Taller 1: Predicting Income
Big data and Machine learning for Economics

Alison Gissell Ruiz Ruiz - Código 202116230
John Daniel Delgado Vargas - Código 202225721
José Julián Parra Montoya - Código 202213144

1. Introducción

La Gran Encuesta Integrada de Hogares (GEIH) es una encuesta recolectada por el DANE de manera mensual, que busca aportar insumos para el análisis del mercado laboral. En este trabajo se busca desarrollar modelos que permitan realizar inferencia y predecir el ingreso individual con base en características sociodemográficas halladas en esta encuesta. En este sentido se desarrolla un modelo que representa la relación entre las ganancias y la edad de las personas, así como la relación entre las ganancias y el género, con el fin de analizar la brecha salarial que podrían o no tener las personas de género femenino con respecto al masculino. Finalmente se elige, entre ocho especificaciones, la más precisa al momento de predecir las ganancias.

Para la estimación de los diferentes modelos se utiliza la GEIH recolectada para Bogotá en el año 2018, y restringida a individuos mayores de 18 años. Los datos empleados fueron obtenidos mediante *webscrapping*, método que facilita la extracción de información de páginas web. El código de obtención de información y el desarrollo de los algoritmos previamente descritos se desarrolla empleando *R* como lenguaje de programación.

2. Descripción de los datos

Para el análisis que se busca realizar, se requiere determinar una(s) variables(s) que mida(n) adecuadamente los *earnings* (ganancias) y el *income* (ingreso). En economía laboral ambas cosas son conceptualmente diferentes: mientras que las ganancias corresponden al salario o ingresos laborales, el ingreso se compone, además de los ingresos laborales, de ganancias de capital (Angrist, 1999). Para medir *earnings* se emplea el ingreso laboral (*y_ingLab_m* en la GEIH). De acuerdo a una exploración previa de los datos, esta variable se puede calcular de dos formas: en primer lugar, como la suma de los salarios percibidos por las ocupaciones primarias (*y_salary_m*) y secundarias (*y_salarySec_m*), beneficios laborales como auxilio de transporte (*y_auxilioTransp_m*), auxilio de vivienda (*y_vivienda_m*), viaticos (*y_viaticos_m*), salarios en especie (*y_especie_m*), bonificaciones (*y_bonificaciones_m*) y la proporción mensual de los diferentes tipos de prima.

Una alternativa para calcular *earnings* es mediante la suma del ingreso percibido por la ocupación primaria (*impa*) y secundaria (*isa*) pero excluyendo los montos mensuales de subsidios alimenticios (*y_auxilioAliment_m*) y subsidios por accidentes (*y_accidentes_m*), así

como las ganancias de los trabajadores cuenta propia (*y_gananciaNeta_m*). Esta variable es más apropiada que simplemente el salario (*y_salary_m*) para medir *earnings* ya que captura también los ingresos laborales de ocupaciones secundarias, cosa que no hace la primera. Para medir *income* se utilizará el ingreso total (*ingtot*). Esta variable, además de incluir el ingreso laboral, contempla las ganancias de los trabajadores cuenta propia, los ingresos por arrendamientos e intereses (*iof6*), jubilaciones y pensiones (*iof2*), transferencias monetarias (*iof3i* y *iof3h*), y el ingreso de los desocupados o inactivos (*imdi*).

Un hecho ampliamente documentado en las encuestas de hogares es que las características de los individuos explican el patrón de datos faltantes en las variables de ingreso (Restrepo, 2010). Esto indica que no es apropiado omitir los datos faltantes pues esto sesgaría la distribución de ingresos en los análisis posteriores. Restrepo (2010), encuentran que el método de imputación óptimo para la GEIH es el método Hot Deck. De acuerdo con la descripción metodológica del DANE, el proceso de imputación de las variables de ingreso ya fue realizado por este método, y estas variables se representan en la base con el sufijo *es* (*impaes* para la imputación de la variable *impa* o ingreso de la ocupación principal, etc.). En el cuadro 1 se presenta un conteo de los datos faltantes en las variables de Ingreso Laboral e Ingreso Total.

Cuadro 1: Datos faltantes

	Faltantes	Total
Ingreso Total	0	24568
Ingreso Laboral	14676	24568
Ingreso Laboral Imputado	4872	24568

Como puede observarse, no existen datos faltantes en el Ingreso Total. Esto se debe a que esta variable corresponde a la suma de las variables descritas previamente teniendo en cuenta, en cada caso, la versión de la variable donde fue necesario realizar la imputación, de manera que esta variable, al ser suma de variables imputadas, ya se encuentra imputada. En el caso del Ingreso Laboral, se observa que existen datos faltantes; no obstante, algunos de sus componentes como los ingresos de las ocupaciones primaria y secundaria fueron imputados por el DANE, de manera que se pueda reducir la cantidad de datos faltantes aprovechando esta imputación al sustraer las variables adecuadas (como la ganancia de los independientes, el auxilio por accidente y el auxilio alimenticio). En la última fila del cuadro 1 se muestra la variable con la corrección. Los demás datos faltantes consisten en casos donde el individuo es un trabajador por cuenta propia para el cual no es posible determinar qué parte de sus ganancias corresponde a ingresos laborales, quedando en total 4872 datos donde no es posible realizar imputación sin sesgar los datos.

