**Universidad de los Andes**

**Facultad de Economía**

**Big Data y Machine Learning para Economia Aplicada**

PhD. Ignacio Sarmiento-Barbieri

**Merit Tejeda:** 202210104

**Celin Hernández:** 202210067

**Estefania Laborde:**

**Problem Set 1: Predicting Income**

# 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]](#footnote-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://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 *“Medici´on de Pobreza Monetaria y Desigualdad Report”* that takes information from the [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 model of individual hourly wages

where 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) = Xβ.

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.

El propósito de este trabajo es encontrar dentro del conjunto de variables seleccionadas, cuales de ellas determinan o aportan al comportamiento del salario por hora. Para abordar la relación entre el salario por hora y sus variables explicativas seguiremos lo planteado por [Mincer (1974](https://link.springer.com/article/10.1023/B:REHO.0000004794.43782.55)), el cual es el paper seminal para el análisis de los ingresos y su distribución en el contexto del mercado laboral.

Siguiendo la propuesta minceriana, [Cardoso et all (2016)](https://www.bportugal.pt/sites/default/files/anexos/papers/re201606_e.pdf) analizan la brecha salarial entre hombres y mujeres para el caso de Portugal, concluyendo que la clasificación entre empresas y puestos de trabajo puede explicar alrededor de dos quintas partes de la brecha salarial de género. [Galassi & Andrada (2009)](https://www.aacademica.org/000-058/48.pdf), estiman la relación entre educación e ingresos por regiones geográficas de Argentina y encuentran que existe una relación inversa entre nivel de desarrollo y rendimiento de la educación. Adicionalmente, [Oxa & Laoyza (2017)](http://doi.org/10.5281/zenodo.1211809), estiman modelos de Mincer para Bolivia, evalúando el rol de capital en la determinación de los ingresos, estos autores señales que la tendencia a la baja de las tasas de retorno a la escolaridad podría explicarse por al aumento de la proporción de población con mayores años de educación formal, lo cual estaría dando una mayor homogeneidad educativa en la población dada la universalización de la educación formal.

Como fuente de datos se dispone de la Gran Encuesta Integrada de Hogares - GEIH 2018, la cual tiene como unos de sus principales objetivos la recopilación de datos del mercado laboral e ingresos de los hogares. Se utilizarán datos de la GEIH 2018 solo para la ciudad de Bogotá, por su robustez estadística en términos de cobertura geográfica, diseño metodológico y su disponibilidad de datos del mercado laboral, la GEIH es idónea para el presente trabajo. Los resultados encontrados siguen los postulados por la literatura económica, donde los rendimientos de la educación tienen un efecto positivo en los ingresos de los trabajadores y la edad y experiencia (lineal) aporta positivamente al ingreso, pero tienen un efecto decreciente en forma cuadrática. Adicionalmente, existe brechas salariales entre ingresos de hombre y mujer, cuyas estimaciones son estadísticamente significativas, pero no tienen mucha significancia económica, ya que la dummy de sexo, no tiene mayor efecto en la media del salario por hora.

* + 1. Data.[[2]](#footnote-2) We will use data for Bogot´a from the 2018 “Medición de Pobreza Monetaria y Desigualdad *Report”* that takes information from the [GEIH](https://www.dane.gov.co/index.php/estadisticas-por-tema/mercado-laboral/empleo-y-desempleo/geih-historicos).

The data set contains all individuals sampled in Bogota and is available at the following website [https://ignaciomsarmiento.github.io/GEIH2018 sample/](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:

* + - * 1. Describe the data briefly, including its purpose, and any other relevant informa tion.

Para estimar la relación entre el salario por hora y las variables explicativas, se utilizarán datos de la

Gran Encuesta Integrada de Hogares - GEIH 2018, implementada por Departamento Administrativo Nacional de Estadísticas de Colombia (DANE), la cual es una base de datos de corte transversal. La GEIH tiene como eje central la recopilación de datos de la estructura mercado laboral y los ingresos de los hogares, con una muestra total con cobertura a nivel nacional [(DANE, 2019).](https://microdatos.dane.gov.co/index.php/catalog/547) Para efectos de este taller, se utilizarán datos de la GEIH 2018, solo para la ciudad de Bogotá, que contiene variables originales y construidas, las que permiten identificar y describir características importantes de los individuos, empresas y sectores productivos en el ámbito del mercado de trabajo.

