**Problem Set 1**

**Autores:** Juan Sebastián Vásquez Acevedo y Walter Leonardo Sanchez Salazar
**Fecha inicio:** 11/06/2022
**Enlace a repositorio:** https://github.com/wsanch92/Problem_set_1_WS_JSV.git

Lo primero que se realizó fue la extracción de los datos. En este proceso se empleó un scraping de la página web https://ignaciomsarmiento.github.io/GEIH2018_sample/, en el cual no se presentó ninguna restricción en el acceso a la información. Para poder realizar este proceso se hizo uso del Software R y de las librerías: tidyverse y rvest.

En primera medida, se incluye en un objeto la URL del problem set, seguido se logra definir esta como un HTML. En el análisis de la pagina web se identifica el "XPATH" el cual brindará la ruta de los datos. Es así como se logra encontrar los nodos de los cuales se extraen los links que contienen las tablas que conforman la base de datos. Por medio de un tibble se unen tanto los nombres de los chunks como los links de interés para la extracción de los datos. En este proceso logramos identificar que el primer link de los chunks no contiene la tabla de la base de datos, por lo cual se hace necesario encontrar el link "verdadero" con el cual lograremos extraer la tabla que contiene los datos de interés. Luego de una exploración del código fuente de la pagina web se logra encontrar el link "verdadero" en el cual se identifica una particularidad, y es que en este link solo cambia el número o indicador del chunk, este número va del 1 al 10 y hace referencia al número de tablas que conforman la base de datos. Con este link "verdadero" logramos encontrar la puerta de entrada al proceso de extracción de los datos. Para lograr este proceso de extracción se hizo uso de un "loop".

El "loop" se encarga de entrar a la pagina y obtener el link "verdadero" para cada una de las tablas que conforman la base de datos que como se menciono anteriormente va del 1 al 10. Luego de ello extrae las tablas de las bases de datos y las convierte en un data frame, almacenándolas por medio una unión vertical en un data frame vacío previamente definido.

La base completa contiene 178 variables y 32,177 observaciones. Se hace un énfasis en la población mayor de 18 años y que manifiesta estar ocupada, esto lleva a que la base pase de 32,177 observaciones a 16,542. El enfoque en esta población se da debido al interés del presente estudio el cual busca crear un modelo que logre predecir el ingreso, es por ello que se seleccionan ciertas variables que según la literatura ayudan a explicar el ingreso individual. La teoría del capital humano resalta variables como el género, manifestando que a lo largo de diferentes análisis se ha logrado evidenciar que existe una diferencia salarial entre los hombres y las mujeres. Esto ha llevado a que investigadores y hacedores de política se vean motivados a crear estrategias que trabajen en disminuir la brecha salarial que existe entre los géneros. Otra variable es la raza, y al igual que el género, la literatura ha demostrado que existen diferencias salariales y que, evidentemente, se deben a la existencia de discriminación.

La edad es una variable que ha estado presente desde los inicios de los análisis de capital humano. En la ecuación de Mincer podemos evidenciar que se encuentra la edad como un determinante de los ingresos individuales. Esto se debe a que a medida que las personas van creciendo logran adquirir ciertas habilidades que les permiten ser más productivas. La edad guarda una relación con la experiencia laboral, pues su análisis es muy parecido. La experiencia les permite a las personas ser más atractivas en el mercado laboral y, como se mencionó con anterioridad, con el pasar del tiempo las personas adquieren un valor agregado que las hace más productivas, lo cual les permite acceder a mejores salarios. Hay que tener en cuenta que, tanto la edad como la experiencia, tienen rendimientos marginales decrecientes que también se deben analizar. Otras variables que la teoría resalta son: el tiempo trabajado, la actividad, ocupación y el nivel educativo, entre otras.