El cuadro 2 contiene algunas estadísticas descriptivas para estas variables. Como puede observarse, tras utilizar la imputación del DANE para el Ingreso Laboral se obtiene una distribución a la izquierda de la distribución de Ingreso Total pues la media, la mediana y el percentil 90 son inferiores a los de esta última. Esto es de esperarse pues el Ingreso Total contiene al Ingreso Laboral junto a otros tipos de ingresos, lo que debería producir valores más grandes para cada individuo. De no utilizarse los valores imputados, como puede observarse en la tercera fila, se tendría una distribución de Ingreso Laboral a la derecha de la distribución de Ingreso Total, lo cual es conceptualmente incorrecto.

Cuadro 2: Estadísticas descriptivas variables dependientes

	Media	Mediana	D.E.	Percentil 90
Ingreso total	1369323.55	900000.00	2387364.35	2921931.33
Ingreso laboral	1745416.34	1032559.84	2403441.13	3250000.00
Ingreso Laboral Imputado	1008801.19	737717.00	2033436.79	2166666.75

Las variables que se utilizarán para predecir los ingresos descritos previamente son: la edad (*age*), la pertenencia al género femenino (*female*), la tenencia de educación universitaria (college_2), la condición de ser trabajador cuenta propia (*cuentaPropia*), la condición de informalidad (*informal*) y el oficio (*oficio*). En el cuadro 3 se presentan algunas estadísticas descriptivas. Este indica que la edad, que es un variable numérica, es en promedio de 42 años y que la distribución tiene un sesgo positivo pues la mediana (40 años) está a la izquierda de la media. Así mismo se observa que la edad más común es de 23 años, y que existe, en promedio, una desviación de 17 años respecto a la media. A diferencia de la edad, la condición de ser mujer, de tener educación universitaria, ser cuenta propia o trabajador informal son todas variables dicótomas. El promedio de las variables dicótomas indica la frecuencia con que la condición descrita por la variable se encuentra en la muestra. Por lo tanto, el 53 % de los individuos son mujeres, el 30 % tiene educación universitaria, el 21 % es trabajador por cuenta propia, y el 59 % es formal; estas frecuencias son consistentes con las modas de cada distribución. Finalmente, el oficio es una variable multinomial; la categoría ocupacional más común es la 45, que consiste en vendedores ambulantes o a domicilio. Se debe notar que la variable de educación universitaria es una variable nueva que debió calcularse debido a que la variable de educación universitaria existente en la base de datos (*college*) no refleja en realidad el nivel de educación superior si no el nivel de educación media.

Cuadro 3: Estadísticas descriptivas variables independientes

	Media	Mediana	Moda	D.E.
Edad	42.31	40.00	23.00	17.27
Mujer	0.53		1.00	
Universidad	0.40		0.00	
Cuenta Propia	0.21		0.00	
Formal	0.59		1.00	
Oficio		45.00		

En el cuadro 4 se presentan las 10 categorías ocupacionales más comunes. Estas incluyen vendedores ambulantes (oficio código 45), axiliares de oficina o bodega (39), conductor de vehículo (98), empleada doméstica (54) y cocinero o camarero (53). Algo común de estas profesiones es su alto grado de informalidad, lo cual es consistente con la frecuencia de informales que se encuentró previamente.

Cuadro 4: Oficios más comunes

Oficio	Conteo	%
45	1763	0.11
39	955	0.06
98	831	0.05
54	771	0.05
53	704	0.04
21	693	0.04
41	662	0.04
58	659	0.04
95	622	0.04
55	610	0.04

Ahora resulta interesante analizar la relación entre las variables dependientes y algunas de las variables independientes. En los cuadros 5 y 6 se realizan pruebas de diferencia de medias para ambos tipos de ingresos agrupados por algunas características de los individuos. Los hallazgos del cuadro 5 resultan sorprendentes: en las columnas 1 y 2 se puede observar que el ingreso total promedio para las mujeres es de $1'144.787 mientras que para los hombres es de $1'622,191, y la diferencia entre ambos promedios en todos los casos es estadísticamente significativa (columna 3). Algo similar sucede con los ingresos laborales, los cuales son en promedio de $830,259 para las mujeres y de $1'232,013 para los hombres; nuevamente la diferencia entre ambos promedios, para los dos tipos de ingresos, es estadísticamente significativa. En los individuos con estudios universitarios se observa el mismo patrón: tanto su ingreso total promedio ($2'210,500) como su ingreso laboral promedio ($1'701,006) es superior respecto a aquellos que no poseen estudios universitarios (818,004 y 526,196 respectivamente). La diferencia entre ambos promedios es nuevamente estadísticamente significativa.