Por su robustez estadística en términos de cobertura geográfica, diseño metodológico y su especialidad en compilar datos del mercado laboral, la GEIH es idónea para estimar la relación entre el salario y el conjunto de variables explicativas. Es importante indicar que aun con las fortalezas de esta encuesta, la base presenta una serie de falencias, en vista de la presencia de muchos valores perdidos en algunas variables y datos atípicos en otras, por lo que previo a las estimaciones econométricas, es pertinente y necesario la implementación de un proceso de revisión y depuración de los datos, a fin de contar con información más consistente.

* + - * 1. Describe the process of acquiring the data and if there are any restrictions to accessing/scraping these data.

La obtención de la base de datos se realizó mediante un web scraping de tablas de una lista de urls y luego se combinaron todas estas tablas descargadas en un solo dataframe. Un supuesto para poder importar la base de datos, es que todas las urls proporcionadas contienen tablas con la misma estructura. En este caso [PS1 (ignaciomsarmiento.github.io)](https://ignaciomsarmiento.github.io/GEIH2018_sample/) contiene 10 tablas con las mismas dimensiones.

* + - * 1. Describe the data cleaning process and

Al importar la base de datos se cuenta con una tabla total que contiene 32,177 observaciones y 178 variables, previo al análisis econométrico, se realizó el siguiente proceso de filtrado, depuración, eliminación e imputación para los valores atípicos, el cual se describe a continuación:

* Excluir de la base de datos los individuos con edades menores a 18 años, con lo cual, la base de datos se redujo en un 25%.
* Eliminar los registros de los individuos que entran en la categoría de desempleados y población económicamente inactiva, alcanzando con esta exclusión un 50% de la base de datos originales.
* Se identificó la variable dependiente (y\_salary\_m\_hu), el salario - real por hora (habitual) - principal occ. (incluye propinas y comisiones), en la revisión de esta variable se encontró que un 40.32% de sus registros son valores perdidos (NA), los cuales fueron excluidos, con lo cual, se obtiene una base de datos de 9,785 observaciones, equivalente a un 30.4% de la base de datos original.
* De las 178 variables disponibles en el set de datos, se conservaron aquellas con un % de valores perdidos (NA) menor al 30%, este proceso dio como resultado 43 variables a analizar.
* Se construyeron dos (2) variables adicionales, remuneraciones extras en forma monetaria y en especie, a partir de la suma de un conjunto de beneficios en términos monetarios y en especie que recibe el trabajador; sin embargo, estas variables presentan una alta variabilidad, asociada a la presencia de muchos valores extremos; por tanto, serán excluidas del análisis de regresión.
* Se analizó e imputó valores para las siguientes variables:
  + *maxEducLevel:* Para los valores perdidos de esta variable, que representaban el 1% del total de las observaciones, se imputó el valor modal.
  + *totalHoursWorked:* Al revisar esta variable se identificó la presencia de valores extremos, el análisis en este caso fue el siguiente, de acuerdo con la normativa colombiana, la jornada de trabajo comprende 48 horas semanales, que a su vez coincide con el promedio de las horas trabajadas; en ese sentido, se calculó un factor umbral para el total de horas trabajadas, que es igual a la media de horas + 2 desviaciones estándar, esta es una regla empírica muy usada en estadísticas, donde se considera que los valores se encuentran dentro de una banda con dos veces la desviación típica respecto a la media, el valor obtenido se imputó para todos los valores por encima del umbral.
  + *p6426(experiencia):* En la revisión de esta variable se detectaron muchos valores extremos, en este sentido, se identificó en la base de datos que la edad máxima es 86, al cual le restamos 18, que son los años de inicio en el mercado laboral, se determinó que el máximo de años de experiencia laboral de un individuo puede alcanzar es 68, este valor fue imputado para todas las observaciones mayores a 68.
    - * 1. 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.