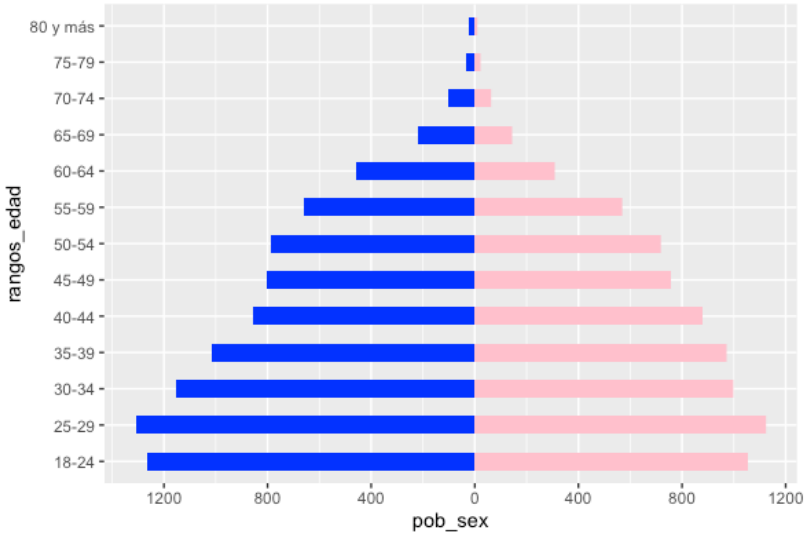
*Tabla 1 estadísticas descriptivas*

| Variables | Missing | Mean | sd | p0 | p25 | p50 | p75 | p100 |
|---|---|---|---|---|---|---|---|---|
| edad | 0 | 39.4 | 13.4 | 18 | 28 | 38 | 50 | 93 |
| ingreso | 0 | 1798186 | 2687702 | 15000 | 800000 | 1069453 | 1750703 | 85833333 |
| niveleduc | 0 | 5.01 | 1.10 | 1 | 4 | 5 | 6 | 9 |
| actividad | 0 | 1.51 | 1.37 | 1 | 1 | 1 | 1 | 6 |
| ocupacion | 0 | - | - | - | - | - | - | - |
| sex | 0 | 0.533 | 0.499 | - | - | - | - | - |
| tiempoempresa | 0 | 63.8 | 89.5 | 0 | 7 | 24 | 84 | 720 |
| oficio | 0 | 49.7 | 28.2 | 1 | 33 | 45 | 70 | 99 |

*Fuente: datos extraídos de la GEIH 2018 elaboración propia.*

Luego de conocer las variables que la teoría del capital humano ha demostrado que explican los ingresos individuales, elegimos cinco que desde la encuesta se pueden capturar, estás son: edad, nivel educativo (medido en años de educación), sexo, tiempo en la empresa (medido en meses) y ocupación (variable categórica). Al analizar el nivel educativo se puede observar que en promedio la población de la base tiene 5.01 años de educación y los datos se encuentran alejados de la muestra en 1.1 años de educación. El 75% de la muestra cuentan con 6 años de educación o menos, y los máximos años de educación reportados en la base corresponde a 9 años. De la variable sexo, podemos decir que el 53% de la muestra son hombres; respecto a tiempo en la empresa, podemos aclarar que se seleccionó esta como un proxy a la experiencia. En promedio, las personas llevan en la empresa 63.8 meses. Los datos se encuentran alejados de la muestra en 89.5 meses. El 50% de la muestra llevan menos de 24 meses trabajando. Por último, tenemos la ocupación, una variable categórica con ocho ocupaciones: obrero o empleado de empresa particular, obrero o empleado del gobierno, empleado doméstico, trabajador por cuenta propia, patrón o empleador, trabajador familiar sin remuneración, trabajador sin remuneración en empresas o negocios de otros hogares y otros. En cada una de ellas hay 9335, 632, 577, 5063, 620, 25, 15 y 10 observaciones, respectivamente.

*Gráfico 1. Pirámide poblacional*




Fuente: datos extraídos de la GEIH 2018 elaboración propia

La edad es una variable que en nuestra base no tiene datos faltantes, al igual que las otras variables que se seleccionaron. La edad promedio corresponde a 39.4 años con una desviación estándar de 13.4. Al observar la pirámide poblacional (gráfico 1) se evidencia que es más ancha en su base, lo que nos indica que el 50% de la población se encuentra en un rango entre los 38 y 18 años. Desde el rango de los 65-69 a la punta de la pirámide, podemos observar una reducción considerable de la población ocupada en los dos sexos, la edad máxima reportada en la base de datos corresponde a 93 años.