Cuadro 5: Diferencia de medias variables independientes (A)

	Mujer			Universitario		
	Media Sí (1)	Media No (2)	Diferencia (3)	Media Sí (4)	Media No (5)	Diferencia (6)
Ingreso total	1144787.68	1622191.16	-477403.48***	2210500.86	818004.12	1392496.74***
	(1945185.64)	(2781507.32)	[30989.16]	(3462752.02)	(899429.22)	[35877.96]
Ingreso laboral	830259.7	1232013.76	-401754.06***	1701006.24	526196.2	1174810.04***
	(1829872.41)	(2242514.38)	[29673.55]	(2936768.86)	(661870.04)	[33221.98]

Nota: desviaciones estándar en paréntesis y erores estándar en paréntesis cuadrados. ***p<0.01, **p<0.05, *p<0.1.

En el cuadro 6 se observa el comportamiento de los ingresos según algunas características de los trabajadores. En la columna 1 se observa que el ingreso total de los trabajadores cuenta propia es de $1'386,842 y su ingresos laboral es de $1'077,734. Esto no es tan distinto al ingreso total promedio de los trabajadores que no son cuenta propia $1'364,727 y a su ingreso laboral promedio $1'006,204, como se observa en la columna 2. En efecto la diferencia entre ambos tipos de promedios no es estadísticamente significativa (columna 3). Una posible explicación de esto es que las características que explican los mayores ingresos vistas hasta ahora (como el género y el nivel educativo) estén distribuidas de forma homogénea entre ambos tipos de trabajadores. Este patrón no se sostiene para los informales, quienes tienen ingresos totales y laborales en promedio inferiores ($944,022 y $786,169, respectivamente) a los de los trabajadores no informales ($2'350,537 y $2'058,596); la diferencia entre ambos promedios, para ambos tipos de ingresos, es estadísticamente significativa, como es de esperarse.

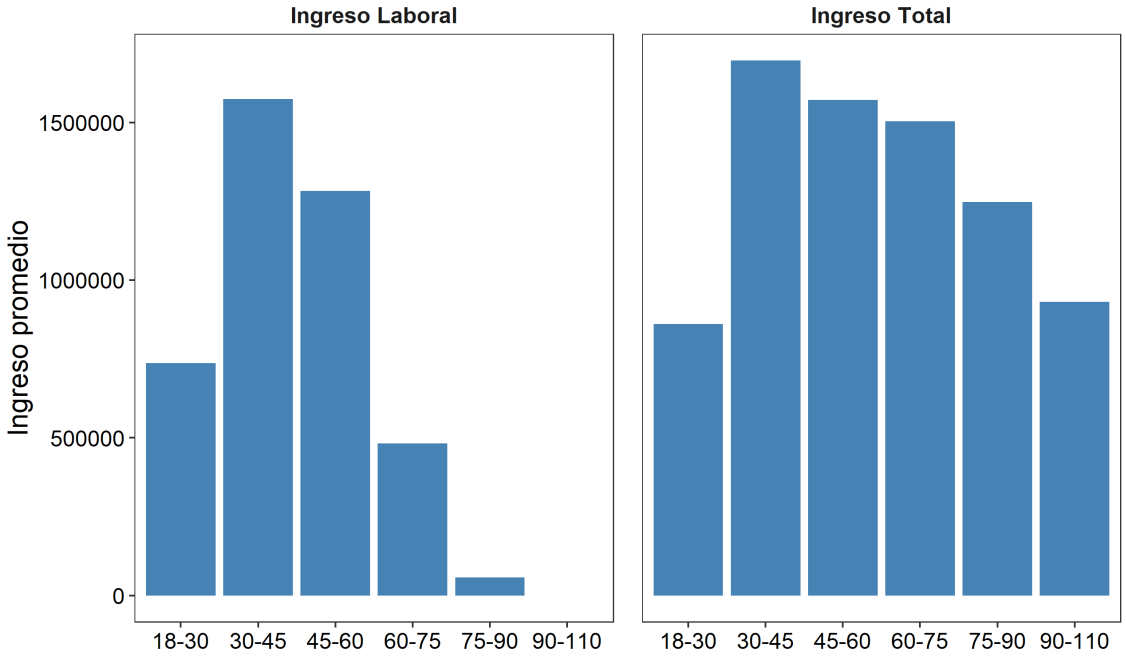
Cuadro 6: Diferencia de medias variables independientes (B)

	Cuenta Propia			Informal		
	Media Sí (1)	Media No (2)	Diferencia (3)	Media Sí (4)	Media No (5)	Diferencia (6)
Ingreso total	1386842.09	1364727.43	22114.66	944022.72	2350537.97	-1406515.25***
	(2054585.85)	(2467277.18)	[33756.84]	(1181637.68)	(3224865.87)	[35716.23]
Ingreso laboral	1077734.94	1006204.51	71530.43	786169.2	2058596.77	-1272427.57***
	(1573625.11)	(2048728.94)	[60699.94]	(1121024.92)	(2666218.2)	[35080.94]

Nota: desviaciones estándar en paréntesis y erores estándar en paréntesis cuadrados. ***p<0.01, **p<0.05, *p<0.1.

