**Taller 1.**

**Predicción de Ingresos.**

**Integrantes:** Isabella Mendez Pedraza. Cód.: 201814239

Manuela Ojeda Ojeda. Cód.:

Juan Sebastian Tellez Melo. Cód.: 201513710

Andres Mauricio Palacio Lugo. Cód.: 201618843

1. **Introducción.**

Para determinar el valor de los impuestos que debe pagar cada persona es fundamental la exactitud en la declaración de los ingresos.    Sin embargo, el fraude fiscal de todos los tipos siempre ha sido un problema que está muy presente. Menos del 90% de los impuestos son pagados voluntariamente y en el tiempo estimado en USA. La razón de esta brecha es que muchas personas no reportan correctamente sus ingresos.

Un modelo predictivo del ingreso podría ayudar a señalar casos de fraude para reducir la brecha. Además, podría ayudar a identificar personas en condiciones de vulnerabilidad que puedan necesitar un apoyo o ayuda adicional.

1. **Datos.**

La Gran Encuesta Integrada de Hogares (GEIH) del DANE es una encuesta mediante la que, además de las características generales de la población (como la edad, sexo, nivel educativo, entre otros), se obtiene información acerca de las condiciones de empleo de las personas (si están empleados, cantidad de horas que trabajan, si tienen seguridad social) y su nivel y fuentes de ingresos. Con esta información se generan indicadores del mercado laboral en Colombia como la tasa de ocupación, la rama laboral en la que se desempeñan los colombianos y la remuneración, y el comportamiento del mercado laboral para grupos poblacionales específicos como los jóvenes.

**Scrapping**

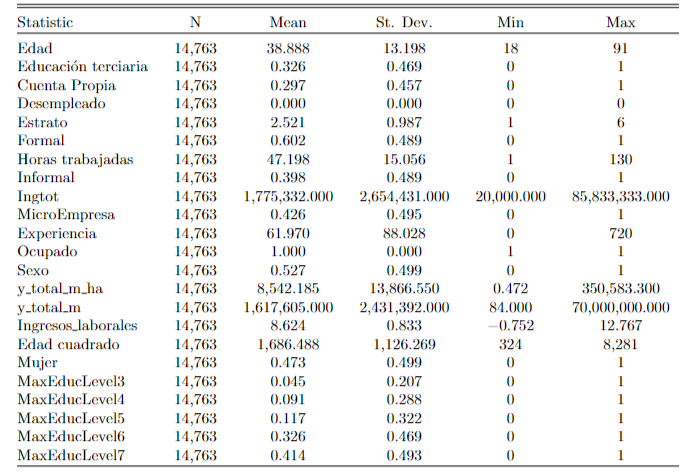
Para el análisis del presente documento, se usaron datos extraídos de la GEIH con una serie de modificaciones hechas por profesores de la Facultad de Economía de la Universidad de Los Andes. Estos datos fueron publicados en 10 páginas de internet diferentes por lo que fue necesario acceder a ellos mediante métodos de *web scraping*, utilizando el software estadístico de R y el código utilizado para ello se encuentra en el repositorio descrito al inicio del documento en el script denominado “*2\_Data\_cleaning*”. En este procesamiento de los datos desde las páginas de internet se identificó que los datos, aunque no tenían restricciones de seguridad, si se encontraban almacenados en objetos que requirieron revisar en detalle los recursos de red de cada página de internet. Una vez se logró obtener los datos, se creó un *loop* en R que permitió descargarlos y almacenarlos unidos en tables locales.

**Limpieza**

En un primer acercamiento con los datos optamos por tomar las variables que consideramos relevantes para el análisis, aquellas que pudieran explicar o relacionarse con el nivel de ingresos de las personas y también aquellas variables características de las personas o del contexto de la muestra. Iniciamos la limpieza eliminando las personas que son desempleadas o menores de 18 años. Para tener mayor claridad sobre los datos a manejar y depurar la base, calculamos el porcentaje de missing values por variable y eliminamos aquellas que tuvieran un porcentaje mayor al 50%. Luego, eliminamos los NA’s restantes. Finalmente, para dar inicio al análisis, generamos las variables que necesitábamos, como “Ingresos\_laborales” y definimos las varaibles catégoricas.