Ya se explicaron las variables independientes con las que se estructuraran los análisis que se van a ir presentando a lo largo del documento. Ahora se hablará de cómo se eligió el ingreso, que es la variable dependiente en el presente análisis. El ingreso total, es la variable que deseamos desde la encuesta poder capturar, esta debe incluir los ingresos laborales, los cuales están constituidos por las remuneraciones laborales o los honorarios, los subsidios al transporte, alimentación, entre otros, así como las ganancias adicionales producto del trabajo de horas extra o las primas y bonificaciones. También se deben incluir los ingresos por pensiones y por ganancias, que hacen referencia a los ingresos económicos que los hogares e individuos pueden obtener producto de arriendo, intereses de préstamos e inversiones. Dentro de los ingresos totales también se encuentran los ingresos de ayudas, en este grupo están los ingresos que reciben los hogares por parte de otros hogares que se encuentran dentro o fuera del país, así cómo los subsidios que reciben los hogares y/o personas por parte de instituciones tanto publicas como privadas. Por ultimo, están otros ingresos, estos hacen referencia a los rubros que se pueden obtener por diferentes razones a las que ya se mencionaron anteriormente, por ejemplo, los ingresos en especie.

Todos estos ingresos se pueden observar desde la encuesta GEIH, y hacen referencia a lo reportado tanto por los hogares como por las personas. Al momento de realizar análisis de pobreza y desigualdad en el país, el Departamento Administrativo Nacional de Estadística (DANE) ha explicado que es importante estudiar tanto los ingresos reportados como los imputados, por eso el DANE publica anualmente los diferentes grupos de ingresos que con anterioridad se mencionaron y crea el ingreso total. Esta variable contiene la suma de todos los ingresos que tiene el hogar tanto reportados como imputados, y lleva el nombre de "ingtot". Dentro de la base de datos se encuentran todas aquellas variables de ingresos observados e imputados que se han mencionado, por esto decidimos tomar como la variable de ingresos a "ingtot" ya que es el ingreso que el DANE utiliza para realizar los diferentes análisis en materia de ingresos. Al analizar esta variable de ingreso nos encontramos con unas observaciones que tienen como ingreso 0, y en ninguno de los ingresos, tanto imputados como observados, de las variables que el DANE utiliza para la construcción de la variable "ingtot", reportan alguno. Por consiguiente, hemos tomado la decisión de borrar estas observaciones lo cual nos lleva a quedarnos con una base de 16,277 observaciones.

Al momento de construir el primer modelo, que busca predecir los ingresos de las personas ocupadas mayores de 18 años en la ciudad de Bogotá D.C., se plateó la siguiente ecuación.

$$ingreso = \beta 0 + \beta 1 edad + \beta 2 edad^2 + \varepsilon$$

Tomando esta ecuación como base realizamos la construcción de 7 modelos en los cuales incluimos las variables que se explicaron en párrafos anteriores, realizando interacciones entre ellas logrando así encontrar el que sería nuestro modelo principal, el cual se define por la siguiente ecuación.

$$ingreso = \beta 0 + \beta 1 edad + \beta 2 edad^2 + \beta 3 niveleduc + \beta 4 tiempoempresa + \beta 5 tiempoempresa^2 + \varepsilon$$

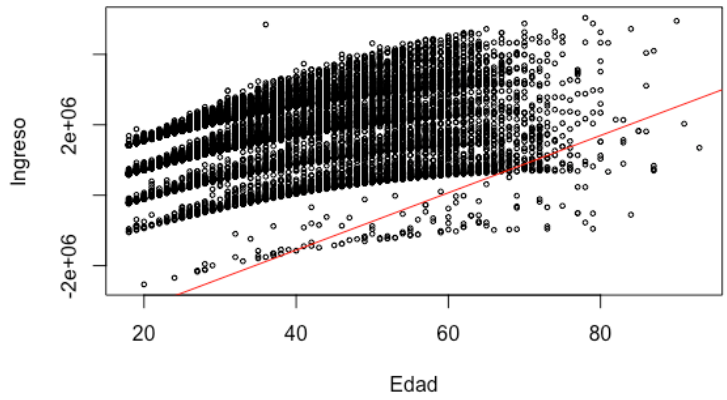
A continuación, se presenta una tabla con los coeficientes que se obtuvieron al correr los anteriores modelos presentados.