Finalmente, en la figura 1 se muestra el comportamiento de los ingresos laborales y totales promedio por intervalo de edad. Para el ingreso laboral se observa un patrón inicialmente creciente hasta los 45 años; posteriormente comienza a decrecer hasta que es nulo para los individuos mayores a 90 años. Este patrón se asemeja al encontrado para el ingreso total, aunque en este caso el ingreso total para los individuos mayores a 45 años decrece de forma mucho más suave. Esto se debe a los programas de seguridad social y transferencias monetarias a personas de la tercera edad, los cuales garantizan un nivel de ingreso mínimo en la vejez.

Figura 1: Ingresos por intervalo de edad



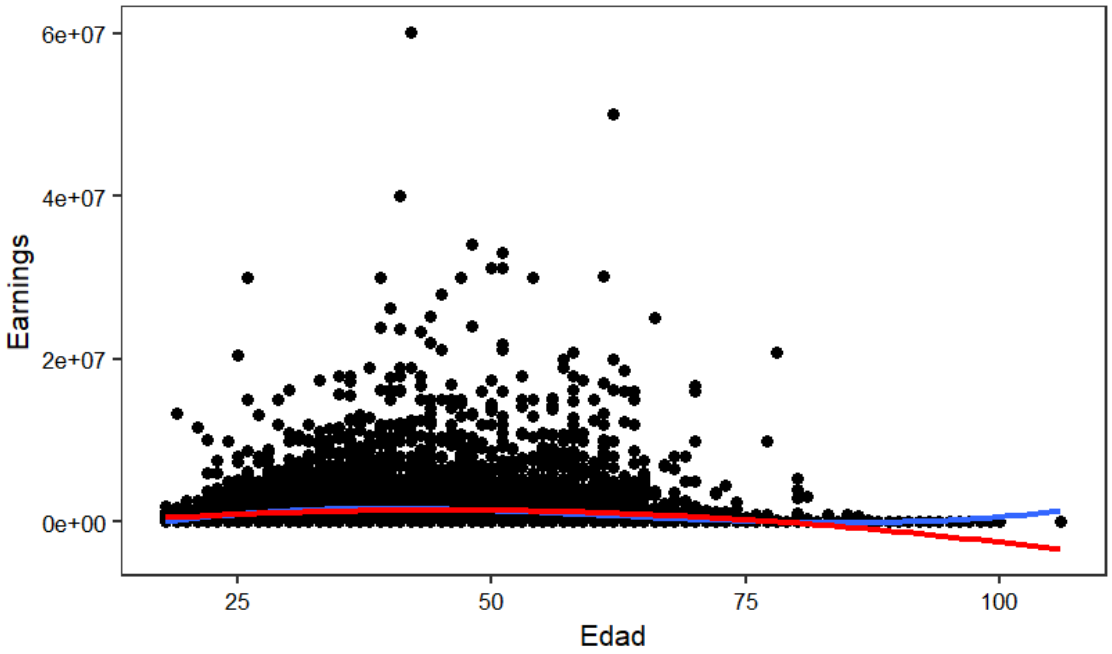
3. Análisis de relación edad-*earnings*

La hipótesis realizada sobre la relación entre estas variables es: "Los salarios tienden a ser bajos cuando el trabajador es joven; aumentan a medida que el trabajador envejece, alcanzando su punto máximo alrededor de los 50 años; y el salario tiende a permanecer estable o a disminuir ligeramente después de los 50 años", sin embargo, como se observa

en la figura 1, el comportamiento es diferente, ya que en nuestra muestra se observa que el punto máximo de ingresos no está alrededor de los 50 años sino mucho antes, en un rango entre los 30 y los 45 años, adicional a esto, no se mantienen estables luego de los 50 años; para el caso de Ingresos Totales, se podría tomar como cierto que disminuye ligeramente luego de esta edad, sin embargo no es un decremento constante, sino que aumenta su porcentaje con el paso de los años, por lo tanto se considera que la variación no es ligera. Para el caso del ingreso laboral se observa la disminución de ingresos de una forma mucho más agresiva, para la cuál tampoco se cumple la premisa.

En la figura 2 se puede observar la distribución de los datos, donde se evidencia que el valor máximo en ganancias se presenta antes de los 50 años, adicional a esto se observan dos regresiones en ella. En rojo, el resultado de la regresión sugerida con la fórmula $Earnings = \beta 1 + \beta 2Age + \beta 3Age^2 + u$, donde se puede observar que por la imputación y su gran cantidad de datos en 0 a edades cercanas a los 100 años tiende a un número negativo, por lo cuál se deduce que aún eliminando la linealidad agregando una variable cuadrática a la regresión, no es suficiente para representar de forma significativa a la distribución de los datos. Se observa además en azul una regresión aumentando el grado de complejidad (aumentando el grado de la regresión), de cuadrado a cúbico: $Earnings = \beta^1 + \beta 2Age + \beta 3Age^2 + \beta 4Age^3 + u$, obteniendo como resultado lo que se observa en la figura 2, en la cual se observa que tiene una forma muy similar a la anterior, salvo porque tiene un aumento cerca de los 100 años, lo cual no representa a la naturaleza de los datos (tampoco los números negativos). Se hicieron experimentos aumentando aún más el grado de complejidad pero el resultado es muy similar al cúbico.

