**Taller\_1\_BigData**

**Grupo 2**

**Introducción**

Este trabajo toma la Gran Encuesta de Ingresos y Hogares - GEIH de 2018, realizada por el DANE.

Con este taller se busca dar explicación a partir de la teoría económica y utilizando análisis básico y

herramientas de machine learning para poder captar el concepto de no buscar el mejor ajuste sino la capacidad del modelo de replicarse y predecir en otros conjuntos muestrales con el menor error posible.

El enfoque de este taller es aplicar la idea anterior, de permitir una mayor variabilidad que tener un modelo que se ajuste y tenga una alta significancia estadística.

Para aplicar lo anteriormente descrito, se busca analizar varios modelos que expliquen la variable

dependiente ingreso en función de las variables disponibles en la GEIH, dicho análisis se realizará

por medio de los conceptos vistos en clase.

Se decidió utilizar el lenguaje de programación R y el interfaz de ejecución R-Studio, dado que R tiene paquetes que permiten una mejor ejecución de lo que se quería probar con la muestra propuesta.

**Desarrollo**

1. Obtención de datos

Se utilizó la metodología de webscrapping con el paquete rvest que nos permitió navegar dentro de la página creada para este taller donde se encontraban anidados los datos. Dicha página no tenía ningún tipo de restricción para el uso del scrapping como pudimos comprobar al validar que no contaba con el documento robots.txt.

El proceso de lo anterior consistió en encontrar el vínculo html que contenía las tablas utilizando las herramientas de desarrollo de inspección de la página en el navegador Chrome, una vez identificado esto, se decidió crear un ciclo que permitiera hacer la descarga de cada uno de los 10 chunks, y se unificaron en un dataframe que se exportó a un archivo Excel para facilidad de acceso a los datos. La muestra unificada es de 321,770 registros y 178 columnas.

1. Limpieza de datos

Se delimito la base de datos a los individuos mayores de 18 años y menores o iguales a 80 años, ocupados que registraron su residencia en Bogotá.

Consideramos la variable de ocupados como la más efectiva para indicar los individuos que están trabajando y con ingreso, ya que esta variable considera que la persona sea de la PEA (Población Económicamente Activa) y tenga cualquier tipo de relación salarial, sea formal o informal, y que le permita tener un ingreso mensual. Se decidió acotar la muestra a 80 años, dado que hay adultos mayores sin pensión que deben seguir ocupados y la esperanza de vida en Colombia oscila entre 77 – 82 años. Adicionalmente, se consideró esta variable dado que era la que presentaba menos missing values, lo nos permite tener una mayor cantidad de observaciones.

Decidimos dejar como variables relevantes:

* Edad
* Género
* Oficio
* Máximo nivel de educación

Respecto a la variable de máximo nivel de educación, en base a ella se generó una nueva variable que contiene la “experiencia potencial” que tiene cada observación, lo anterior, dado que revisando la literatura de Mincer nos indica que la experiencia es una variable importante al momento de explicar el ingreso de un individuo.

(Describir la asignación y creación de esta variable)

1. Estimación de modelo y perfiles

Dado el enunciado, se propone inicialmente el siguiente modelo:

(poner en formulas)

Dicho modelo se estimó por OLS, en donde se encontró que el ajuste del modelo es muy bajo, a pesar de que es estadísticamente significativo como se muestra en la siguiente tabla:

Tabla

Descripción generada automáticamente

En la siguiente gráfica se muestra que los valores predichos precios a partir de la ecuación propuesta, sugieren una edad de máximo ingreso de aproximadamente 46 años (45.96 años) con un intervalo de confianza del 95% entre 44.19 y 47.68 años.

Gráfico

Descripción generada automáticamente

1. Existencias de brecha de género

Usando el siguiente modelo sugerido:

Imagen que contiene Gráfico

Descripción generada automáticamente poner bien

Se encontró que las mujeres sin clasificar por oficio tienen una brecha respecto a los ingresos de los hombres, es decir, mientras mayor edad tengan las y mas ingresos, hay una brecha que se mantiene, como se muestra a continuación:

Gráfico

Descripción generada automáticamente

Aplicando la misma metodología del punto anterior, se observa que las mujeres en promedio tienen un pico de ingresos a los aprox. 42 años (41.87 años) con un intervalo de confianza del 95% entre 40.84 y 42.96 años; Y en los hombres, se observa que las mujeres en promedio tienen un pico de ingresos a los aprox. 50 años (49.87 años) con un intervalo de confianza del 95% entre 44.74 y 54.29 años.