*Tabla 2 Primeros modelos de regresión*

| | *Dependent variable:* | |
|---|---|---|
| | ingreso | |
| | (1) | (2) |
| edad | 89,516.850*** | 81,327.340*** |
| | (9,078.271) | (8,843.928) |
| edad_sqr | -771.785*** | -545.215*** |
| | (105.215) | (102.393) |
| niveleduc | | 822,174.900*** |
| | | (19,129.560) |
| tiempoempresa | | 3,952.434*** |
| | | (582.102) |
| tiempoempresa_sqr | | -1.782 |
| | | (1.517) |
| Constant | -391,598.400** | -4,809,133.000*** |
| | (181,969.700) | (204,622.100) |
| Observations | 16,277 | 16,277 |
| $R^2$ | 0.018 | 0.132 |
| *Note:* | | *p<0.1; **p<0.05; ***p<0.01 |

Como se puede observar en la tabla 2, tanto el modelo uno (1) como el dos (2) cuentan con un total de 16,277 observaciones. En el modelo uno solo se incluyen dos variables de control, las cuales son la edad y la edad al cuadrado; los signos de estos coeficientes son los esperados según la literatura, los cuales nos indican que ante un aumento de un año en la edad los ingresos también lo hacen, sin dejar de lado los rendimientos marginales decrecientes, por tanto, se espera que el signo de la edad al cuadrado sea negativo. Según el modelo (1), un aumento de la edad genera un aumento en promedio del ingreso en 89,517 pesos con todo lo demás constante, mientras que en el modelo (2) el valor del coeficiente corresponde a 81,327 pesos, ceteris paribus. En el segundo modelo nos encontramos con un total de 5 variables explicativas, se mantiene la edad y la edad al cuadrado, variables que mantienen sus signos en sus coeficientes, pero el valor de estos disminuye respecto al modelo uno. Esto se debe a una reducción en el sesgo debido a que se incluyen más controles. Se agrega nivel educativo y el proxy de la experiencia. Podemos interpretar que un aumento de un año de educación genera un incremento en promedio de 822,175 pesos en los ingresos con todo lo demás constante. Al comparar los dos modelos, el modelo (2) cuenta con un R2 mayor, el cual indica que las variables de este modelo explican en un 13,2% nuestra variable de interés (ingreso). En cuanto a la significancia, tanto para el modelo (1) como del (2), casi todos los coeficientes son significativos a un nivel de confianza del 99% salvo el coeficiente de la variable tiempo en la empresa al cuadro en el modelo (2).

*Gráfico 2. Ingresos predichos por edad*




Fuente: elaboración propia.

Como se evidencia en el gráfico 2, no se observa un buen ajuste del modelo planteado (línea roja), esto se relaciona con el error promedio del modelo el cual corresponde a 6.271577e+1. Se estima el error estándar de los coeficientes de la regresión del modelo dos de la tabla 2 a través de Bootstrap, con los cuales se construye los intervalos de confianza, como se evidencia en la tabla 3. Se evidencia que la mayoría de los coeficientes de la regresión desde b0 hasta b4 caen dentro de los intervalos al 95% de confianza, salvo el b5 que no resulta significativo.