Figura 2: Distribución de datos y regresión para relación edad-earnings

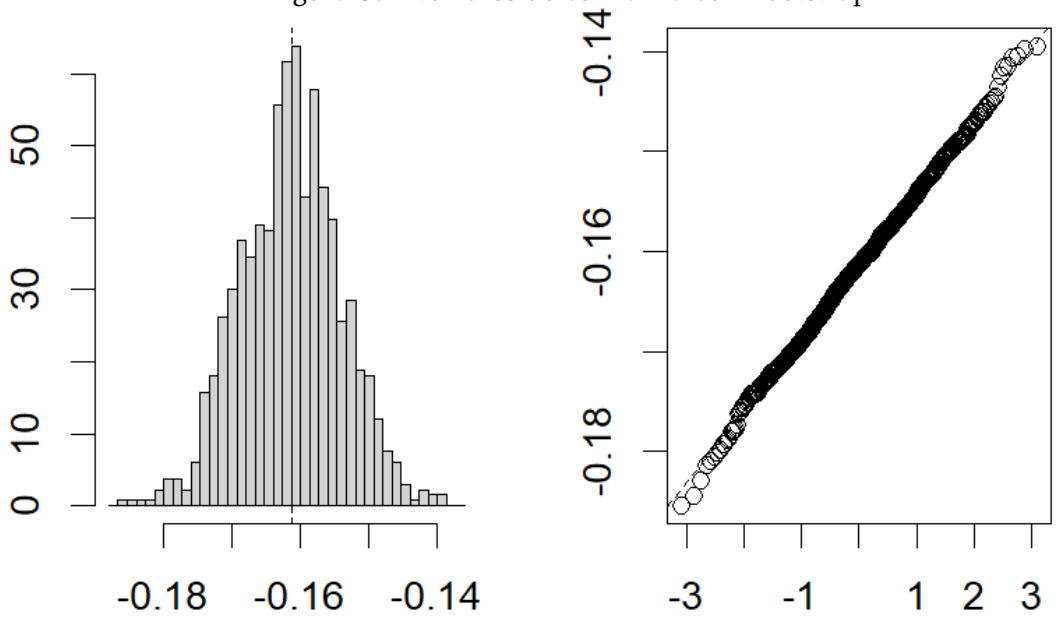


En general aumentar la complejidad solo modificó el comportamiento en datos superiores a 80 años, donde los datos en su mayoría son cero, de manera que la ecuación de regresión sugerida al cuadrado funciona de forma aceptable, se observan bastante más bajas que muchos datos, pero sucede debido a que hay muy pocas personas con ganancias

tan grandes, la mayoría de estos datos corresponden a earnings bajos.

Se realiza bootstrap 1000 veces, obteniendo la distribución que se observa en la figura 3, donde se realiza además el cálculo de los intervalos de confianza por medio de los métodos normal, básico y por percentiles, donde los resultados son en los 3 casos, con un nivel del 95 %, e intervalos de confianza pequeños, en un rango de (-0.17, -0.14), como se observa en la figura 3.

Figura 3: Intervalos de confianza con Bootstrap



4. Análisis de relación género-*earnings*

En esta sección se analiza como la variable de ingresos se ve explicada por el genero y si esta brecha se ve afectada por la edad y por los distintos tipos de oficios. El set de datos se utiliza la variable ingtot (ingreso total), considerando no solo el salario sino los diferentes ingresos que tiene las personas, esta es la variable objetivo a predecir. Inicialmente como variable dependiente para realizar la estimación del coeficiente se utiliza la variable factor Female, que indica con un 1 si el individuo es femenino.

Figura 4: Distribución de genero

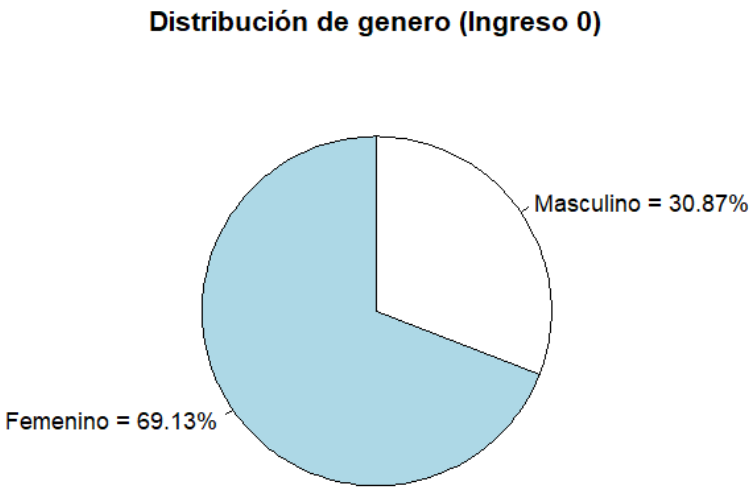
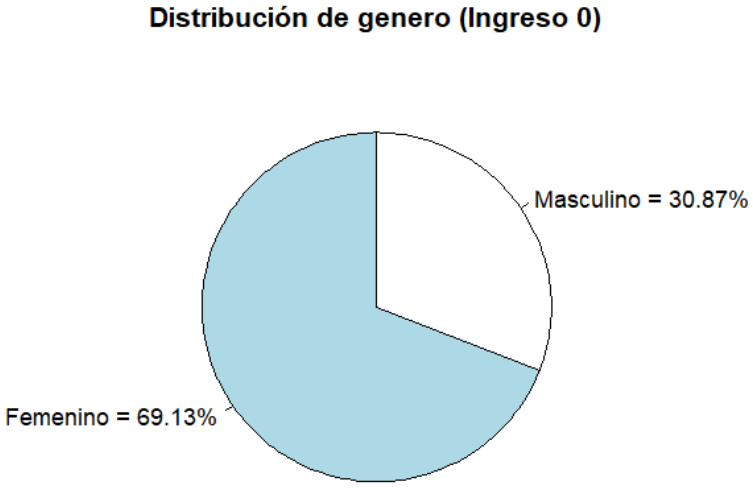