Se hará un análisis descriptivo de los ingresos reportados por los individuos en la GEIH 2018

**Tabla 1:** Estadísticas Descriptivas de los Ingresos Reportados en la GEIH 2018

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ============================================================================ | | | | | |
| Ingresos | N | Mean | St. Dev. | Mínimo | Máximo |
| ------------------------------------------------------------------------------------------------------------------------------------- | | | | | |
| Salario Mensual | 9,785.0 | 1,566,234.0 | 2,158,107.0 | 10,000.0 | 34,000,000.0 |
| Ingreso Monetario 1 | 9,785.0 | 1,739,064.0 | 2,314,670.0 | 0.0 | 40,000,000.0 |
| Ingreso Monetario 2 | 9,771.0 | 16,861.5 | 216,344.8 | 0.0 | 10,000,000.0 |
| Ingreso Total | 9,785.0 | 1,884,787.0 | 2,519,838.0 | 40,000.0 | 41,333,333.0 |
| Rem. Adicionales en Especie | 9,740.0 | 85,637.7 | 163,287.8 | 0.0 | 6,300,000.0 |
| Rem. Adicionales Monetarias | 9,767.0 | 1,139,646.0 | 3,285,939.0 | 0.0 | 130,000,000.0 |
| Salario por Hora Real | 9,785.0 | 7,984.3 | 11,629.4 | 151.9 | 291,666.7 |
| Ingreso Laboral Mensual | 9,785.0 | 1,757,077.0 | 2,413,729.0 | 30,000.0 | 60,100,000.0 |
| Ingreso Laboral por Hora | 9,785.0 | 8,868.2 | 12,917.1 | 326.7 | 350,583.3 |
| ------------------------------------------------------------------------------------------------------------------------------ | | | | | |

La GEIH 2018 reporta una serie de indicadores que reflejan la fuente de ingresos de los hogares, a excepción del salario por hora real, los demás ingresos están expresados en términos nominales, es decir, su valor describe la composición de efecto precios y volumen. Para obtener el salario por hora real, la encuesta contempla la construcción de un deflactor que permite ajustar el gasto y eliminar las diferencias atribuibles a los precios relativos, el valor de este índice de precios depende de una canasta de bienes que representa los gastos comunes de todos los dominios (DANE, 2019).

Las estadísticas descriptivas presentadas en la tabla 1, denotan una variabilidad alta en estos ingresos, atribuible a la presencia de valores extremos en los datos y a la heterogeneidad de los ingresos, según edad, sexo, estrato, nivel educativo, entre otros. Al observar los ingresos promedios, un aspecto a destacar es que todas las categorías de ingresos están por encima del salario mínimo, que para 2018 era de [$ 781,242.0](https://www.salariominimocolombia.net/2018), lo cual pueda reflejar un importante diferencial de ingresos si comparamos con otras regiones de Colombia, según datos del DANE, Bogotá tiene uno de los PIB per cápita más alto de Colombia[[3]](#footnote-3) y dado que solo estamos utilizando información de esta departamento, algunas de las conclusiones de este trabajo puedan carecer de validez externa, al no considerar la dinámica de ingresos de otras zonas del país.

La teoría del capital humano, desarrollada principalmente por Gary Becker (1964), nos indica una relación entre el nivel educativo e ingresos, los datos reportados en esta encuesta se pueden representar en la siguiente gráfica:



Tal y como se esperaría teóricamente, existe una relación directa y creciente entre el ingreso y la educación, la grafica 1 describe que el menor nivel de ingreso salarial por hora (en forma logarítmica) lo reciben los individuos que carecen de educación; en cambio, las personas con educación universitaria presentan los salarios más altos de Bogotá. Adicionalmente, es importante destacar que el cambio en el salario de secundaría completa a universitario es mayor que el cambio observado de primaria completa a secundaría completa, es decir, que un año adicional es educación universitaria, tiene un mayor efecto de un año adicional de educación secundario y primaria.

La siguiente tabla presenta la distribución porcentual de los ingresos, según los niveles educativos en Bogotá

**Tabla 2:** Estadísticas Descriptivas de Ingreso por Nivel Educativo

|  |  |  |  |
| --- | --- | --- | --- |
| ========================================================= | | | |
| Nivel Educativo | Salario por hora | Salario Mensual | Ingreso Total |
| Ninguno | 0.19 | 0.19 | 0.20 |
| Primaria Completa | 3.66 | 3.79 | 3.97 |
| Primaria Incompleta | 1.62 | 1.71 | 1.76 |
| Secundaria Completa | 19.41 | 20.63 | 20.95 |
| Secundaria Incompleta | 4.81 | 5.00 | 5.07 |
| Universitaria | 70.31 | 68.69 | 68.04 |

Del total de salarios por hora recibidos por los trabajadores en Bogotá, un 70.3% se concentró en individuos con educación universitaria, similar relación porcentual se observa en los salarios mensuales y en los ingresos totales; definitivamente, los datos indican con claridad y precisión que la educación y sus retornos juegan un rol crucial en el mejoramiento de los ingresos en Bogotá.

