Universidad de los Andes

Maestría Economía Aplicada: Big Data and Machine Learning for Applied Economics

Grupo: Laura Natalia Capacho (202121025), Sebastián David Beltrán (202121021) y

Yurani Gonzalez (201212100)

***Problem Set 1: Predicting Income***

El objetivo es construir un modelo predictivo para el ingreso individual, a partir de la Gran Encuesta Integrada de Hogares del DANE para 2018 en la ciudad de Bogotá.

**1. Adquisición de datos**

(a) El proceso de adquisición de datos está descrito en el *script* de R adjunto al documento.

(b) No hay restricciones para acceder a la información puesto que no existe ningún archivo (robots.txt) asociado a la raíz del sitio web que indique restricciones para rastrear la información de la página. Sin embargo, al momento de realizar el *web scraping* para obtener las bases de datos, nos enfrentamos a un problema debido a que la página web es dinámica y no estática. Por lo anterior, no es posible extraer las tablas de forma directa, en la medida que la información no carga inmediatamente. Las tablas realmente provienen de una promesa *json*, asociada a un link externo que es de donde pudimos extraer los datos.

(c) Para adquirir los datos seguimos el siguiente proceso:

1. Crear una lista vacía “Links” en la que se van a guardar los links de cada una de las páginas que contienen las tablas con la información
2. Crear una lista vacía “base” en la que se van a almacenar las tablas que se extraen de cada página web
3. Crear un *loop* que para i = 1, 2, …, 10, ejecuta los siguientes pasos:
   1. Guardar en la lista “Links” cada uno de los links de las 10 páginas web, definiéndolo como un documento de .*html*
   2. Descargar y preprocesar las páginas web descargadas de cada link (*read\_html)*
   3. Convertir los nodos de tipo tabla de la página web en tablas de R (*html\_table)*
   4. Almacenar cada una de las tablas en la lista “base”
4. Renombrar la primera columna de todas las bases, que está vacía, para evitar errores
5. Convertir cada una de las tablas en la lista “base” a un formato *tibble*
6. Hacer un *append* de las 10 bases almacenadas en la lista “base”, para obtener una base de datos completa con toda la información

**2. Limpieza de datos**

Luego de hacer una exploración de las variables disponibles en la base de datos se hizo una selección de aquellas que consideramos más relevantes para explicar el ingreso individual de una persona. Destacamos que en este trabajo se hace referencia al ingreso total, constituido tanto por ingresos laborales provenientes de salarios o independientes, como por ingresos de ayudas, subsidios, bonificaciones y diversas fuentes que serán detalladas más adelante.

En este sentido, los determinantes del ingreso son tanto características de la persona o su hogar, como del tipo de trabajo que desempeñe (rama de actividad económica, condiciones de formalidad, tipo de empresa, entre otros). Así como otros factores que pueden aumentar la probabilidad de recibir otros ingresos (subsidios o auxilios, pensiones, entre otros).

Variables de interés

En primer lugar, se tomó un conjunto de características sociodemográficas de los individuos: **edad, edad al cuadrado** (para capturar el efecto decreciente de la edad sobre el ingreso), **género, máximo nivel educativo, relación con el jefe del hogar** y **estrato**. Dentro de estas se hicieron algunas modificaciones y se crearon nuevas variables:

1. Creamos una variable de **educación** medida en años que construimos a partir del máximo nivel educativo alcanzado que se reportó en la encuesta. En efecto, si este nivel es ninguno se asignaron 0 años, si es preescolar 3 años, si es primaria incompleta 7 años, si es primaria completa 8 años, si es secundaria incompleta 13 años, si es secundaria completa 14 años y si es terciaria 19 años.
2. Construimos un *proxy* de **experiencia laboral**[[1]](#footnote-1). Debido a que no se tiene información reportada de los años de experiencia laboral, utilizamos el concepto de “experiencia potencial”, que se crea a partir de la edad, los años de educación y los años de iniciación en el mercado laboral (Aristizábal & Ángel, 2017).
3. Generamos una variable *dummy* que caracteriza si la persona es **jefa del hogar** o no.

