**BIG DATA MACHINE LEARNING FOR APPLIED ECONOMICS**

**TALLER 1 – PREDICTING INCOME**

INTEGRANTES:

Luisa Cuellar - 201613942

Isabella Riveros - 201923015

Miguel Contreras - 202116189

CONTENIDO

Introducción

Datos

Perfil edad-ingresos

Brecha salarial de género

Predicción de ganancias

1. **Introducción**

En el presente documento se indaga acerca de las características que determinan el ingreso de un trabajador en Colombia. Para ello, se realizan una serie de análisis econométricos utilizando datos de la Gran Encuesta Integrada de Hogares (GEIH) del 2018. Se efectuaron regresiones con variables que determinan las características de las personas como, por ejemplo, la edad, el sexo, el estrato socio económico y el nivel educativo, entre otros.

Algunos de los resultados destacables de este estudio son que, la edad es un factor relevante en los salarios, teniendo un signo positivo en el estimador, lo cual indica que al momento de ingresar al mercado laboral (para el caso de este estudio, 18 años) la remuneración es más baja, y tiende a incrementarse con el paso de los años. Por su parte, el coeficiente de la edad al cuadrado cuenta con un signo negativo, lo que indica que el salario de las personas de la muestra aumenta, pero a un ritmo cada vez más bajo, hasta el punto en que tiende a estabilizarse o incluso, a declinar (a excepción de algunos datos atípicos de individuos de estratos altos). En el caso de la edad y la edad al cuadrado, variables utilizadas en el primer modelo, estas son estadísticamente significativas.

Otro de los hallazgos es que, las mujeres presentan en general, menores salarios que los hombres (xx % según estima el modelo), aunque la brecha se reduce al momento de comparar individuos con el mismo oficio. La variable sexo también es estadísticamente significativa en este análisis. Tras estimar un modelo FWL los resultados muestran que xxxxxxxxxx

Además, se puede observar que tal como lo predice la teoría, la inclusión de más variables en la regresión del logaritmo del salario tiende a aumentar la varianza del modelo, y también se corrobora que el R2 no es un instrumento del todo confiable para medir el ajuste del modelo, dado que aumenta a medida que se incorporan más variables explicativas.

Por otra parte, mediante el método de Bootstrap los resultados indican que xxxx

1. **Datos (tabla de estadísticas)**

Tras el proceso de depuración de la base de datos se obtiene que, la muestra seleccionada para el estudio corresponde a hombres y mujeres mayores de 18 años que se encuentran en el mercado laboral colombiano. La muestra se restringió a 16.541 personas. En cuanto a las variables de las características personales inicialmente se plantearon las de la edad, el sexo, el nivel educativo (con 1 missing value), el estrato socioeconómico, el régimen de salud al cual se encuentra afiliado (con 1.420 missing value) y si es cotizante en pensiones.

Por otro lado, en cuanto a las características laborales se plantearon las variables de oficio, de tamaño de la firma, si es una microempresa, las horas trabajadas usualmente, las horas trabajadas en el segundo trabajo actual, si es trabajador informal y la relación laboral. Como medida de ingreso las variables que se contemplaron fueron ingreso total mensual, y la de cuándo ganó el mes anterior por la actividad principal (con 4.535 missing values). Tras un análisis conjunto se decidió utilizar las variables que se describen a continuación:

* Ingtot: es una variable continua numérica, la cual representa el ingreso total por persona en pesos colombianos. Este resulta de sumar cada una de las fuentes de ingresos tanto observadas como imputadas. Consideramos que esta es una mejor medida para determinar el ingreso dado que, la opción alterna que contemplamos inicialmente no tenía en cuenta los ingresos por actividades secundarias. Esto es importante pues, al momento de hacer un análisis con fines tributarios, es necesario considerar otros ingresos como, por ejemplo, auxilios de alimentación, auxilios de vivienda, subsidios familiares, subsidios educativos, ingresos por concepto de arriendos, por intereses, dividendos y utilidades, entro otros.
* Age: variable continua numérica, restringida a mayores de 18 años para el análisis. Tomamos la mayoría de edad como partida dado que asumimos que es desde este momento cuando los individuos ingresan a las actividades laborales formales y, por tanto, son sujetos de obligaciones fiscales. Si bien la legislación nacional permite trabajar legalmente desde los 15 años, los menores requieren la respectiva autorización expedida por el Inspector de Trabajo o, en su defecto, por el ente territorial local, y dicha información no se encuentra especificada en la GEIH.
* Sex: variable dummy la cual toma valores de 1 en caso de ser mujer y 0 en caso de ser hombre. Esta variable es de gran interés en el análisis dado que se busca estimar la brecha salarial de género en Colombia.
* Estrato1: variable discreta numérica que va de 1 a 6, número que corresponde al estrato de la persona encuestada. Para el análisis se toma como base el estrato 1, con el fin de evitar problemas de multicolinealidad perfecta en la regresión. La inclusión de esta variable nos permite ver el nivel en el cual la distribución del ingreso se asocia con el lugar de residencia de las personas. Este también puede servir como base para determinar medidas fiscales diferenciadas por estratificación.
* Maxeduclevel: variable discreta numérica que va de 1 a 7, e indica el nivel máximo que alcanzó la persona en el sistema escolar. Va desde ningún nivel educativo hasta nivel terciario. Para el análisis se toma como base ningún nivel educativo, con el fin de evitar problemas de multicolinealidad. Con su inclusión se busca entender si la educación es un factor que incide de manera significativa en el salario.
* College: variable dummy la cual toma valores de 1 si es graduado de la universidad y 0 de lo contrario. Su inclusión permite estimar qué tanto pueden cambiar los ingresos de una persona en Colombia si esta logra graduarse de la universidad. Esta variable es más específica que la anterior
* regSalud: variable discreta numérica la cual indica si la persona se encuentra afiliada a seguridad social en salud el régimen, 1 cuando es contributivo, 2 cuando es especial (fuerzas armadas, Ecopetrol, universidades públicas), y 3 cuando está en el régimen subsidiado. Se define un nivel base para evitar multicolinealidad. Su inclusión permite caracterizar mejor la población, por ejemplo, al ver cómo se distribuyen en cada régimen según la edad, el sexo o el nivel educativo, entre otros.
* cotPension: variable discreta numérica la cual indica la situación en materia de afiliación al sistema pensional de la persona. Para este caso, 1 indica que está cotizando actualmente a un fondo de pensiones, 2 que no está cotizando y 3 que ya está pensionado. Se define un nivel base para evitar multicolinealidad. Con esta variable podemos analizar entre otras cosas, la disparidad entre hombres y mujeres en materia de ingresos una vez que estos superan la edad de pensión, además de darnos la posibilidad de caracterizar mejor las personas que no cotizan al cruzar con otras variables.

Xxxxxxx cuadro variables xxxxxxxxxxxx

Del análisis de las variables se puede resaltar que: distribución, medias, medianas, varianza, variables que se eliminaron

El mayor porcentaje de personas de la muestra se encuentra en la categoría de afiliado a régimen de salud contributivo y a la vez pertenece al grupo de graduados de educación secundaria o con estudios de educación terciaria. El 40,2 % de los afiliados a régimen de salud contributivo son de educación terciaria, mientras que el 25,6 % son graduados de educación secundaria. Esto es evidencia de que a mayor nivel de estudios las personas van a tener preferencia por el régimen contributivo de salud.

La mayor cantidad de afiliados al régimen contributivo de salud tienen 26 años, mientras que la mayor cantidad de afiliados al subsidiado tienen 44 años. Hay evidencia de que los más jóvenes de la muestra tienden a estar en el régimen contributivo.

Del total mujeres afiliadas al régimen de salud, el 82 % están en el régimen contributivo, mientras que en el caso de los hombres dicho porcentaje es del 81,2 %. Tanto hombres como mujeres prefieren ampliamente el régimen contributivo de salud.

En materia de pensiones se puede observar que los individuos están en su mayoría en el grupo de cotizantes a pensión, a la vez que están en el régimen contributivo de salud. Esto es el 59,7 % de la muestra. También es de destacar que hay un porcentaje importante de personas que están en el régimen subsidiado de salud y no cotizan a pensión (15 %). Este grupo podría considerarse como población vulnerable.

La mayor cantidad de horas trabajadas en un segundo trabajo, tanto para hombres como para mujeres de la muestra, es de 8 horas. Además, quienes tienden a reportar más horas trabajadas en un segundo trabajo están en su mayoría, en el nivel de educación terciaria.