Figura 5: Distribución de genero, ingreso total 0



Los datos están distribuidos como muestra la figura 4, donde las mujeres representan un 53 % y Los hombres 47 %. No existen datos faltantes en las dos variables. Al

realizar la estimación se encontró que ser mujer afectaba en una reducción del -190 %, esto debido a que muchos de los salarios muy altos eran de hombres (ver figura 6) y se realizó el ajuste de la distribución (ver figura 7). La distribución de las personas con ingresos 0 (Figura 5) muestra que el 69 % de los registros con este nivel de ingresos son mujeres, por tanto se realizó una limpieza de outliers para correr nuevamente la estimación.

Figura 6: Distribución de Ingresos - outliers

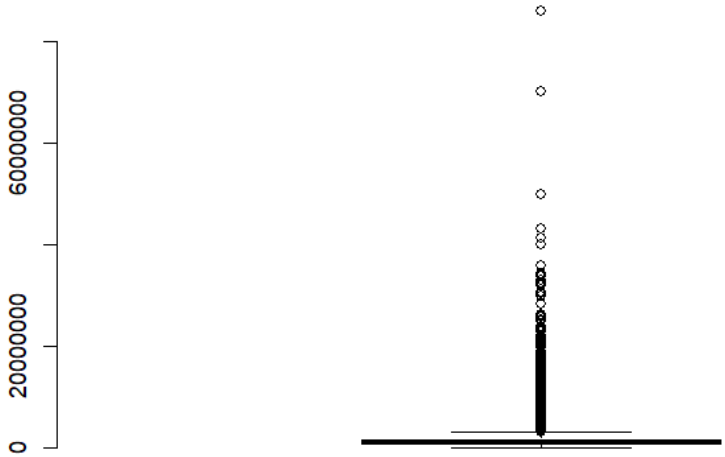
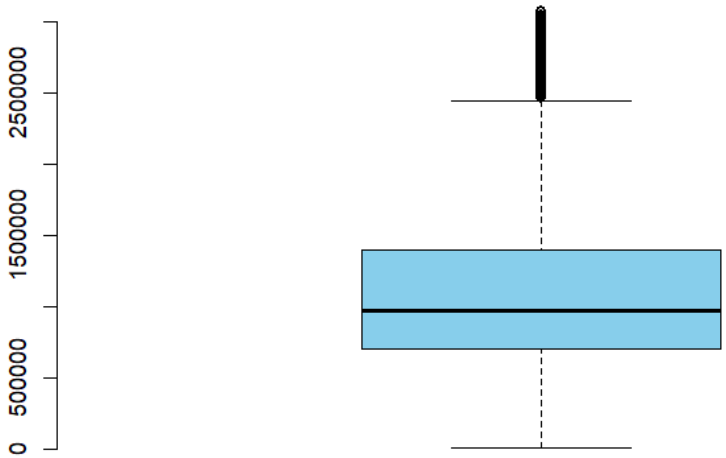