**Estadísticas descriptivas**

**Tabla 1.** Estadísticas descriptivas



 Como parte de la limpieza de datos, se restringió la información a sujetos con edades a partir de los 18 años y que estuviesen ocupados (utilizando la variable “ocu =1”). Así, llegamos a una muestra de 14,763 observaciones. De esta muestra, el 47.3% corresponde a mujeres mientras que el 52.7% son hombres.

La edad promedio de las personas es 39 años con una desviación estándar de 13 años y la persona ocupada con mayor edad tiene 91 años.

Gráfico, Gráfico de líneas, Histograma

Descripción generada automáticamente

En la figura 1 observamos la distribución por edad, allí se evidencia que solo tenemos personas mayores de edad y hay muy pocas personas que tienen más de 75 años. También se observa que las edades con mayor frecuencia están entre los 20 años y 30 años.

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Gráfico, Gráfico de líneas

Descripción generada automáticamente

En la figura 2 se observa la distribución de los salarios sin ningún tratamiento, sin embargo, en la figura 3 se evidencia cómo es la distribución del logaritmo del salario de los individuos, que es la variable que se utilizará en las estimaciones de este documento. Se observa que la media del logaritmo del salario está entre $8 y $9 mil pesos por hora.

Gráfico, Gráfico de líneas

Descripción generada automáticamente

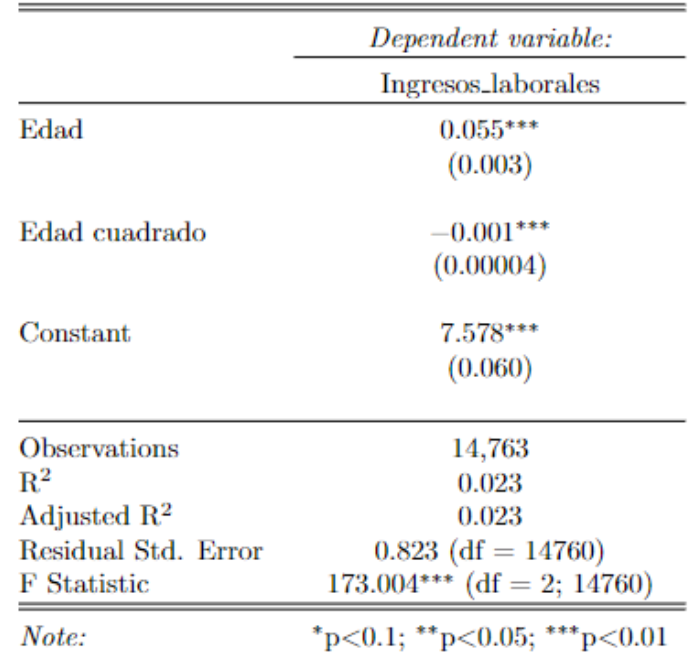
A partir de la tabla 1 y la figura 4 se observa que en promedio los individuos trabajan 47 horas habitualmente en la semana, la persona que menos trabaja es 1 hora y la que más hora trabaja llega a 130 horas. Los individuos trabajan con mayor frecuencia entre 40 y 50 horas habitualmente en la semana.

1. **Perfil Edad-Salario**

Para estimar el perfil edad-salario de los individuos vamos a estimar:

Tomamos nuestra variable Ingresos\_totales que es el logaritmo de y\_total\_m\_ha (ingresos asalariados + independientes total - nominal por hora). Al estimar el modelo encontramos que el incremento de 1 año de edad es asociado a un cambio en los ingresos de 5.5%. Observamos que nuestro modelo tiene un R2 de 0.02, es decir, que solamente el 20% de la varianza total de nuestro resultado objetivo está siendo explicado por nuestro modelo, por lo que podemos pensar que no es el mejor modelo.

