*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

**Big Data y Machine Learning para Economía Aplicada**

**Taller 1**

**2023-02**

*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

Angie Ariza Quitian– Código

**Presentado por**

Andrés Felipe Diaz Barreto - Código

Duvan Javier Farfán López – 201317299

Daniel Eduardo Sandoval - 200712968

**Introducción - Dani**

El salario de una persona depende de diferentes variables y contextos particulares. En Colombia, el salario (…)

**Descripción de las fuentes de datos**

**Gran Encuesta Integrada de Hogares**

La Gran Encuesta Integrada de Hogares, se viene desarrollando desde el año 2005 y es realizada por el Departamento Administrativo Nacional de Estadística (DANE). Esta encuesta, analiza los hogares como unidad de análisis para obtener información sobre cómo están formados estos, además de los aspectos socioeconómicos de cada uno; de esta forma la información recolectada permite describir estas características que evidencia la realidad de los hogares colombianos. La encuesta que recopila información sobre las condiciones de empleabilidad de las personas en Colombia a nivel regional, departamental y sus capitales. De esta manera, la GEIH ayuda a generalizar la situación sobre el empleo, salarios, formalidad e informalidad, así como actividades económicas adicionales que puedan realizar las personas.

Dadas las características de la base de datos, sus observaciones son de gran utilidad para realizar análisis estadístico de las variables asociadas a los ingresos de los colombianos, permitiendo la interacción de variables asociadas a las condiciones laborales y las características sociodemográficas de cada individuo. Esta información ya ha sido utilizada por el mismo DANE para hacer estudios de brecha salarial en Colombia y brindando un panorama con línea base que permita establecer políticas públicas orientadas a mejorar la calidad laboral de las personas, sus ingresos, la reducción de brechas salariales en varias categorías, la estimación de la informalidad y nociones básicas de la distribución de la pobreza a lo largo del territorio nacional.

Según la revisión bibliográfica hecha para la elaboración de este trabajo, muchas de las variables asociadas a la determinación del salario están contenidas en la GEIH, por lo cual es un buen insumo para la estimación de brecha salarial entre hombres y mujeres. De hecho, el DANE ya ha usado esta base de datos para realizar un estudio mucho más robusto sobre está brecha, teniendo en cuenta otras categorías. La suficiente cantidad de observaciones y todas las variables de la base permiten seleccionar variables adicionales que pueden servir como controles en las regresiones y ser tenidas en cuenta para predicciones del salario más acertadas y cercanas a la realidad de Colombia.

**Adquisición de datos**

El sitio web indicado en el enunciado (https://ignaciomsarmiento.github.io/GEIH2018 sample/) contiene la siguiente información:

1. Acceso a las diez partes (*data chunks*) en las que se dividió la muestra del GEIH 2018

2. Enlace al GEIH del 2018

3. Diccionario de las variables de la muestra

4. Descripción del DANE de la metodología y de las variables utilizadas en el GEIH 2018.

Durante la exploración de las páginas en las que estaban los *data chunks*, se evidenció que las tablas con las observaciones se demoraban en cargar. Al intentar raspar su información en R, mediante las funciones del paquete *rvest*, se obtuvo que no había ninguna tabla; esto sucedió porque al momento de crear el objeto *hmtl* para el respectivo *data chunk* la tabla todavía no había sido cargada, por tanto, el objeto carecía de la información que se quería obtener.

Luego de identificar que la razón por la que había una demora en el cargue de las tablas era porque estas páginas estaban trayendo la información de otros sitios web, se cambiaron las URL que estaban utilizándose para raspar las observaciones, logrando conseguir las 32,177 observaciones de la muestra mencionada.

**Descripción del proceso de limpieza de datos y selección preliminar de variables**

Una vez se realizó a la adquisición de la data y se exportó como archivo .Rda, se creó un nuevo script para realizar la limpieza de la misma. Una vez cargada, se identificó que la base de datos estaba conformada por 24.054 observaciones y 178 variables.

En la primera revisión de la estructura de lavase de datos, se evidenció que algunas de las variables poseen campos con NAN, lo cual distorsionaría el resultado de las estimaciones a realizar, sin embargo, previo a verificar el contenido de cada variable, se procedió a verificar en la literatura cuales son los criterios identificados en diferentes estudios para explicar el salario de una persona.

Dentro de la literatura asociada, se puede concluir que el trabajo realizado por Mincer en 1974 para el estudio de la determinación del salario fue el punto de quiebre para llegar a un acuerdo sobre las variables a estudiar. La ecuación minceriana se basa en la teoría del capital humano pues considera que los salarios están determinados por la educación y la experiencia de las personas, así, los individuos con mayor productividad y capacidad obtienen mejores salarios (Guataqui, García, Rodriguez, 2009). Esta ecuación es usada principalmente para el estudio del retorno en la educación, pero está condicionada a supuestos muy fuertes que la pueden hacer inviable para casos diferentes como el colombiano.