*Tabla 3. Errores estándar e intervalos de confianza por Bootstrap*

| Estadísticos del Bootstrap | | | | |
|---|---|---|---|---|
| Coeficientes | | Error estándar | Lim. Inferior | Lim. Superior |
| b0 | -4809133.0 | 227674.7286 | -5255375.468 | -4362890.532 |
| b1 | 81327.3 | 11721.5755 | 58353.05202 | 104301.628 |
| b2 | -545.2 | 147.69733 | -834.7015668 | -255.7280332 |
| b3 | 822174.9 | 23201.10642 | 776700.7314 | 867649.0686 |
| b4 | 3952.4 | 945.1997 | 2099.842588 | 5805.025412 |
| b5 | -1.8 | 3.19731 | -8.0489246 | 4.4845306 |

Fuente: elaboración propia.

Como se explicó con anterioridad, los análisis de ingreso se han enfocado en estudiar las diferencias prexistentes de ingreso entre los hombres y las mujeres. En el presente documento se estimaron modelos diferenciados por género para determinar si en la muestra estudiada existen brechas de ingresos entre los géneros como se presenta en la literatura. Igualmente, se identifica que la edad donde se maximiza el ingreso es de 41 años.

Se corrió un primer modelo, con la siguiente ecuación.

$$\log\_ingreso = \beta 0 + \beta 1 mujer + \varepsilon$$

*Tabla 4. Regresiones 1*

| | *Dependent variable:* |
|---|---|
| | log_ingreso |
| mujer | -0.193*** |
| | (0.014) |
| Constant | 14.064*** |
| | (0.009) |
| Observations | 16,277 |
| $R^2$ | 0.012 |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 |

Al analizar los coeficientes (tabla 4), nos encontramos con que el signo del coeficiente de la variable mujer es negativo, esto nos indica que una mujer gana en promedio 19% menos que un hombre si todo lo demás permanece constante, es decir, existen diferencias por género en los ingresos. Al ver el R2 podemos evidenciar que este es muy pequeño, es de 0.012, lo cual nos dice que nuestra variable independiente explica en un 0.1%la variabilidad del modelo. En cuanto a la evaluación del modelo en muestra se estima un error de predicción promedio de 0.7608782.

Luego de evidenciar que existe una diferencia entre los ingresos de los hombres y las mujeres, y que ser mujer disminuye los ingresos, desarrollamos un modelo que incluye un análisis por género. Para el desarrollo de este análisis, dividimos la base entre hombres y en mujeres; luego, se estimó el modelo inicial por género para el logaritmo del ingreso con nuestras variables de control ya elegidas, obteniendo los siguientes resultados:

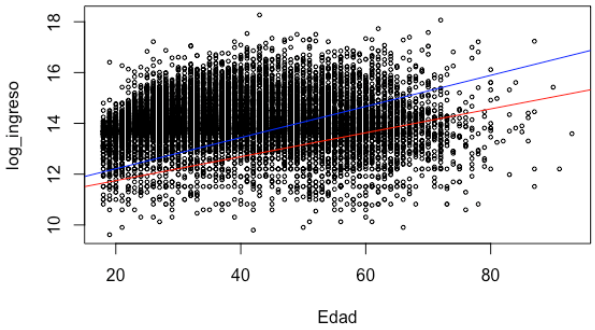
*Tabla 5. Regresiones del mejor modelo por género*

| | *Dependent variable:* | |
|---|---|---|
| | log_ingreso | |
| | (Mujer -1) | (Hombre -2) |
| edad | 0.047*** | 0.061*** |
| | (0.004) | (0.003) |
| edad_sqr | -0.001*** | -0.001*** |
| | (0.0001) | (0.00004) |
| niveleduc | 0.387*** | 0.321*** |
| | (0.009) | (0.007) |
| tiempoempresa | 0.003*** | 0.002*** |
| | (0.0003) | (0.0002) |
| tiempoempresa_sqr | -0.00000*** | -0.00000*** |
| | (0.00000) | (0.00000) |
| Constant | 10.804*** | 10.989*** |
| | (0.101) | (0.078) |
| Observations | 7,595 | 8,682 |
| $R^2$ | 0.234 | 0.234 |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 | |

