Universidad de los Andes

Maestría en Economía Aplicada

Andres Felipe Martinez - XXXXXX

Angela Paola Morales Guio – 201015503

Oscar Cortes - XXXXXX

Problem Set 1: Predicting Income

MECA 4107

# 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. 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 llaman las librerías necesarias para la totalidad del proceso: pacman, tidyverse, rvest, datasets y data.table.
3. Para identificar la url requerida, se inspecciona el código html de la página web https://ignaciomsarmiento.github.io/GEIH2018\_sample , evidenciando que en cada enlace de los chunks de esta página web no se observan las tablas directamente como html table, por lo tanto, se encuentra en el código html el enlace que si contiene la tabla directamente para cada chunk.
4. Una vez identificada la url en la que se encontraba cada una de las tablas, 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>

* 1. Para realizar el loop se crea el valor nombrado “url” que contiene el link anterior sin el numero de la página con el fin de que el loop se encargue de llamar página por página y traer cada una de las 10 bases de datos, así:

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

* 1. Se crea un data frame nombrado “data” que va a servir para guardar la información de cada tabla generada por el loop, de la 1 a la 10.
  2. Se ejecuta el loop pegando el valor creado con cada iteración “i” del 1 al 10. Posteriormente, se carga cada tabla en la lista “data” creada así:

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

1. Considerando que la primera columna de las 10 tablas no tiene nombre, se procede a renombrarla como “id” para poder unir las tablas y convertirlas a tibble.

Captura de pantalla de computadora

Descripción generada automáticamente

Fin del proceso

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

* + 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.

En el proceso de data cleaning se seleccionó la variable ocupados la cual define 1= ocupado; 0= no ocupado, de acuerdo con los datos y el directorio, se deduce que esta variable corresponde a las personas que tiene trabajo independientemente de las características de este. Esto quiere decir que cuando la variable ocupado toma el valor de 1 el individuo está empleado, por otro lado cuando la variable toma el valor de 0 el individuo puede estar desempleado y/o también puede estar inactivo, por esta razón la variable “dsi” que considera los desempleados y los no desempleados, no es una variable idónea para elegir las personas empleadas ya que los no desempleados no necesariamente tienen un trabajo sino que pueden estar inactivos.

El objetivo del trabajo se enfoca en los individuos empleados (ocu=1) y mayores de 18 años. Se elige la variable “ocu” ya que relaciona todas las personas que tienen una ocupación y por lo tanto devengan un ingreso salarial, el enfoque en analizar el ingreso de los individuos que se encuentran empleados y esta variable recoge esa información.

Inicialmente, la base general de la GEIH contiene 32177 observaciones y una vez realizando el filtro mencionado anteriormente se reduce a 16397 observaciones.

* 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[[1]](#footnote-1), sin embargo, existen variables omitidas en la ecuación de Mincer y que controlan el ingreso individual como lo son: la habilidad del individuo[[2]](#footnote-2) (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, la raza, las horas de trabajo, el estrato, el tamaño de la empresa, el tipo de trabajo si es formal o informal y el tipo de ocupación.

De acuerdo con lo anterior, seleccionamos las siguientes variables:

* Ingreso total

Para seleccionar la variable de ingreso se analizaron las variables “y\_total\_m” e “ingtot”. Por un lado, “y\_total\_m” considera el ingreso salarial y el ingreso total de los independientes, por otro lado, “ingtot” relaciona el ingreso total que agrupa el ingreso salarial, arriendos, dividendos, pensiones, ingresos en especie, entre otros.

De acuerdo con el análisis realizado, se encontró que la variable “ingtot” realiza imputación a los registros faltantes[[3]](#footnote-3) y considera los ingresos no laborales, estas dos características de “ingtot” pueden afectar la inferencia estadística a realizar pues el objetivo a analizar es el ingreso laboral de los individuos ocupados sin realizar imputaciones a los datos [[4]](#footnote-4). Teniendo en cuenta lo anterior, se considera que la variable “y\_total\_m” recoge el ingreso laboral de los individuos ocupados sin imputar ingresos adicionales, variable escogida para el análisis.

* Educación

Considerando la revisión de literatura, Mincer (1958) menciona que existe una relación positiva entre los años de educación y el salario, Becker (1975) propone analizar las actividades que influyen en el ingreso, coincidiendo con Mincer en el argumento de incluir la educación, este autor plantea ver la escolarización como una inversión y por lo tanto se esperan retornos a la educación. De acuerdo con un estudio realizado por Tarazona y Remolina (2017), la tasa de retorno de la educación para Colombia es de 9,1%, indicador de cuánto puede crecer el salario por un año adicional de educación.

