**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 utilizados en el primer modelo son estadísticamente significativos.

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. También se estimó un modelo FWL en el cual 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. Así ocurre 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.

Por otra parte, mediante el método de Bootstrap

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 tamaño de la firma, si es una microempresa, el oficio, 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 discreta numérica

…

La motivación es predecir para ver quién tiene que pagar impuesto, por eso es mejor usar ingresos totales, pues esta incluye otros ingresos como, por ejemplo

Age

Sex

Collage: muy puntual. Pero eduterciar incluyo

Estrato1: var se toma como base el estrato 1, especificar y justificar, qué implica

Horas en el actual segundo trabajo indica que la gran mayoría 15.980 no tienen segundo trabajo\*

regSAlud: justificación de haberla metido, porque es característica que da señal de sus ingresos, dado que las personas de mayores ingresos y trabajos formales tienden a estar en régimen contributivo.

regLos que usualmente cotizan pensión y salud es los que tienden a tributar, pues al estar en sector formal

cotPension: esta es una variable discreta numérica

Ojo. Decir cuáles son las categóricas, hablar de la distribución de las variables.

Estadísticas descriptivas: interpretar media, varianza, desviaciones estándar, distribuciones, saber bien qué tipo de variables eliminamos

Ver paper de Ignacio de cómo sale

dinero por intereses, dividendos, utilidades

patrones de la gente que no cotiza, en qué edad está

**Latex : para presentar bonito**

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

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

Para primera regresión, la edad máxima es 57 años (sale de la foto que tomé)

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

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. //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

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

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%)

……………………………..

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)