En la tabla 5, el modelo uno corresponde a la base exclusiva para mujeres y el modelo dos para los hombres. La base de datos de las mujeres cuenta con un total de 7,595 observaciones y la de los hombres con 8,682. Es interesante comparar los coeficientes, y así evidenciar las variables que afectan en mayor medida tanto positiva como negativamente el ingreso de los hombres y el de las mujeres. Se interpretarán dos coeficientes significativos al 95% de confianza que llamaron la atención; el primero es la edad en el cual podemos ver que un año más de edad genera un incremento promedio del 6.1% en el ingreso de los hombres, manteniendo todo lo de más constante, mientras que en las mujeres este incremento es del 4.7%. El nivel educativo nos muestra que un año más de educación genera un incremento en el ingreso de las mujeres de 38.7%, en promedio, manteniendo todo lo demás constante; mientras que en los hombres un año más de educación brinda un incremento en el salario en promedio del 32.1% manteniendo todo lo demás constante, a un nivel de confianza del 95%. Para estos dos modelos, el coeficiente de variación obtenido es de 23,4% para ambos.

Así mismo, se encuentra que el intercepto y la pendiente en los modelos divididos por sexo no es la misma. Buscando ser más visuales, se presenta el gráfico 3, en el cual podemos ver en los puntos los ingresos por edades tanto de los hombres como de las mujeres. La línea roja hace referencia al modelo de las mujeres y la línea azul al de los hombres, el gráfico es claro al mostrar que tanto la pendiente como la intercepción con el eje “del ingreso” no es el mismo entre los dos modelos.

*Gráfico 3. Logaritmo del salario por edad y sexo*




Fuente: elaboración propia.

Se evidencia que la edad promedio en donde las mujeres alcanzas el máximo es de 41 años, mientras que los hombres lo obtienen a los 50 años. Además, al calcular los errores estándar con la técnica de Bootstrap para construir los intervalos de confianza al 95% de significancia diferenciados por género, se identifica que todos coeficientes para ambos géneros caen dentro de los intervalos; es de resaltar que se observa una sobreposición de los intervalos entre hombres y mujeres (Tabla 6 y 7).

*Tabla 6. Errores estándar e intervalos de confianza por Bootstrap: mujeres*

| Estadísticos del Bootstrap. Modelo mujeres | | | | |
|---|---|---|---|---|
| Coeficientes | | Error estándar | Lim. Inferior | Lim. Superior |
| b0 | 10.8044 | 0.1090924 | 10.5905789 | 11.0182211 |
| b1 | 0.04705803 | 0.00477229 | 0.03770435 | 0.05641171 |
| b2 | -0.0005172 | 5.8337E-05 | -0.0006315 | -0.0004028 |
| b3 | 0.3869648 | 0.01072128 | 0.36595109 | 0.40797851 |
| b4 | 0.00296669 | 0.00037047 | 0.00224057 | 0.00369281 |
| b5 | -3.538E-06 | 1.1563E-06 | -5.804E-06 | -1.271E-06 |

*Tabla 7. Errores estándar e intervalos de confianza por Bootstrap: hombres*

| Estadísticos del Bootstrap: modelo hombres | | | | |
|---|---|---|---|---|
| Coeficientes | | Error estándar | Lim. Inferior | Lim. Superior |
| b0 | 10.98856 | 0.09119838 | 10.8098112 | 11.1673088 |
| b1 | 0.06126898 | 0.00412515 | 0.0531837 | 0.06935426 |
| b2 | -0.0005889 | 4.9916E-05 | -0.0006867 | -0.0004911 |
| b3 | 0.3207861 | 0.00840337 | 0.3043155 | 0.3372567 |
| b4 | 0.00198248 | 0.00027449 | 0.00144448 | 0.00252048 |
| b5 | -2.645E-06 | 7.7665E-07 | -4.167E-06 | -1.123E-06 |