En segundo lugar, se consideraron un conjunto de variables correspondientes al trabajo del individuo: trabajador **asalariado o independiente, formalidad, oficio** y **tamaño de la empresa** en que labora. En este caso también se realizaron modificaciones y se crearon variables:

1. Se genera una variable *dummy* que indica si la persona trabaja en una **microempresa** o no, a partir del tamaño de la empresa, medido por el número de trabajadores. En concreto, se denotan como microempresas aquellas compañías que tengan personal no superior a 10 trabajadores (Ministerio de Comercio, Industria y Turismo [MINCIT], 2007).

Finalmente, se toman todas las variables referentes a **ingresos** **totales** de la persona: **ingreso laboral** (salario o independiente) y **otros ingresos**, conformados por: ingreso por intereses o dividendos, por jubilaciones o pensiones, de ayudas de hogares e instituciones, por arriendos, por horas extra, por bonificaciones, por auxilios alimentarios o de transporte, por subsidios familiares o educativos, por primas de servicios, navidad o de vacaciones, por pensiones alimentarias, viáticos, accidentes o de cualquier otra fuente. En este conjunto de variables se hicieron algunas agrupaciones:

1. Se crea una variable de **ayuda de hogares** a partir de la suma de variables que denotan dinero recibido de otros hogares o personas residentes en el país y fuera del país.
2. Se crea la variable de **primas** que agrupa los ingresos por primas de servicios, navidad y de vacaciones.

Tratamiento de valores faltantes

En el procesamiento de los datos encontramos que hay muchas observaciones con datos faltantes (*missings)*, en este sentido, dimos un tratamiento especial a las variables correspondientes de acuerdo con ciertos criterios. En principio se dio un tratamiento a las variables existentes en la base de datos, a continuación, se detalla el procesamiento:

* La variable **oficio** es una variable categórica que denota la ocupación de la persona y toma valores de 1 a 99. En este caso los *missings* se reemplazaron por 0, que corresponde a una nueva categoría de ocupación que agrupa “otras ocupaciones”.
* La variable **formal** es una variable categórica que toma el valor de 1 si se trata de un trabajador formal y 0 en otro caso. Para los *missings* en este caso se asignó el valor de la variable p6090 (¿Está afiliado, es cotizante o es beneficiario de alguna entidad de seguridad social en salud?). Lo anterior, puesto que el empleo informal se refiere a los trabajadores que, entre muchos factores, pertenecen a una empresa o desempeñan un trabajo sin contrato laboral y sin aportes a seguridad social (Departamento Administrativo Nacional de Estadística [DANE], 2009).
* La variable **iof1es**corresponde al ingreso por intereses y dividendos imputado, y para sus valores faltantes se asignó el valor de la variable iof1 que es la misma variable, pero antes de imputación.
* La variable **iof2es**corresponde al ingreso por jubilaciones y pensiones imputado, y para sus valores faltantes se asignó el valor de la variable iof2 que es la misma variable, pero antes de imputación.
* La variable **y\_vivienda\_m** representa la renta por vivienda, en este caso a los *missings* se les asignó el valor reportado en la variable p7500s1 (Valor asociado a ¿El mes pasado, recibió pagos por: a. arriendos de casas, apartamentos, fincas, lotes, vehículos, equipos etc.?)
* Las variables **y\_horasExtras\_m**, **y\_bonificaciones\_m**, **y\_auxilioAliment\_m**, **y\_auxilioTransp\_m**, **y\_subFamiliar\_m**, **y\_subEducativo\_m**, **y\_primaServicios\_m**, **y\_primaNavidad\_m**, **y\_primaVacaciones\_m**, **y\_primas\_m**, **y\_viaticos\_m**, **y\_accidentes\_m**, **y\_total\_m** son ingresos por conceptos de horas extra, bonificaciones, auxilios de alimentación y transporte, subsidios familiar y educativo, primas de servicios, navidad y vacaciones, primas totales, viáticos, accidentes e ingreso salarial y de independientes, respectivamente. Para los valores faltantes en estas variables se asignó el valor cero (0). Lo anterior, partiendo del supuesto de que, si no hay valor en estos ingresos, es porque no existió o fue igual a cero.