Basados en la ecuación de Mincer, se han realizado otros estudios en el mundo con variantes interesantes que eliminan el sesgo de selección, como por ejemplo Soon (1987) en Malasia, cuyo trabajo se centró en la determinación de los ingresos para asalariados e independientes. Este estudio es importante, ya que encuentra que la experiencia no es significativa para las personas independientes, mientras que para los asalariados si lo es. Otros estudios se presentan en Turquía por Tansel (2000), en el que se relaciona el género y probabilidad de pertenecer a un sector de la economía (inactivo, empleado, formal, informal o cuenta propia) (Guataqui, García, Rodriguez, 2009).

Para tener cercanía con la información colombiana, este trabajo usó el estudio generado por el DANE en 2020, el cual hace una investigación de la brecha salarial de género en Colombia. Aunque su principal foco es ver las diferencias entre los salarios ganados por hombres y mujeres, y partiendo de la Gran Encuesta integrada de Hogares - GEIH 2019 y del Registro Estadístico de Relaciones Laborales – RELAB, se muestra información relacionada a las diferencias entre el sector de las personas de la encuesta (rural o urbano), la edad, el tamaño de las empresas, la formalidad o informalidad, el tipo de ocupación, el nivel educativo, el tipo de relación laboral, entre otros. Es importante resaltar que en dicho estudio se concluye que existe una brecha salarial entre hombres y mujeres en casi todas las categorías de análisis.

Para robustecer el ejercicio de selección de variables se revisaron también informes de empresas consultoras, en las que se evidencia la diferencia entre los salarios por sector de la economía, cargo de las personas y tamaño de la empresa (Michael Page, 2022). Se puede concluir que existen diferencias sustanciales entre los salarios de estas categorías, siendo los sectores de mayor demanda en este momento (Tecnología y finanzas empresariales), los que presentan una mayor remuneración.

Teniendo en cuanta la literatura, y con el ánimo de identificar las características sociodemográficas de la muestra, se realizó la **primera selección de variables** que conformarían la base de datos para el desarrollo de estimación de parámetros y predicción de salario y brechas salariales por género. Las variables seleccionadas, su descripción y justificación de uso basada en la teoría económica y consideraciones propias se puede observar en la tabla 1.