**Tabla 2**. Perfil edad-salario



En la figura 5 observamos que como era de esperarse y según la evidencia de la economía laboral los salarios tienden a ser bajos cuando el trabajador es joven y aumentan a medida que el trabajador crece llegando a un máximo, en donde el salario empieza a disminuir.

Continuando con el modelo anterior y para la construcción del bootstrap utilizamos intervalos de confianza en los cuales encontramos que la edad mínima es 41.24 años y la máxima es 42.95 años. Adicionalmente, se observa que la edad máxima es 42.09 años. Es decir, que en promedio aproximadamente a los 42 años los individuos obtienen su logaritmo de salario máximo de aproximadamente $8.750 por hora con una confianza del 95%.

Gráfico, Gráfico de líneas

Descripción generada automáticamente

1. **Brecha salarial por género**

Al analizar la distribución de los salarios por género, encontramos que las mujeres obtienen, en promedio, el mayor salario hacia los 45 años, mientras que los hombres obtienen su mayor salario, promedio, hacia los 39 años. La figura 8, a continuación, muestra que las edades pico para cada caso (mujer-hombre) son estadísticamente diferentes, notando que con un intervalo de confianza del 95% estas dos estimaciones no se cruzan en la edad más alta en cada género.

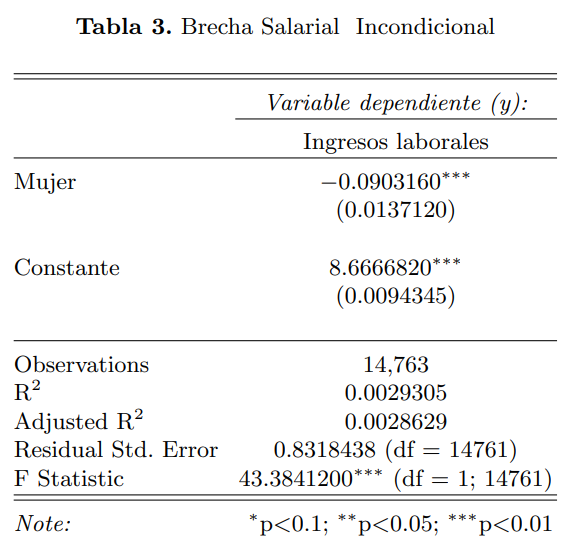
Gráfico

Descripción generada automáticamente

Ahora bien, para el análisis se utilizará el siguiente modelo y estimaremos el impacto del género sobre los ingresos laborales, utilizando el Teorema de FWL (Frisch-Waugh-Lovell ):

El análisis consistirá en introducir variables de control en este modelo y corroborar el efecto que tiene el género sobre los ingresos evaluados, dentro de estos controles se tienen las siguientes variables: Edad, persona independiente o no, es informal o no, máximo nivel de educación, si es una microempresa o no y el tiempo que lleva trabajando.

La regresión inicial consideró solamente la relación entre los ingresos laborales y el hecho de ser mujer o no (Modelo 1) y con ello encontramos un coeficiente de -0.09032, esto muestra que el hecho de ser mujer tiene un impacto negativo en los salarios, estimado en dicha magnitud.