Posteriormente, procedimos a crear algunas variables de interés, y así mismo se hizo un tratamiento sobre sus valores faltantes.

* La variable **Micro\_empresa** que creamos para denotar si la persona trabaja en una microempresa o no. En este caso, se decidió eliminar las observaciones para las cuales esta variable tenía *missings,* esto puesto que no tenemos información suficiente para asignar a microempresa o no arbitrariamente.
* En la variable **educ** que creamos y mide el nivel de educación en años, utilizamos el máximo nivel de educación, y en ese caso solamente se encontró un valor faltante, al cual se le asignó el valor de cero (0).

Estadísticas descriptivas

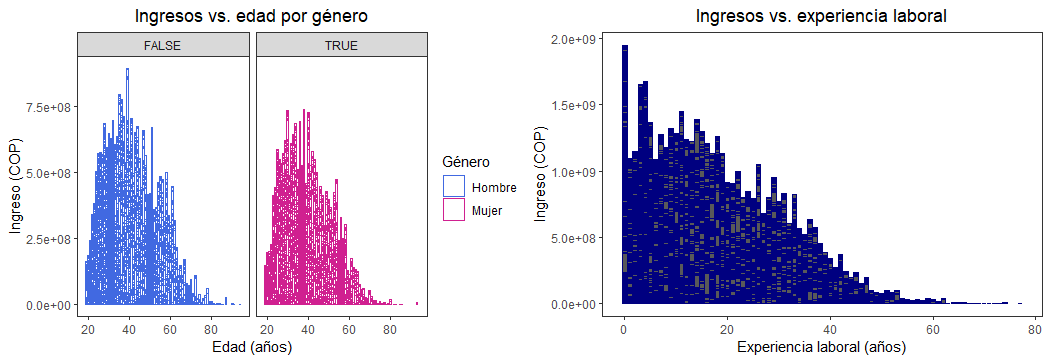
* La Tabla 1 presenta las principales estadísticas descriptivas de los datos. Contamos con una muestra de 16,397 observaciones para todas las variables. Encontramos que el ingreso promedio es de COP2,698,186 millones (mn), con una desviación estándar de COP5,144,721 mn y se observa un valor máximo atípico de más de COP160 mn. Del mismo modo, encontramos que el 47% de la muestra son mujeres y 48% jefes de su hogar, y en cuanto a caracteristicas laborales, 31% son trabajadores independientes, 59% son formales y 51% trabajan en microempresa.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **N** | **Media** | **Desv. Est.** | **Min** | **Máx** |
| Ingreso | 16,397 | 2,698,186 | 5,144,721 | 0 | 160,833,333 |
| Edad | 16,397 | 40 | 13 | 19 | 94 |
| Educación | 16,397 | 15 | 4 | 0 | 19 |
| Experiencia | 16,397 | 19 | 15 | 0 | 77 |
| Oficio | 16,397 | 50 | 28 | 1 | 99 |
| Estrato | 16,397 | 3 | 1 | 1 | 6 |
| Ingreso laboral | 16,397 | 1,451,278 | 2,359,661 | 0 | 70,000,000 |
| Otros ingresos | 16,397 | 1,127,691 | 3,674,663 | 0 | 130,000,000 |
| Mujer (=1) | 16,397 | 7715 | 47.1% | - | - |
| Cuenta propia (=1) | 16,397 | 5080 | 31.0% | - | - |
| Formal (=1) | 16,397 | 9676 | 59.0% | - | - |
| Jefe hogar (=1) | 16,397 | 7872 | 48.0% | - | - |
| Microempresa (=1) | 16,397 | 8443 | 51.5% | - | - |

**Tabla 1. Estadísticas descriptivas de las variables\***

\*Para las variables categóricas la media corresponde al número de personas en la muestra que toman el valor de 1 en cada variable. Por su parte, la desviación estándar indica la proporción de estos individuos sobre el total.