No obstante, para un análisis más detallado de los salarios e ingresos de los hogares en Bogotá, es pertinente observar otras dimensiones o incluir otras variables que nos un espectro mas amplio de los determinantes del ingreso, en tal sentido, la siguiente tabla presenta los mismos indicadores de ingreso, considerando otras variables adicionales y definiendo un rango de edades que nos ilustran algunas diferencias importantes.

**Tabla 3:** Ingresos por Rango de Edades, Educación, Estrato y Tamaño Empresas

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| ====================================================================================== | | | | | | | |
| Rango de Edad | Salario por hora | Salario Mensual | Ingreso Total | Nivel Educativo Medio | Estrato Predominante | Tamaño Empresas Predominante | Número de Individuos |
| 18-25 | 5,020.3 | 969,589.4 | 1,146,842.0 | 6 | 2 | 9 | 2,056 |
| 26-35 | 6,975.5 | 1,368,351.0 | 1,605,467.0 | 6 | 2 | 9 | 1,738 |
| 36-45 | 8,772.8 | 1,743,375.7 | 2,076,154.0 | 6 | 2 | 9 | 2,694 |
| 46-55 | 9,575.7 | 1,904,861.9 | 2,271,153.0 | 6 | 2 | 9 | 1,785 |
| 56-65 | 9,742.3 | 1,891,623.7 | 2,328,023.0 | 6 | 3 | 9 | 1,190 |
| 66-90 | 10,437.1 | 1,882,181.9 | 2,723,351.0 | 5 | 2 | 9 | 322 |

De la table 3 podemos observar que el ingreso es creciente en la edad; sin embargo, a medida que se incrementa la edad, el crecimiento del salario se vuelve de decreciente, tal y como se esperaría. Para relacionar el nivel educativo con los salarios y las edades, se calculó el promedio del nivel educativo para cada uno de los rangos de edades, en tal sentido, se observa que, en promedio, las personas en Bogotá se encuentran principalmente en la categoría 6, que los identifica con secundaria completa y predominan personas del estrato 2 y que trabajan en empresas grandes, de más de 100 trabajadores[[4]](#footnote-4).

Los estratos 2 y 3 se denomina bajo y medio-bajo, respectivamente y corresponden a segmentos poblacionales que son beneficiarios de subsidios en los servicios públicos domiciliarios[[5]](#footnote-5), la mayoría de personas encontradas en esta encuesta se ubican en esta escala, pero laboran en su mayoría en empresas, lo cual es un factor positivo que debería de incidir en mejores remuneraciones.

Otro aspecto a destacar de la tabla 3, es que el mayor % de los individuos se ubican en el rango de 36-45 años y estos

**Tabla 4:** Estadísticas Descriptivas de Ingreso por Sexo

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ===================================================================== | | | | |
| Sexo | Salario Real por Hora | Salario Nominal Mensual | Ingreso Laboral | Ingreso Total |
| Hombre | 8,178.4 | 1,651,645.1 | 9,042.9 | 1,845,881.8 |
| Mujer | 7,788.7 | 1,480,209.2 | 8,692.3 | 1,667,633.6 |

**Tabla 5:** Estadísticas Descriptivas Variables Seleccionadas

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| =============================================================================== | | | | | |
| Variables | N | Mean | St. Dev. | Mínimo | Máximo |
| ---------------------------------------------------------------------------------------------------------------------------- | | | | | |
| Horas trabajadas | 9,785 | 49.943 | 9.708 | 24 | 72 |
| Educación | 9,785 | 5.136 | 1.009 | 1 | 9 |
| Edad | 9,785 | 36.438 | 11.937 | 19 | 86 |
| Experiencia | 9,785 | 20.193 | 15.813 | 0 | 65 |
| Tamaño empresa | 9,785 | 6.439 | 2.872 | 1 | 9 |
| Estrato | 9,785 | 2.513 | 0.977 | 1 | 6 |
| -------------------------------------------------------------------------------------------------------------------------------- | | | | | |