El 31, 4 % de la muestra es graduado de universidad,

La mediana del máximo nivel educativo es 6 (secundaria completa), mientras que la media también se acerca mucho a este valor, lo cual es evidencia de que su distribución tendería a ser normal.

La edad mínima y máxima son 18 y 94 años respectivamente, la mediana es 38 años y la media 39,4. Esto muestra una distribución truncada y con bastante dispersión en la cola derecha. La mayoría de la población es menor de 40 años.

Para el estrato, la mediana es 2, mientras que la media es 2,55 lo cual puede dar indicios de una distribución sesgada.

El 52,9% de la muestra corresponde a mujeres

Para el ingreso se puede ver que la mediana de la muestra es 1.051.160 mientras que la media es 1.769.432. Esto puede dar indicios de datos atípicos. Hay individuos con salarios altos (como el máximo de 85,8 millones mensuales) que jalan la media. Sin embargo, el 75 % de la muestra está por debajo de 1.5 millones, mientras que el 25 % de la muestra reporta ingresos muy altos.

En la muestra, las varianzas de las variables continuas ingreso (7.159794e+12), edad (1.818.076), horas trabajadas (2.416.055), y horas trabajadas en el segundo trabajo (7.134.566), dan indicios de que los datos están demasiado dispersos.

xxxxx

……..

Estadi descrip comparar varianzas de continuas (ingreso, edad, horas de trabajo). Línea 346

En variable dummy, por ejemplo college, solo usar la media (dice en promedio )

Hablar de la mediana del nivel de educación (que haya mucha gente en ese nivel de educación)

Age continua incluir todo, es porcentajes de la población, la mediana, analizar la varianza

Estrato categórica (tables)

En incomen, media mediana (hay gente con salarios altos jala la media, el máximo ingreso es 85 millones, el 75 % de la población está por debajo del 1.5. 25% ingresos muy altos)

La mayoría categóricas (tables del scrip para las variables que falten)

Incluir en estadísticas descriptivas las diferencias de medias, ej: entre los mayores de 50 y los mayores de 50), y diferencia de medias de sexo. Decir que hay muchos menores de 40

El gráfico de ingreso por edades (línea 335). colorear cuáles son mujeres y cuáles hombres, y analizar lo que me dé

1. Perfil edad-ingresos

Modelo 1: ing-vs-age-age2

De los resultados del modelo se puede extraer que, la edad en la cual los individuos tienen el mayor ingreso es al llegar a los 57 años (sale de la foto que tomé). //VOLVER A HABLAR DE LA SIGNIFICANCIA.

El ajuste del modelo no es el mejor, dado que tan solo el 0,1 % de la variación del ingreso es explicado por la edad

El modelo, según el r ajustado, dice que la edad solamente explica 1,7% de la variación del ingreso.

1. Brecha salarial de género

Modelo 2: loging-vs-sex.

Modelo 3: loging-vs-sex-age-age2)

Modelo 3: loging-vs-sex-age-age2-ctrlvar

Modelo 4: loging-vs-sex-age-age2-ctrlvar FWL

al incluir variables como el máximo nivel educativo alcanzado, si fue a la universidad, el estrato socioeconómico, el régimen de salud al cual se encuentra afiliado, si es cotizante a pensión, el oficio en el que se desempeña, si es trabajador formal, el tamaño de la empresa donde labora y las horas trabajadas, entre otras.

1. Predicción de ganancias

Con el fin de evaluar el poder de predicción de la especificación, se dividió la muestra en dos: de entrenamiento (70%) y de prueba (30%)

**Latex : para presentar bonito**

……………………………..

Construcción de conjunto de datos y proceso de adquisición

Descripción de los datos, variables utilizadas, tabla y por qué usamos esas

Regresión ingreso vs edad. Interpretar el coeficiente, el ajuste

Gráfica perfil ingresos por edad

Discusión de la "edad máxima" con sus respectivos intervalos de confianza. (use bootstrap para construir los intervalos de confianza).

Estimación brecha salarial condicional con otras variables (características laborales )

Ver distribuciones, si hay outliers

Es cóncava o convexa

Calcular el error estándar con el Bootstrap (resampleo, cómo varía con cada)