Luego de armar la base de datos y analizar las estadísticas descriptivas principales, se procedió a explorar cómo se caracterizan los individuos en la muestra utilizada. En la Gráfica 1 se observa que los hombres tienden a tener ingresos más altos que las mujeres a una edad más temprana. En ambos casos, el ingreso tiene un declive con el aumento de edad en alrededor de los 46 años. Además, se demuestra un comportamiento similar para la experiencia laboral, donde a mayor número de años el ingreso aumenta y luego comienza a decrecer después de los 20 años.

**Gráfica 1. Ingresos vs. edad y experiencia laboral**

* **…**

**Fuente:** GEIH 2018, construcción propia.

La Gráfica 2 muestra el comportamiento del ingreso por estrato, este tiene una media similar entre estratos 1-2, y a partir del estrato 3 comienza a aumentar el ingreso promedio al igual que la varianza, indicando observaciones atípicas. Asimismo, a pesar de la edad, el estrato es factor determinante, estratos más altos reciben más ingresos incluso con la misma edad.

**Gráfica 2. Ingresos por estrato socioeconómico**

* Chart, box and whisker chart

  Description automatically generated

**Fuente:** GEIH 2018, construcción propia.

Igualmente, la formalidad y tipo de trabajador demuestran ser factores importantes en la varianza de ingresos, es decir, aquellos trabajadores con un trabajo formal o que son dependientes, tienen mayores ingresos de aquellos que no lo son (ver Gráfica 3).

Chart, box and whisker chart

Description automatically generated**Gráfica 3. Ingresos por tipo de trabajo (formalidad)**

**Fuente:** GEIH 2018, construcción propia.

**3. Perfil de edad-ingreso *(Age-earnings profile)***

Como se enunció anteriormente, en este documento se hace referencia al ingreso total de los trabajadores. De esta forma, se analiza un ingreso que está constituido tanto por ingresos laborales provenientes de salarios o independientes, como por ingresos de ayudas, subsidios y otras fuentes.

En este sentido, la variable de **Ingreso** que se utilizará para las estimaciones se construyó sumando: **ingreso laboral** (salario o independiente) y **otros ingresos**, conformados por: ingreso por intereses o dividendos, por jubilaciones o pensiones, de ayudas de hogares e instituciones, por arriendos, por horas extra, por bonificaciones, por auxilios alimentarios o de transporte, por subsidios familiares o educativos, por primas de servicios, navidad o de vacaciones, por pensiones alimentarias, viáticos, accidentes o de cualquier otra fuente.

La selección de las variables se fundamenta en que nuestro interés está en aproximar el ingreso total del individuo, proveniente de cualquier fuente, tanto laboral como no laboral. Con esto, las variables seleccionadas agrupan todas las posibles fuentes de ingreso monetario de la persona, dentro de las variables disponibles en la base.