Se analizará el efecto de la variable “edu” en el ingreso laboral esperando que a mayor nivel de educación mayor será el ingreso laboral, esta variable se creó teniendo en cuenta las variables “maxEducLevel” y “p6210s1[[5]](#footnote-5)”, asignando el ultimo grado aprobado a cada nivel educativo en el cual se encuentra el individuo, así:

Tabla 1. Criterios variable Edu

| **maxEducLevel** | **Nivel** | **Edu  (años de educación finalizados)** |
| --- | --- | --- |
|  |
| 1 | Ninguno | 0 |  |
| 2 | Preescolar [[6]](#footnote-6) | 2 |  |
| 3 | Primaria incompleta | p6210s1 + 2 |  |
| 4 | Primaria completa (5 años) | 7 |  |
| 5 | Secundaria incompleta (6-10 años) | p6210s1 + 2 |  |
| 6 | Secundaria completa (11 años) | 13 |  |
| 7 | Terciaria | p6210s1 + 13 |  |

Elaboración propia

Fuente: XXXXX

* Oficio

Se considera importante incluir esta variable pues se puede controlar por cada una de las ocupaciones que existen, el ingreso de los individuos, bajo el supuesto que los oficios con los ingresos más bajos pueden tener requisitos mínimos de formación, lo que lleva a responsabilidades limitadas para los individuos, lo contrario ocurre con los oficios que requieren de formaciones avanzadas o especializadas de los individuos, siendo esta formación valorada a través del aumento potencial de los ingresos.

De acuerdo con esto, se considera que dependiendo del oficio pueden existir diferencias en los ingresos. Esta variable será importante para el objetivo de inferencia estadística en la brecha salarial de género y también para realizar pronósticos de ingreso.

La base de datos cuenta con la variable “oficio” la cual esta dividida en 82 ocupaciones, esta variable se incluirá en la regresión como un factor lo que implica una dummy para cada oficio.

* Edad

De acuerdo con la literatura, la edad es un factor importante que incide en la generación de ingreso, un año adicional genera incrementos en los ingresos hasta un punto determinado en el cual un año adicional de edad genera efectos marginales negativos en los ingresos.

Por lo anterior, se incluye en el análisis las variables “age” y edad al cuadrado “age2” para modelar estos rendimientos marginales decrecientes y de esta manera poder encontrar el peak age para el ingreso.

* Sexo

Según el DANE (DANE, 2020) para el año 2019 en Colombia de acuerdo con la GEIH, la brecha salarial general entre hombres y mujeres es de media de 12,9%, adicionalmente, existen numerosos trabajos y estudios que concluyen que las mujeres reciben salarios por debajo del que perciben los hombres, esto hace que sea muy importante incluir la variable sexo “female” para identificar las posibles brechas salariales de género.

“female” =1 mujer =0 hombre

* Horas trabajadas:

Se incluye esta variable ya que dependiendo de las horas que trabaje un individuo puede generar o no mayores ingresos comparados con otros individuos que trabajen menos horas y que tengan características similares, por lo tanto, horas trabajadas, variable “totalHoursWorked” puede ayudar a controlar el efecto de la brecha salarial que se busca estimar.

* Estrato

Se crean dos variables dummy a partir de la variable “estrato1” que informa el estrato de energía, la variable “estrato\_medio” recoge los estratos medio 3 y 4 y la variable “estrato\_alto” corresponde a los estratos altos 5 y 6, estas variables[[7]](#footnote-7) se incluyen en el análisis ya que puede estar evidenciando diferencias salariales por la pertenencia o no a un estrato socioeconómico, se espera que los estratos altos generen mayores ingresos comparados con los estratos bajos.

* Tamaño empresa

Se considera importante incluir esta variable porque puede controlar el ingreso dependiendo del tamaño de la empresa en donde trabaja el individuo, pueden existir diferencias en los ingresos por tamaño de empresa. Esta variable será importante para el objetivo de inferencia estadística en la brecha salarial de género y también para realizar pronósticos de ingreso.

La base de datos cuenta con la variable “sizeFirm” la cual está dividida en 5 categorías: cuenta propia, de 2-5 trabajadores, 6-10 trabajadores, de 11 a 50 trabajadores y más de 50 trabajadores, esta variable se incluirá en la regresión como un factor lo que implica una dummy para 4 categorías.

