia del 1%, el grado de ajuste del modelo (R2) es muy bajo esto derivado a que hay otras variables que explican el ingreso de las personas tal como lo plantea Jacob Mincer (1974) que también depende de los años de escolaridad.

Adicionalmente varios autores han planteado la relación entre salario y género, puesto que hay una diferencia entre los salarios que reciben los hombres y las mujeres bajo las mismas condiciones de edad y educación (Nordin & Persson, 2010).

* Plot the predicted age-earnings profile implied by the above equation.

Gráfico

Descripción generada automáticamente

* What is the “peak age” suggested by the above equation? Use bootstrap to calculate the standard errors and construct the confidence intervals.

Con el modelo se obtienen los siguientes valores de los coeficientes:

* b1 = -961012
* b2 = 124116
* b3 = -1331.928

Se obtiene que la edad pico es 46.59 años

ORDINARY NONPARAMETRIC BOOTSTRAP

Call:

boot(data = dbIncome, statistic = eta\_mod.fn, R = 1000)

Bootstrap Statistics :

original bias std. error

t1\* 46.59261 0.0791927 0.9472881

* + 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, oficio*

* + - 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]]

1. De acuerdo con Rosen (1992), las habilidades del individuo se traducen en su capacidad innata, la cual afecta al valor marginal de la educación y al entorno familiar que afecta a las oportunidades de financiación. [↑](#footnote-ref-1)