Basados en esta aproximación, se realiza la estimación del siguiente modelo de perfil edad-ingreso, y los resultados se observan en la Tabla 2:

**Tabla 2. Modelo: perfil edad-ingreso**

|  |  |
| --- | --- |
|  | **Ingreso** |
|  | **(1)** |
| Intercepto | -1,317,388.5\*\*\*  (355023.24) |
| edad | 190,025.0\*\*\*  (17584.57) |
| edad2 | -2,008.7\*\*\*  (202.56) |
| R2 | 0.008 |
| Adj. R2 | 0.008 |
| Estad. F | 67.670\*\*\* |
| N | 16,397 |

*\*\*\*p < 0.001; \*\*p < 0.01; \*p < 0.05*

**Fuente:** GEIH 2018, Cálculos propios.

Al analizar el ajuste del modelo vemos que tanto el **R2** como el **R2** ajustado son de 0.008, es decir, las variaciones en las variables explicativas explican en tan solo 0.8% las variaciones en el ingreso del individuo. Lo anterior, indica que esta estimación no permite ajustar de manera sobresaliente la variable de interés, y esto podría deberse a que existen muchos otros factores no incluidos en el modelo que pueden estar explicando el ingreso.

De otro lado, en cuanto al **estadístico F**, que indica significancia conjunta de las variables independientes, podemos afirmar que existe suficiente evidencia estadística para rechazar la hipótesis nula, con lo cual, las variables incluidas son, en conjunto, significativas para explicar el ingreso. Asimismo, en los resultados de la estimación se observa que individualmente las variables de edad y edad al cuadrado resultan significativas al 1% como determinantes del ingreso.

Como se evidencia en la Gráfica 4, la relación entre el ingreso y la edad no es lineal, con lo cual, sabemos que se tiene un impacto positivo sobre el ingreso con cada año adicional, hasta un punto, en el cual cada año adicional de edad empiezan a impactar negativamente. Procedemos a encontrar ese “pico de edad” que sugiere la ecuación estimada, utilizando el método de remuestreo *bootstrap* para calcular los errores estándar del modelo y construir así los intervalos de confianza para la edad.

En primer lugar, se encuentra que la edad\*, es decir, el punto de inflexión en la edad es 47, con lo cual, cada año adicional aumenta el ingreso, y esto ocurre hasta los 47 años, edad a partir de la cual cada año adicional de edad empieza a reducir el ingreso de la persona. Lo anterior, se puede confirmar visualmente en la Gráfica 8. Por otro lado, se encuentra un error estándar de 0.953, y se utiliza la fórmula estándar[[2]](#footnote-2) para construir un intervalo de confianza de la media con un nivel de significancia del 5%. Se encuentra para la media de la edad un , que indica que con un 95% de probabilidad podemos afirmar que el verdadero valor del pico de edad promedio de todos los individuos se encuentra entre 45 y 49 años.

Gráfico

Descripción generada automáticamente**Gráfica 4. Ingreso estimado vs. edad**

**Fuente:** GEIH 2018, construcción propia.

Con el fin de poder comparar el modelo de perfil de edad-ingreso aquí desarrollado con los modelos que se estimarán más adelante, se decide hacer una transformación de la variable de interés, ingreso, y se toma el logaritmo de este. Del mismo modo, la transformación se hace con el fin de capturar de forma más precisa el crecimiento del ingreso en términos relativos, en la medida que el logaritmo permite eliminar el efecto de las unidades de la variable sobre los coeficientes. Con esto, se busca aportar estabilidad en los regresores, y reducir las observaciones atípicas (Wooldridge, 2010). De este modo, se realiza la estimación del modelo (1), y en términos de la significancia de las variables y el ajuste del modelo se obtienen los mismos resultados.

De la estimación de ambos modelos se puede ver en la Gráfica del Anexo 1 que existe una varianza importante entre los datos estimados y observados, lo que nos indica que este modelo no permite un muy buen ajuste.

1. Para las personas con educación se utiliza la siguiente aproximación de experiencia (X): Si 18 < edad < 22, X = edad – 18; si edad > 22, X = edad – educación – 6. Y para personas sin educación terciaria se aproxima como sigue: si edad > 18, X = edad – 18. [↑](#footnote-ref-1)
2. La fórmula para IC de media en muestras grandes es: . El valor de 1.96 proviene de la distribución normal estándar, donde 1.96 es el valor crítico asociado al grado de confianza de 95%. [↑](#footnote-ref-2)