1. INTRODUCCCIÓN

En Colombia, de acuerdo con la Dirección de Impuestos y Aduanas Nacionales (DIAN) la evasión de impuestos es cercana a los $65 billones de pesos, estimaciones indican que la evasión esta cercana a 5,4 puntos del PIB, de los cuales 0,7 puntos corresponden evasión de impuestos de personas naturales. [[1]](#footnote-1) Algunos trabajos han estimado el subreporte de ingresos en 1.66 veces mas que el reporte inicial. (Rocha (2014)). Por lo anterior, usando metodologías aprendizaje estadístico, aun incipientes en economía, estimamos un modelo de predicción de salarios basado en información de la GEIH para el año 2018 en Bogota. Este modelo predictivo es potencialmente útil para encontrar casos de fraude fiscal y adicionalemente, apoyar la toma de decisiones de política para intervenir familias vulnerables.

1. DATOS

Usamos datos de la Gran Encuesta Integrada de Hogares GEIH en Bogota para el año 2018, la cual contiene información sobre las condiciones de empleo de las personas, características generales de la población como sexo, edad, estado civil y nivel educativo. Contiene también información sobre fuentes de ingreso.

El proceso de adquisición de la información se realizo mediante el scraping la información contenida en el sitio web <https://ignaciomsarmiento.github.io/GEIH2018_sample/>, sobre esta información no existe ningún tipo de restricción. El sitio web contiene información extraída y clasificada de la GEIH con variables originales y construidas que proporcionan datos relevantes para el análisis. Se realizo un web scrapping de los 10 bloques de información los cuales estaban particionados en diferentes URL´s. Estos bloques de información se consolidaron para terminar con una sola base de datos integrada de la información relevante.

* 1. *Limpieza de datos*

Considerando que la GEIH en lo concerniente a los procedimientos de conformación, depuración, imputación y empalme del ingreso contiene 137 variables con las características de los individuos pero sobre todo aproximadamente 40 variables que contienen información relacionada a ingreso, es necesario realizar un procedimiento de limpieza de la base en el cual se restringe el universo de datos en aquellos relevantes para los objetivos del trabajo, en particular usamos procedimientos para limpiar la base y dejar únicamente las siguientes variables:

1. *Sexo*; Variable dicótoma que toma valor 1 si es hombre y 0 si es mujer. Usamos esta variable con la finalidad de identificar brechas salariales entre hombres y mujeres.
2. *Edad;* relación laboral, educación, horas de trabajo por semana y Se realizo una depuración de información en la base de datos, la cual contenía información detallada de los individuos encuestados y una cantidad significativa de variables que contenían diferentes tipos de ingreso de los individuos por tanto se analizo la conveniencia de cada variable para el objetivo del estudio.

**4. The Gender Earning Gap**: Con el fin de estudiar la posible brecha salarial entre hombres y mujeres se realizaron las siguientes estimaciones:

**a.** **Brecha por género con modelo lineal simple**: Se utilizó una regresión lineal para estudiar si la variable dicotómica *Female* (Female=1 si es mujer) explica diferencias en el salario que gana una persona por hora. El modelo utilizado fué el siguiente:

Aunque el predictor explica solamente el 0.1% de la variación del logaritmo del salario por hora, la variable *Female* es significativa. Ser mujer disminuye el salario en 4.5%. El intervalo de confianza es (-0.072, -0.0175) por lo cual el descuento por ser mujer es significativamente diferente de cero. El resultado confirma que sí existe una brecha salarial por género.

Table

Description automatically generated



**b. Equal pay for equal work**: Existen diferentes variables que influyen en el salario que recibe una persona y aunque cada empleador ofrece un salario diferente, este no debería variar por el hecho de ser mujer u hombre. Las personas que realizan el mismo tipo de trabajo deberían recibir el mismo salario. Para estudiar si la brecha que se encontró en el apartado anterior se mantiene cuando se controla por tipo de trabajo y características similares del empleado se utilizó un modelo lineal en el que además de la variable *Female*, se incluyeron: *relab*=tipo o relación de trabajo, *maxEduLevel*=máximo nivel educativo, *tam\_empresa*=tamaño de la empresa, *edad* y *edad al cuadrado*.

Los siguientes son los resultados de la regresión utilizando MCO.

Table

Description automatically generated

Este modelo explica el 31.08% de la variación del logaritmo del salario. En este modelo *Female* continúa siendo significativa y afecta negativamente al logaritmo salario en un porcentaje mayor: 9.127%. Es decir que este modelo predice una brecha por genero aún más grande.

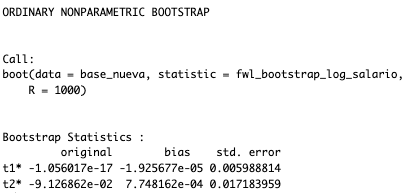
La regresión también se llevó a cabo siguiendo los pasos del teorema Frish-Waugh-Lovell (FWL). En este caso se utilizará *Female* como y *relab, maxLevelEduc, edad, edad\_sqr y tam\_empresa* como El teorema especifica tres pasos:

* Paso 1: Se realiza la regresión de la variable dependiente, en este caso log(w), respecto a De esta manera se obtiene la parte de *log(w)* que no explican las variables
* Paso 2: Se realiza la regresión de respecto a Se obtiene entonces la parte de que no explican las variables
* Paso 3: Se realiza la regresión de los residuales del paso 1 respecto a los residuales del paso 2. De esta manera se obtiene la parte del logaritmo del salario que sólo explica . El Teorema predice que el coeficiente de MCO de *Female* debe ser igual al coeficiente del paso 3. A continuación se muestran los resultados obtenidos por MCO y por FWL que confirman lo que pronostica el teorema:

Table

Description automatically generated

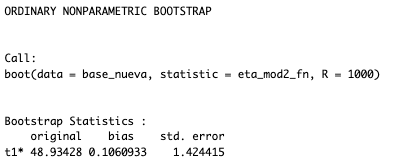
**b.ii.** FWL con Bootstrap: Se estimaron los coeficientes de FWL utilizando *bootstrap*:



Comparación de coeficientes de *Female* y errores:

|  |  |  |  |
| --- | --- | --- | --- |
|  | MCO | FWL | BOOTSTRAP |
| Coeficiente Female | -0.09127 | -0.09127 | -0.09126 |
| Errores | 0.012 | 0.012 | 0.0017 |

**4.c Predicted Wage**: A continuación se muestran dos gráficos de edad-salario. El primero toma en cuenta todo el rango de edad de la muestra. Se puede apreciar que existe una alta variabilidad después de los 50 años de edad que puede ser el motivo por el que se genera un alto *peak age* de 60.2 años. En el gráfico de la derecha se restringe la edad para personas entre 18 y 55 años. En ese rango el *peak age* es de 48.9 años con un CI (46.04, 51.58). El CI se calculó utilizando Bootstrap. Se utilizó el modelo lineal:



Text

Description automatically generated

Chart, scatter chart

Description automatically generatedChart, line chart

Description automatically generated

1. Ver:<https://www.larepublica.co/especiales/reforma-tributaria-2022/segun-la-dian-y-el-minhacienda-la-evasion-de-impuestos-es-cercana-a-80-billones-3422523> [↑](#footnote-ref-1)