Tabla 1 Primera selección de variables

| Variable | Nombre de la variable | Descripción de la variable | Justificación de uso |
| --- | --- | --- | --- |
| age | Edad | Edad de la persona – Solo se tienen en cuenta mayores de 18 años | La ecuación Minceriana plantea una relación entre los salarios y algunas variables explicativas como el sexo, la raza y la experiencia. Esta ultima es una función parabólica, ya que en el inicio de la vida laboral de una persona no se recibe una alta remuneración hasta que se gana la suficiente experiencia, y luego de alcanzar el máximo salario ganado este disminuye por razones de edad y retiro. En este estudio, no se cuenta con información de experiencia, por lo cual se usará la edad como variable equivalente.  Adicional a lo anterior, en DANE (2020) se comparan cifras de salarios por edad (por quinquenios y grupos etarios) y sexo, evidenciando como a mayor edad se obtiene una mejor remuneración siempre existiendo una brecha salarial a favor de los hombres. Esta brecha salarial es mayor en promedio a medida que la edad también aumenta, sin embargo, la brecha del salario por hora disminuye con el aumento de la edad.  La edad y el sexo son variables fundamentales para estimar y predecir el salario. Varios estudios han demostrado su relación en Colombia, por lo cual se incluyen en el análisis y se les presta demasiada atención para verificar sus estadísticas descriptivas y la distribución de la muestra a regresar. |
| clase | Sector de residencia | Indica si la persona vive en el sector rural o urbano | Según estudios de la CEPAL (2007) y el DANE (2020), se evidencian diferencias entre los salarios que se perciben a nivel rural y urbano, siendo los primeros más bajos. Lo anterior asociado al nivel de experticia requerida en los trabajos relacionados con el campo, que principalmente están enfocados agrícola y agropecuaria. La brecha salarial ha disminuido durante los últimos años en Colombia, principalmente por la caída de ingresos de los sectores urbanos y la mejora en el acceso a educación de la población rural, sin embargo, hay evidencia sobre la brecha que se presenta entre los dos sectores. |
| college | Educación terciaria | Indica si la persona tiene o no estudios de educación terciaria (educación universitaria y/o formación profesional) | En el modelo de Mincer (1958) se indica que el logaritmo del ingreso depende linealmente de la escolaridad. Luego, en 1974, Mincer propone un modelo similar incluyendo la experiencia, que genera mucha más confianza dados los fuertes supuestos asociados únicamente a la educación de una persona. Adicional a lo anterior, Aristizabal&Lopez (2017), en su trabajo de grado, hacen una descripción de antecedentes de la literatura en al que se evidencia la relación entre los estudios y el salario de una persona, encontrando que a mayor educación mayor es el salario para el caso colombiano, aunque el impacto entre un año más de escolaridad se ha reducido durante los últimos años. Al igual que Mincer, los autores reconocen que se debe tener en cuenta la variable de experiencia para tener mejores resultados. |
| cuentaPropia | Independiente | Indica si la persona trabaja de forma independiente o recibe salario | Según el estudio del DANE (2020), un análisis superficial de información de mercado laboral e identificación de brechas salariales en Colombia permite establecer una diferencia por tipo de relación laboral. Según la investigación, tanto hombres como mujeres asalariadas perciben un mayor salario en relación con aquellas personas que son independientes. El tipo de relación laboral si puede influir en el salario percibo debido a razones como la estabilidad de los trabajos asalariados y su asociación con la formalidad, así como una remuneración de acuerdo con mercados formales comparables. |
| dsi | Desempleado | Indica si la persona se encuentra desempleada o no | Aunque esta variable no debería ser tenida en cuenta para la estimación del salario, se establece para controlar la base y realizar control de calidad de la información, ya que la base de datos no debería contemplar a ningún individuo desempleado con ingresos que se reporten en la variable de salario a regresar. |
| estrato1 | Estrato | Estrato de energía para las 13 a.M., y sextil de icv para otras cabeceras | La variable de estrato tiene conceptualmente algunos inconvenientes: dados los incentivos que pueden tener las familias por pertenecer a un estrato bajo, puede haber personas calificadas en un nivel que no le corresponde. Sin embargo, se puede evidenciar que los perfiles demográficos de las personas clasificadas en dichos grupos son similares, y por lo tanto asociarse a características económicas de ingreso. Acosta&Ramos (2017) concluyen que todos los estratos socioeconómicos presentan diferencias en los ingresos por costumbres asociadas a la discriminación de la mujer, especialmente en los estratos más bajos. Esto implica que puedan dedicar menos horas al trabajo por cumplir con tradiciones patriarcales como el cuidado del hogar y los hijos. |
| hoursWorkUsual | Horas trabajadas por semana | Horas usuales trabajadas por la persona a la semana | El número de horas es un determinante dentro del salario, ya sea porque a mayor número mayor será la retribución o porque indica el nivel de productividad asociado a cierta actividad. En el estudio del DANE (2020) sobre brecha salarial por género, los datos sugieren una visión contraintuitiva en la que a mayor número de horas trabajadas el salario por hora es menor.  Adicional a lo anterior, muchas mujeres acomodan sus horarios para poder dedicar parte del tiempo al cuidado del hogar y de los hijos, reflejando un mejor ingreso total por tiempo trabajado y aumentando la brecha salarial. Por estos motivos se selecciona la variable para hacer parte de la base de datos. |
| informal | Informal | Define si la persona es informal: personas ocupadas en  empresas que empleen en total 5 personas o menos, excluyendo trabajadores/as independientes que se dedican a su oficio y los empleados/as del gobierno (DANE,2020) | En el mismo estudio del DANE, los datos permiten establecer una relación entre los salarios y el desarrollo de actividades en el sector formal e informal, presentando una brecha amplia entre los mismos. Dada la definición de la informalidad adoptada por el DANE, es posible que la diferencia salarial obedezca al tamaño de la empresa, que a su vez refleja la capacidad financiera que esta puede tener. |