* + 1. *Age-wage profile*. A great deal of evidence in Labor economics suggests that the typical worker’s age-wage 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 this subsection we are going to estimate the *Age-wage profile* for the individuals in this sample:

(2)

When presenting and discussing your results, include:

* A regression table.
* An interpretation of the coefficients and it’s significance.
* A discussion of the model’s in sample fit.
* A plot of the estimated age-earnings profile implied by the above equation. Including a discussion of the “peak age” with it’s respective confidence intervals. (Note: Use bootstrap to construct the confidence intervals.)
  + 1. *The gender earnings GAP*. Policymakers have long been concerned with the gender wage gap, and is going to be our focus in this subsection.

1. Begin by estimating and discussing the unconditional wage gap:

where *Female* is an indicator that takes one if the individual in the sample is identified as female.

1. *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 wage gap should exist. Estimate a conditional earnings gap incorporating control variables such as similar worker and job characteristics. In this section, estimate the conditional wage gap:
2. First, using FWL
3. Second, using FWL with boostrap. Compare the estimates and the standard errors.
4. Next, plot the predicted age-wage profile and estimate the implied “peak ages” with the respective confidence intervals by gender.

Para obtener la edad máxima, optimizamos derivando respecto a la Edad:

Dado que la variable dummy es: mujer = 1 y hombre =0; por tanto, se pueden obtener las siguientes edades máximas, según género:

De acuerdo con datos de la Gran Encuesta Integrada de Honduras – GIEH, Colombia, 2018, los resultados para la Ecuación 4 son los siguientes:

When presenting and discussing your results, include:

* A regression table, with the estimates side by side of the conditional and unconditional wage gaps, highlighting the coefficient of interest. ”Nuisance” controls, should not be included in the table but dutifully noted.[[6]](#footnote-6)
* An interpretation of the “Female” coefficients, a comparison between the models, and the in-sample fit.
* A discussion about the implied peak ages and their statistical similarity/difference.
* A thoughtful discussion about the unconditional and conditional wage gap, seeking to answer if the changes in the coefficient are evidence of a selection problem, a ”discrimination problem,” a mix, or none of these issues.
  + 1. *Predicting earnings. In the previous sections, you estimated some specifications with inference in mind. In this subsection, we will evaluate the predictive power of these specifications.*

1. Split the sample into two: a training (70%) and a testing (30%) sample. (Don’t forget to set a seed to achieve reproducibility. In R, for example you can use set. seed(10101), where 10101 is the seed.)
2. Report and compare the predictive performance in terms of the RMSE of all the previous specifications with at least five (5) additional specifications that explore non-linearities and complexity.
3. In your discussion of the results, comment:
   1. About the overall performance of the models.
   2. About the specification with the lowest prediction error.
   3. 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?
4. 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 and explore the potential links with the influence statis- tic. (Note: when attempting this subsection, the calculations can take a long time, depending on your coding skills, plan accordingly!)

# Additional Guidelines

* Turn a .pdf document in Bloque Neo´n.
* The document must include a link to your GitHub Repository.
  + The repository must follow the [template](https://github.com/ignaciomsarmiento/PS_Repo).
  + 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](https://project-awesome.org/) 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. I encourage you to follow the [tidyverse style guide](https://style.tidyverse.org/).

* 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.](https://www.aeaweb.org/journals/aer/styleguide)

1. See <https://www.irs.gov/newsroom/the-tax-gap> [↑](#footnote-ref-1)
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. [↑](#footnote-ref-2)
3. / <https://www.dane.gov.co/index.php/estadisticas-por-tema/cuentas-nacionales/cuentas-nacionales-departamentales> [↑](#footnote-ref-3)
4. / Para las variables estrato y tamaño de empresa de la tabla 3, se calculó la moda de sus datos. [↑](#footnote-ref-4)
5. <https://www.dane.gov.co/index.php/servicios-al-ciudadano/servicios-informacion/estratificacion-socioeconomica#preguntas-frecuentes> [↑](#footnote-ref-5)
6. Tip: Look how applied papers construct their results tables. These papers usually present comparable results in the same table with coefficients side by side, which helps the reader follow the discussion. [↑](#footnote-ref-6)