Figura 7: Distribución de Ingresos - sin outliers

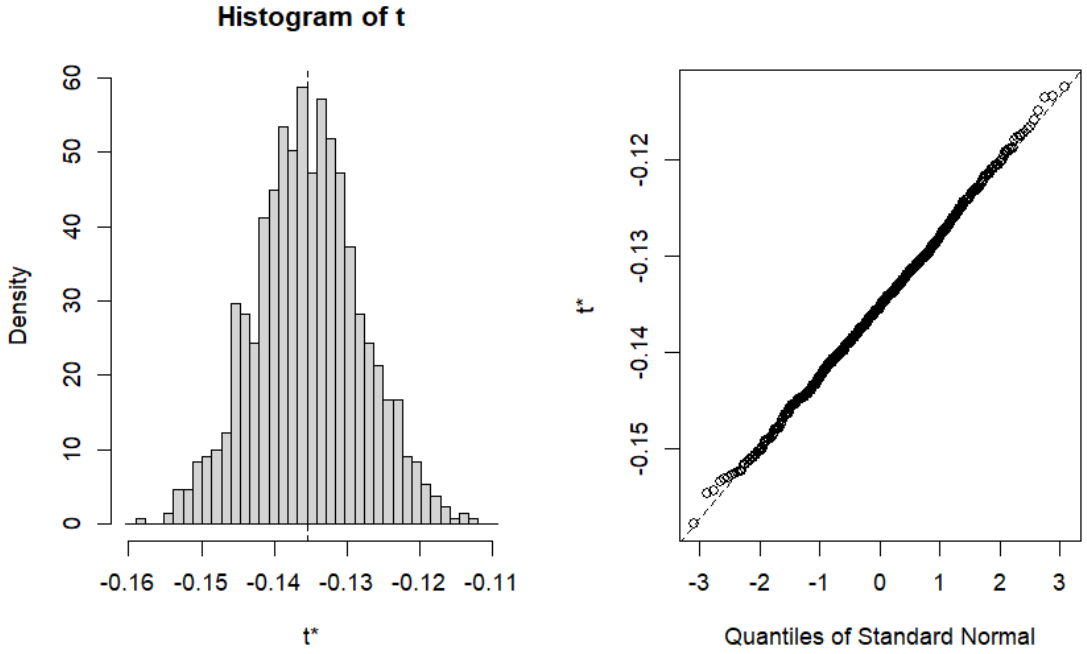


En la estimación con el nuevo set de datos se observa con una alta significancia (p value=0.00002) que las mujeres tienen un 23 % menos de ingresos. El poder predictivo del modelo incluyendo únicamente la variable genero es muy bajo, presentando un R2 de 0.025. Es ligeramente mejor utilizar el modelo que un promedio simple para predecir.

Al incluir en el modelo la variable de las edades, se puede observar como a medida que aumentan las edades también existe una reducción de los ingresos. Analizando en tres grupos las edades, de 18 a 30 años la brecha es de 11 %, de 30 a 50 la brecha es de 24 % y para personas mayores de 50 la brecha aumenta hasta llagar a 30 %. Es evidente como con la edad avanzan están brechas, pero también se evidencia como en las nuevas generaciones se ve una reducción paulatina de esta desigualdad.

Al incluir variables de los diferentes oficios se incrementa el R2 a 0.16, manteniéndose bajo. Un modelo lineal sigue siendo muy básico para estimar el ingreso total con esas variables, pero si se observa que para los diferentes oficios la brecha del genero es diferente.

Figura 8: Análisis con Bootstrap



5. Análisis de predicción para *earnings*

En esta sección se busca la especificación que permita generar la mejor predicción del ingreso laboral. En el cuadro 7 se muestran las diferentes especificaciones empleadas, las cuales utilizan las características de los individuos descritas en la primera sección.

Cuadro 7: Especificaciones para la predicción de *earnings*

Modelo	Especificación
1	$IngresoLaboral = \beta_0 + \beta_1\, Edad + \beta_2\, Edad^2 + e_1$
2	$Log(IngresoLaboral + 1) = \beta_0 + \beta_1\, Mujer + e_2$
3	$IngresoLaboral = \beta_0 + \beta_1\, Edad + \beta_2\, Edad^2 + \beta_3\, Mujer + e_3$
4	$IngresoLaboral = \beta_0 + \beta_1\, Edad + \beta_2\, Edad^2 + \beta_3\, Mujer + \beta_4\, CuentaPropia + e_4$
5	$IngresoLaboral = \beta_0 + \beta_1\, Edad + \beta_2\, Edad^2 + \beta_3\, Mujer + \beta_4\, CuentaPropia + \beta_5\, Universitario + e_5$
6	$IngresoLaboral = \beta_0 + \beta_1\, Edad + \beta_2\, Edad^2 + \beta_3\, Mujer + \beta_4\, CuentaPropia + \beta_5\, Universitario + \beta_6\, Formal + e_6$
7	$IngresoLaboral = \beta_0 + \beta_1\, Edad + \beta_2\, Edad^2 + \beta_3\, Mujer + \beta_4\, CuentaPropia + \beta_5\, Universitario + \beta_6\, Formal + \beta_7\, Formal * Universitario + e_7$
8	$IngresoLaboral = \beta_0 + \beta_1\, Edad + \beta_2\, Edad^2 + \beta_3\, Mujer + \beta_4\, CuentaPropia + \beta_5\, Universitario + \beta_6\, Formal + \beta_7\, Formal * Universitario + \alpha' Oficio + e_8$

Nota: α es un vector de parámetros que contiene los parámetros de cada una de las variables dicótomas que se pueden crear a partir de la variable multinomial de oficios. Así mismo, "Oficio" es un vector de variables dicótomas. Para la estimación de todas las especificaciones fue necesario descartar la categoría 78 de oficio a la cual solo correspondía un individuo en la muestra.

La Raíz Cuadrada del Error Cuadrático Medio (RMSE) de todas las especificaciones se reporta en el cuadro 8. Esta métrica de desempeño fue escogida pues penaliza de forma creciente los errores lo cual es deseable al estimar un ingreso (los errores de predicción más grandes son más problemáticos); así mismo, permite obtener los errores en las unidades de medida originales de la variable.