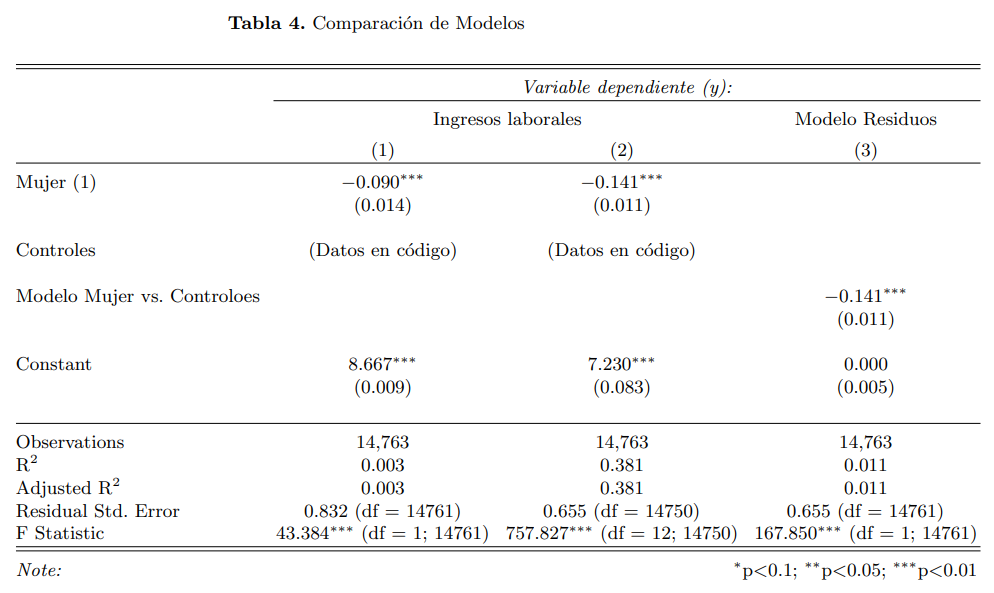


Para los análisis siguientes se estudiaron los siguientes modelos:

* Modelo 2. Relación de los ingresos laborales frente al hecho de ser mujer y las variables de control mencionadas previamente.
* Modelo 3. Para este modelo se realizó una regresión entre la variable que indica si se es mujer o no y las variables de control (R1), una regresión entre los ingresos laborales y las variables de control (R2) y, finalmente, estimamos la regresión de R2 en R1.

Cuando estimamos el Modelo 1 encontramos una relación entre los salarios y el hecho de ser mujer de -0.09032 y cuando incluimos variables de control (Modelo 2), ésta relación inicial pasa a -0.1413, esto ocurre porque dichas variables se encuentran correlacionadas. Por lo tanto, a partir del Teorema de FWL, podemos deducir que existe un efecto de las variables de control que se debe tener en cuenta al momento de analizar las predicciones de los salarios, pues la disminución sobre estos salarios no ocurre solamente por el hecho de ser mujer, aquí se puede evidenciar que las demás variables también generan efectos que deben ser limpiados de los modelos en el momento de realizar las predicciones, es decir, existe un efecto que no es explicado por la variable mujer sobre los salarios y que puede corresponder a variables no observadas. Nótese que al estimar el Modelo 2 encontramos un coeficiente exactamente igual al coeficiente que tendría la variable mujer cuando se regresiona únicamente contra todos los controles.

Ahora bien, teniendo en cuenta que al incorporar variables de control en los modelos el coeficiente que relaciona el hecho de ser mujer con los ingresos laborales aumenta, es posible inferir que esta muestra presenta un problema económico de selección y no de discrimación, pues cuando se utiliza solo la variable mujer sobre los salarios, la diferencia entre mujeres y hombres es más pequeña. Así, las diferencias salarias obedecen a otras variables que condicionan la selección y asignación de salarios.



Para corroborar estos datos se utilizó un *boostrapping* sobre los datos de la distribución de la muestra, con el fin de validar el coeficiente encontrado de relación de la variable cuando se es mujer y los controles frente a los ingresos estimados, con esto se estima la incertidumbre en dicho coeficiente.

El resultado del Bootstrap mostró un error de estimación de 0.01559 y un sesgo de 0.00060 y se identificó exactamente el mismo coeficiente estimado de los Modelos 2 y 3.

Los cálculos de este punto se encuentras en el script guardado en el repositorio como *4\_Gender\_Earnings.*

1. **Predicción de ingresos**