* Formal:

El tener un trabajo formal implicaría mayores ingresos para un individuo respecto a un individuo con trabajo informal, por lo tanto se incluye la variable “formal” para controlar por ese factor el ingreso laboral.

* Jefe del Hogar:

Esta variable se crea con base en la información reportada en la variable “p6050”, toma el valor de 1 cuando el individuo es jefe del hogar, se considera relevante incluirla en el análisis ya que se espera que las personas jefe de hogar trabajen más al tener como motivación mayores responsabilidades por el rol en la familia y esto puede generar impactos en el ingreso

* 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.

1. Partimos de una base de datos de la GEIH con 32177 observaciones, en la cual la variable de ingreso seleccionada “y\_total\_m” cuenta con 17309 NA.
2. Al realizar el filtro de personas mayores de 18 años y ocupados, se tiene un total de 16397 observaciones y para “y\_total\_m” se reduce la cantidad de NA a 1765.

Adicionalmente, se observa en la base que existen valores que se consideran no consistentes para el “y\_total\_m”, los cuales son $84 y $97, se decide convertir estos valores a NA, aumentando la cantidad de NA a 1767.

1. Se eliminan los 1767 valores NA encontrados porque se considera que para realizar el análisis de inferencia no se deben imputar datos pues pueden afectar las conclusiones.

Nota: El proceso anterior se encuentra en el archivo 2\_data\_cleaning.R incluido en la carpeta 2. Scripts.

* 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.

Tabla 2. Estadísticas descriptivas

| **Estadisticas** | **nbr.val** | **nbr.null** | **nbr.na** | **min** | **max** | **range** | **median** | **mean** | **std.dev** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **age** | 14.629,00 | - | - | 19,00 | 91,00 | 72,00 | 37,00 | 39,07 | 13,11 | |
| **female** | 14.629,00 | 7.695,00 | - | - | 1,00 | 1,00 | - | 0,47 | 0,50 | |
| **estrato1** | 14.629,00 | - | - | 1,00 | 6,00 | 5,00 | 2,00 | 2,52 | 0,99 | |
| **estrato\_medio** | 14.629,00 | 8.430,00 | - | - | 1,00 | 1,00 | - | 0,42 | 0,49 | |
| **estrato\_alto** | 14.629,00 | 14.039,00 | - | - | 1,00 | 1,00 | - | 0,04 | 0,20 | |
| **formal** | 14.629,00 | 5.766,00 | - | - | 1,00 | 1,00 | 1,00 | 0,61 | 0,49 | |
| **sizeFirm** | 14.629,00 | - | - | 1,00 | 5,00 | 4,00 | 4,00 | 3,22 | 1,65 | |
| **maxEducLevel** | 14.629,00 | - | - | 1,00 | 7,00 | 6,00 | 6,00 | 5,95 | 1,21 | |
| **edu** | 14.629,00 | 97,00 | - | - | 28,00 | 28,00 | 13,00 | 13,40 | 4,33 | |
| **oficio** | 14.629,00 | - | - | 1,00 | 99,00 | 98,00 | 45,00 | 50,17 | 28,07 | |
| **totalHoursWorked** | 14.629,00 | - | - | 1,00 | 130,00 | 129,00 | 48,00 | 47,65 | 15,12 | |
| **jefe\_hogar** | 14.629,00 | 7.522,00 | - | - | 1,00 | 1,00 | - | 0,49 | 0,50 | |
| **y\_total\_m** | 14.629,00 | - | - | 5.000,00 | 70.000.000,00 | 69.995.000,00 | 996.556,50 | 1.626.617,46 | 2.440.446,29 | |

De la tabla 2 se infiere que del total de 14.629 observaciones finales de individuos mayores de 18 años y ocupados, el 47% corresponde a mujeres el 53% restante a hombres con edad promedio de 39 años y 13 años de estudio en promedio. El 61% de los individuos cuentan con un trabajo formal.

Entre los estratos 1 y 2 se encuentra el 54% de los individuos, en los estratos 3 y 4 el 42%, el 4% entre los estratos 5 y 6, con un promedio de ingresos de $1.626.617.