Cuadro 8: RMSE para diferentes especificaciones de *earnings*

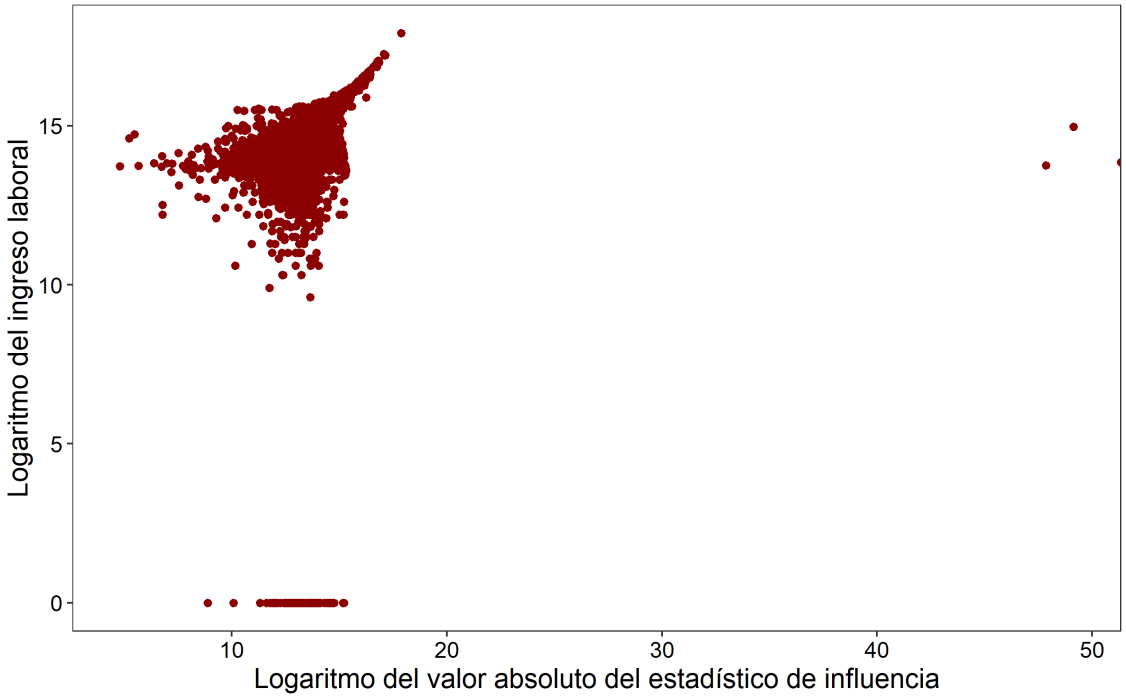
Modelo	RMSE
1	155183853
2	166030377
3	154976001
4	154390798
5	144029580
6	143075426
7	142393393
8	134801931

Como puede observarse, la mejor especificación es la dada por el modelo 8 pues esta arroja el RMSE más bajo. La segunda mejor es la dada por el modelo 7. Ahora se procederá a hacer un análisis de las observaciones más influyentes. Para esto se emplea el estadístico de influencia definido como:

$$\hat{\alpha} = \frac{\hat{u_j}}{1 - h_j}$$

Donde $\hat{u_j}$ corresponde a los residuales del modelo 8 obtenidos para la predicción de los datos de testeo, y h_j a la diagonal principal de la matriz de proyección para los datos de testeo. En la figura 9 se muestra el logaritmo del valor absoluto contra el logaritmo del ingreso laboral. Como puede observarse, no existe un patrón discernible entre ambas variables. Las observaciones de mayor influencia parecen ser valores atípicos relacionados a ingresos laborales altos sin ser los más altos de la muestra. Ignorando los valores atípicos, parecen haber valores similares del estadístico de influencia para valores muy diferentes del ingreso laboral.

Figura 9: Estadístico de influencia vs ingreso laboral



A continuación se realiza el ejercicio de estimar el RMSE de los dos mejores modelos utilizando Leave One Out Cross Validation (LOOCV), el cual consiste en estimar los modelos para cada subconjunto posible de los datos tales que una de las observaciones quede por fuera de la muestra de entrenamiento. En el cuadro 9 se reporta el RMSE de los modelos 7 y 8 del cuadro 8 calculado por este método.

Cuadro 9: RMSE por LOOCV

Modelo	RMSE LOOCV
7	234404932
8	219100926

Como puede observarse, el RMSE del modelo 8 es inferior al del modelo 7. No obstante, al compararlo con el RMSE del cuadro 8, se observa que, para ambos modelos, el RMSE calculado por LOOCV es mayor. Como el RMSE calculado por LOOCV tiene menor sesgo es posible afirmar que el RMSE calculado usando el subconjunto de prueba subestima el verdadero RMSE.

6. Repositorio

El repositorio se puede encontrar en el siguiente enlace:
`https://github.com/AlisonRuiz/Predicting-Income`

Referencias

Angrist, J. (1999). Empirical strategies in labor economics. In Ashenfelter, O., editor, *Handbook of labor economics*, pages 1277–1366. Elsevier.

Restrepo, M. I. y Marin, J. M. (2010). Imputación de ingresos en la gran encuesta integrada de hogares (geih) de 2010. *Revista Desarrollo y Sociedad*, (70):219–243.