La brecha de ingresos que existe entre los géneros ha sido muy estudiada, esto ha llevado a la creación de teorías como el techo de cristal y los pisos pegajosos, las cuales evidencia que no solamente hay diferencias en los ingresos si no que también, en muchas ocasiones, para las mujeres, hay mayor dificultad en avanzar y llegar obtener mejores ingresos. Esto ha llevado también a identificar que existen sectores de la economía y oficios que se han masculinizado y, en los cuales, para las mujeres, es más difícil entrar a participar.
Buscando controlar por estas diferencias que pueden existir entre sectores desarrollamos el siguiente modelo, agregando controles por ocupación de las personas:

$$ingreso = \beta 0 + \beta 1 mujer + \beta 2 ocupación + \varepsilon$$

*Tabla 8. Regresiones con ocupación en el modelo de logaritmo del ingreso*

| | *Dependent variable:* |
|---|---|
| | log_ingreso |
| mujer | -0.169*** |
| | (0.013) |
| as.factor(ocupacion)2 | 0.741*** |
| | (0.034) |
| as.factor(ocupacion)3 | -0.448*** |
| | (0.036) |
| as.factor(ocupacion)4 | -0.452*** |
| | (0.014) |
| as.factor(ocupacion)5 | 0.356*** |
| | (0.034) |
| as.factor(ocupacion)6 | -1.453*** |
| | (0.164) |
| as.factor(ocupacion)7 | -0.698*** |
| | (0.212) |
| as.factor(ocupacion)9 | -1.536*** |
| | (0.260) |
| Constant | 14.170*** |
| | (0.010) |
| Observations | 16,277 |
| $R^2$ | 0.126 |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 |

Este modelo cuenta con 16,277 observaciones, el coeficiente de mujer es negativo y nos indica que una mujer, en promedio, tiene un ingreso 16.9% menor que un hombre si todo lo demás permanece constante, siendo este coeficiente significativo al 95% de confianza. Podemos observar que, al controlar por la ocupación de los individuos, la brecha de ingresos se reduce de 19.3% a 16.9% pero no hay evidencia de que sea un problema de selección debido a que sigue existiendo una diferencia en el salario entre las mujeres y los hombres aun controlando por ocupación. Al analizar por ocupaciones podemos evidenciar que tan solo las ocupaciones de obrero o empleado del gobierno y patrón y empleador presentan ingresos mayores al compararlas con obrero o empleado de empresa particular. El R2 de este modelo es de 0.126, lo que nos indica que nuestro modelo explica la varianza del ingreso en un 12.6%.

Ahora, se busca comprobar el teorema Frisch-Waught-Lovell (FWL) el cual estable que para una observación j cualquiera de interés, se estima el modelo original de mujer agregando una variable dummy para esa observación y el resultado del coeficiente de interés puede obtenerse y será igual al coeficiente de una regresión de los residuales de este modelo contra los residuales de un modelo alterno con la dummy como variable dependiente contra la variable independiente de mujer. Para corroborarlo, tomamos una observación cualquiera de nuestra base y estimamos los modelos referidos (modelos 2 y 3 de la tabla 9) encontrando que, el coeficiente de mujer de -0.194, es igual para ambos modelos.

*Tabla 9. Prueba del Teorema Frisch-Waught-Lovell*

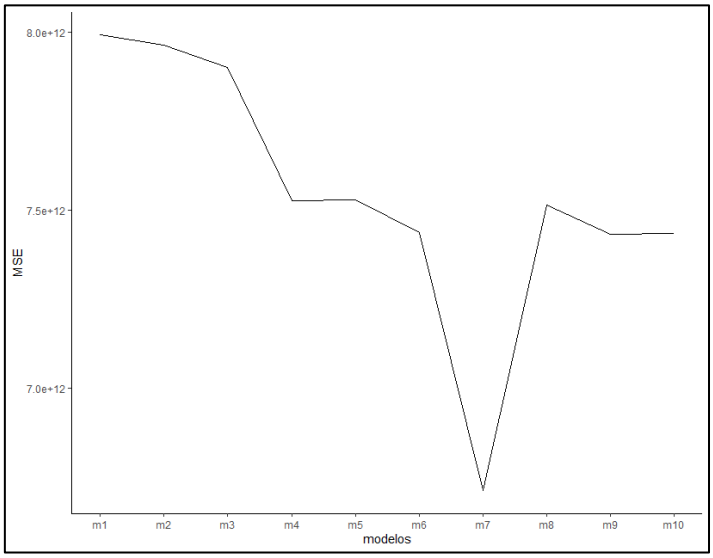
| | *Dependent variable:* | | |
|---|---|---|---|
| | log_ingreso | | res_y_e |
| | (1) | (2) | (3) |
| mujer | -0.193*** | -0.194*** | |
| | (0.014) | (0.014) | |
| ej | | 1.866** | |
| | | (0.872) | |
| res_x_e | | | -0.194*** |
| | | | (0.014) |
| Constant | 14.064*** | 14.064*** | 0.000 |
| | (0.009) | (0.009) | (0.007) |
| Observations | 16,277 | 16,277 | 16,277 |
| $R^2$ | 0.012 | 0.012 | 0.012 |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 | | |