Tabla 3. Ingreso por genero de acuerdo al nivel educativo

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **Hombre** | **Mujer** |
| **Nivel Educativo** | **Sin educación** | $ 698.217,10 | $ 434.099,90 |
| **Primaria** | $ 940.441,30 | $ 635.466,20 |
| **Media** | $ 1.033.321,30 | $ 663.663,70 |
| **Secundaria** | $ 1.172.380,10 | $ 881.530,90 |
| **Universitaria** | $ 2.822.667,60 | $ 2.359.281,90 |
|  | **Ingreso promedio Total** | $ 1.741.044,00 | $ 1.499.633,00 |
|  | **Años de educación promedio** | 13,14 | 13,68 |

|  |  |  |
| --- | --- | --- |
| **Nivel Educativo** | **Hombre (0)** | **Mujer (1)** |
| Sin educación | $ 698.217,10 | $ 434.099,90 |
| Primaria | $ 940.441,30 | $ 635.466,20 |
| Media | $ 1.033.321,30 | $ 663.663,70 |
| Secundaria | $ 1.172.380,10 | $ 881.530,90 |
| Universitaria | $ 2.822.667,60 | $ 2.359.281,90 |

La tabla 3 nos demuestra que la intuición de la brecha salarial entre hombres y mujeres es real, los ingresos promedio incrementan a medida que el nivel educativo es mayor, sin embargo, el incremento de los ingresos en los hombres es mayor al de las mujeres, existiendo esa brecha salariar por nivel educativo que lleva a que en promedio el ingreso de los hombres sea de $1.741.044 y el de las mujeres $1.499.633, brecha media que representa el 14% de los ingresos promedio. Esto, iría en contravía de los años de estudio de las mujeres, los cuales en promedio son más que los de los hombres.

Gráfica 1. xxxxxx

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Gráfica 2. Xxxxxx Gráfica 3. xxxxxx

Gráfico, Gráfico de dispersión

Descripción generada automáticamenteGráfico, Gráfico de dispersión

Descripción generada automáticamente

Gráfica 4. xxxxxx

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Gráfica 5. xxxxxx Gráfica 6. xxxxxx

Gráfico

Descripción generada automáticamente Gráfico, Gráfico en cascada

Descripción generada automáticamente

Gráfica 7. Xxxxxx Gráfica8. xxxxxx

Gráfico, Gráfico de dispersión

Descripción generada automáticamenteGráfico

Descripción generada automáticamenteEsquemático

Descripción generada automáticamente con confianza media

* + 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.

Gráfica 7. Xxxxxx

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?

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

Dependent variable:

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

logingtot

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

female -0.239\*\*\*

(0.015)

Constant 13.990\*\*\*

(0.010)

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

Observations 14,629

R2 0.018

Adjusted R2 0.018

Residual Std. Error 0.880 (df = 14627)

F Statistic 269.451\*\*\* (df = 1; 14627)

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

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

El modelo arroja que la variable que define el género “female” es relevante 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 bajo, lo que lleva a inferir que existen variables omitidas que también explican el ingreso de las personas, variables relacionadas en el punto 2 del presente problema-set.

De acuerdo con la tabla anterior, el ingreso de las mujeres se encuentra en promedio un 23% por debajo del de los hombres, ceteris paribus.

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

XXXX

Gráfica 7. Xxxxxx

Gráfico

Descripción generada automáticamente

* 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. La experiencia profesional no es tenida en el análisis dado que la GEIH no cuenta con información relacionada con los años trabajados en total sino con el tiempo que ha durado el individuo 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. Tendiendo en cuenta que para su cálculo se basa en los años de educación, podría generarse una correlación alta entre la experiencia profesional y los años de escolaridad, generando problemas de multicolinealidad, por esta razón consideramos que las variables a incluir edad y años de escolaridad, recogen la experiencia potencial por lo que no es necesario incluirla. [↑](#footnote-ref-1)
2. 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-2)
3. De acuerdo con la descripción de la variable en el archivo nacional de datos del DANE para la GEIH. [↑](#footnote-ref-3)
4. Estos impactos en la inferencia estadística se verán en detalle en un ejercicio adicional realizado para el punto 4 de análisis “The earnings GAP.” [↑](#footnote-ref-4)
5. Variable que refleja el grado escolar aprobado, esto, de acuerdo con la descripción de la variable en el archivo nacional de datos del DANE para la GEIH. [↑](#footnote-ref-5)
6. Se tiene como supuesto que las personas con máximo nivel educativo escolar en Preescolar realizaron dos años educativos. [↑](#footnote-ref-6)
7. No se incluye la variable estrato bajo correspondiente a los estratos 1 y 2 porque generaría multicolinealidad perfecta entre estas variables dummy creadas. [↑](#footnote-ref-7)