De lo anterior se puede concluir que a nivel general según la muestra, las mujeres llegan a su pico de ingreso más rápido que los hombres y esto genera la brecha, sin embargo, puede ser que esto se relacione con la escolaridad y el oficio por el que se perciben los ingresos.

4. Predicción y comparación de modelos

5.

Habiendo divido aleatoriamente la muestra que tenemos 70% entrenamiento y 30% prueba, a continuación, se muestra cuadro resumen con los MSE para los 9 modelos (el primer modelo corresponde a una constante).

a.i Modelo con constante

a.ii Estimación de los modelos previos

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 | Model 6 | Model 7 | Model 8 | Model 9 |
| (Intercept) | 1684514.03 \*\*\* | 1684514.03 \*\*\* | 1684514.03 \*\*\* | 1684514.03 \*\*\* | 1684514.03 \*\*\* | 1684514.03 \*\*\* | 1684514.03 \*\*\* | 1495776.10 \*\*\* | 1498063.03 \*\*\* |
|  | (26324.29) | (26287.92) | (25102.78) | (26149.63) | (24654.52) | (26298.07) | (26118.62) | (35424.95) | (35309.15) |
| age |  | 137102.46 \*\*\* |  | 1685030.88 \*\*\* | 1678425.98 \*\*\* |  |  |  | 3857436.11 \*\*\* |
|  |  | (26289.31) |  | (155980.15) | (147062.06) |  |  |  | (483072.14) |
| Escol |  |  | 771414.84 \*\*\* |  | 897957.07 \*\*\* |  |  | 1036488.06 \*\*\* | -249143.33 |
|  |  |  | (25104.11) |  | (26105.76) |  |  | (30127.54) | (163777.73) |
| agesqr |  |  |  | -1570153.06 \*\*\* | -1268549.01 \*\*\* |  |  |  |  |
|  |  |  |  | (155980.15) | (147323.11) |  |  |  |  |
| exp |  |  |  |  |  | -117246.40 \*\*\* |  |  |  |
|  |  |  |  |  |  | (26299.46) |  |  |  |
| `poly(exp, 2)`1 |  |  |  |  |  |  | -117243.30 \*\*\* | 471515.06 \*\*\* | -3966972.47 \*\*\* |
|  |  |  |  |  |  |  | (26119.31) | (30015.05) | (556642.51) |
| `poly(exp, 2)`2 |  |  |  |  |  |  | -299464.80 \*\*\* | -228954.60 \*\*\* | -165804.42 \*\*\* |
|  |  |  |  |  |  |  | (26119.31) | (24675.46) | (25834.22) |
| sex |  |  |  |  |  |  |  | 364942.18 \*\*\* | 360520.19 \*\*\* |
|  |  |  |  |  |  |  |  | (49304.32) | (49144.65) |
| N | 9463 | 9463 | 9463 | 9463 | 9463 | 9463 | 9463 | 9463 | 9463 |
| R2 | 0.00 | 0.00 | 0.09 | 0.01 | 0.12 | 0.00 | 0.02 | 0.13 | 0.13 |
| All continuous predictors are mean-centered and scaled by 1 standard deviation. \*\*\* p < 0.001; \*\* p < 0.01; \* p < 0.05. | | | | |  |  |  |  |  |

|  |  |
| --- | --- |
| **Model** | **MSE** |
| 1 | 4.99E+12 |
| 2 | 4.97E+12 |
| 3 | 4.33E+12 |
| 4 | 4.90E+12 |
| 5 | 4.19E+12 |
| 6 | 4.97E+12 |
| 7 | 4.88E+12 |
| 8 | 4.11E+12 |
| 9 | 4.08E+12 |

De la tabla anterior se concluye que el mejor modelo es el 9. En dicho modelo XXX