Predicción

Para la predicción de las ganancias o de los ingresos se divide la muestra en un 70% de entrenamiento y un 30% de prueba o testeo. A partir de esto, se entrenan los modelos previamente construidos donde la variable dependiente puede ser es el ingreso o el logaritmo del ingreso. Se realizan las predicciones de los modelos en la muestra de testeo con el fin de calcular el promedio de los errores al cuadrado; como criterio de identificación entre las diferentes variaciones y complejidades de los modelos.

Se toma como base el modelo de regresión lineal entre el logaritmo del ingreso y la constante, para ser comparado con otros 9 modelos de más, donde se incluyen los modelos iniciales ya mencionados y otros con transformaciones que buscan darle complejidad a la estructura del modelo. A continuación, se muestra el promedio de los errores de predicción en la muestra de prueba de los modelos estimados (gráfica 4):

*Gráfico 4. Promedio de errores de predicción*




Fuente: elaboración propia.

Como se observa en la gráfica 4, el modelo con mejor desempeño, es decir, que presenta un promedio de error de predicción en la muestra de testeo menor, es el modelo m7, que corresponde al inicialmente elegido que estima el ingreso en niveles contra la edad, la edad al cuadrado, el nivel educativo, el tiempo de la empresa y el tiempo de la empresa al cuadrado.

Para este modelo m7, se calculó el estadístico Laverage para cada observación en la muestra de prueba. Al inspeccionar la muestra de prueba ordenada por este estadístico, se evidencia datos atípicos de ingreso con un alto valor de Laverage. Consideramos que podría ser un producto de un modelo defectuoso ya que un valor elevado en el Laverage puede ser generado por un incremento en los residuales de estas observaciones atípicas del modelo.

De igual manera, se quiere revisar los resultados a través de la técnica de validación cruzada. Se estiman los mismos 10 modelos anteriores que arronjan los siguientes errores de predicción promedio (Tabla 10):

*Tabla 10. Validación Cruzada: Errores de predicción promedio*

| Modelos | MSE-Cross Validation |
|---|---|
| m1 | 0.8775754 |
| m2 | 0.877503 |
| m3 | 0.8775745 |
| m4 | 0.8774391 |
| m5 | 0.8773431 |
| m6 | 2684902 |
| m7 | 2681683 |
| m8 | 0.8775524 |
| m9 | 0.8774645 |

Fuente: elaboración propia.

En este caso, comparamos por separado los errores de predicción promedio de los modelos con la variable de ingreso en niveles como dependiente (m6 y m7 de la tabla 10) y los errores de predicción promedio de los modelos con el ingreso en logaritmos, encontrando que, el mejor modelo en niveles es el m7 y el mejor modelo en logaritmos es el m5. No obstante, se analizará el modelo m7, ya que es el modelo que se ha discutido en todo el documento y ha demostrado el mejor desempeño.

Comparando las dos técnicas empleadas, a pesar de que se infiere el mismo resultado en términos de elección del modelo, los errores de predicción son diferentes.

Se calculó el estadístico Leave-One-Out Cross Validation (LOOCV) para toda la muestra y se realizó el gráfico 5 en donde se compara con el estadístico Laverage, logrando evidenciar que el comportamiento de los dos estadísticos da resultados similares.

*Gráfico 5. Comparación LOOCV vs Laverage*