| ingtotob | Ingreso total observado | Ingreso total observado | Se tomaron varias variables de ingreso de la recolección de datos original con el fin de verificar cual era la indicada para realizar los ejercicios de estimación y predicción. Se incluye el ingreso total para verificar la relación entre el salario y los ingresos totales y establecer controles que permitan eliminar observaciones que reciban ingresos, pero no un salario (Fuente principal del estudio) |
| maxEducLevel | Nivel educativo (incluye estado) | Indica el máximo nivel alcanzado de escolaridad y si la persona finalizó los estudios asociados al nivel | Al igual que la variable *Educación terciaria*, la ecuación de Mincer y los estudios colombianos sobre el efecto de la escolaridad en el salario hacen importante la inclusión de esta variable en la selección inicial. Esta variable presenta la ventaja de tener categorías para diferenciar los efectos, mientras que la *Educación terciaria* es dicótoma y solo se centra en el efecto de tener educación terciaria.  Esta variable es importante por su estructura, ya que las categorías permiten diferenciar el efecto de la educación concluida con grado sobre la que no fue finalizada. |
| microEmpresa | Microempresa | Indica si la persona trabaja en una empresa de 5 o menos empleados o en una con más de 5 empleados | Al igual que con la variable *informal*, y según la información recopilada y tratada por el DANE, el tamaño de la empresa puede influir en el salario de las personas debido a su capacidad financiera. Una empresa grande también es sinónimo de ingresos que pueden soportar toda su operación. En el mismo estudio se argumenta que las empresas más grandes tienen más probabilidad de sobrevivir, además de tener primas salariales más altas. |
| ocu | Ocupado | Indica si la persona está ocupada o activa laboralmente | Aunque esta variable no debería ser tenida en cuenta para la estimación del salario, se establece para controlar la base y realizar control de calidad de la información, ya que la base de datos no debería contemplar a ningún individuo que no esté ocupado pues el salario es el resultante de la ejecución de alguna actividad económica. |
| oficio | Ocupación | Indica la ocupación o profesión de la persona | Varios estudios como Michael Page (2022) y DANE (2020) muestran cifras de salario por sector y ocupación. Esta información permite identificar diferencias entre las ocupaciones, principalmente en trabajos asociados a la agricultura, actividades agropecuarias, construcción, transporte y oficios de almacenamiento. Así mismo, se evidencia que los mayores salarios están en ocupaciones relacionadas con el sector de finanzas empresariales, explotación de minas, producción de hidrocarburos y tecnología. Lo anterior está asociado al nivel de conocimiento que se requiere para ejecutar estas ocupaciones y también por la baja oferta de profesionales y la alta demanda del mercado laboral (especialmente en el sector de tecnología). |
| p6210 | Nivel educativo (no incluye estado) | Indica el máximo nivel alcanzado de escolaridad. No incluye estado de finalización | Esta variable fue seleccionada por la misma razón que fueron seleccionadas las variables *college* y *maxEducLevel*. La educación está relacionada con el ingreso salarial, sin embargo, debe controlarse por la experiencia, ya que las características de la escolaridad no son iguales para toda la población. Aunque ya se tienen dos variables asociadas a la educación, se incluye esta para verificar calidad de las otras dos variables, teniendo en cuenta que esta no cuenta con información de la finalización o no del máximo nivel educativo alcanzado.  La selección de la variable depende entonces de la completitud de la información (NAN y outliers) |
| p6620s1 | Estimado de ingreso | Estimación de ingresos recibidos por la persona | Se tomaron varias variables de ingreso de la recolección de datos original con el fin de verificar cual era la indicada para realizar los ejercicios de estimación y predicción. Se incluye la estimación de ingresos por la persona para verificar la relación entre el salario y los ingresos totales y establecer controles que permitan eliminar observaciones que reciban ingresos, pero no un salario (Fuente principal del estudio) |
| relab | Tipo de ocupación | Tipo de ocupación y categoría de empleabilidad (cuenta propia, empleado, otros) | Al igual que en la variable oficio, se sustenta la selección de esta variable porque las ocupación y sectores requieren diferentes niveles de experticia. Esta variable, además, presenta una ventaja frente a la variable oficio, ya que está condensada en menos categorías e incluye el tipo de relación laboral que se presenta, lo cual es otro factor que impacta el salario. |
| sex | Sexo | Sexo: Hombre o mujer | La ecuación de Mincer, el estudio del DANE sobre Brecha Salarial de género en Colombia del 2022 y otra literatura consignada al final de este documento evidencian con información concreta la relación entre el salario y el sexo de la persona. Esta diferencia se debe principalmente a las costumbres generacionales sobre el control de los hombres en la sociedad y el relevo de las mujeres hacia el cuidado del hogar. Aunque como sociedad han cambiado paradigmas y el rol de la mujer se consolida mucho más en el mercado laboral, la información de salarios indica que aun se presentan brechas importantes entre hombres y mujeres, siendo esta más pronunciada en sectores rurales y en países no desarrollados. Adicional a lo anterior, y más allá de las cifras, se ha demostrado que muchas mujeres se enfrentan a *techos de cristal*, en los que independientemente de su preparación y experiencia, no consiguen obtener aumentos ni promociones laborales, mientras que sus colegas hombres si, lo anterior, asociado al hecho de su sexo. |
| sizeFirm | Tamaño de la empresa | Tamaño de la empresa por número de empleados | Varios estudios como Michael Page (2022) y DANE (2020) muestran cifras de salario por sector, ocupación y tamaño de la empresa (pequeña, mediana, grande). Al igual que en la justificación de la variable microEmpresa, el tamaño de una firma si puede influir en los salarios, pues refleja la solidez de la misma y sus ingresos. Empresas con más de 100 empleados demuestran una alta actividad económica asociada al requerimiento de personal, por lo cual pueden tener trabajos formales remunerados igual o mayor a los salarios promedio del mercado. Adicional a lo anterior, las empresas pequeñas pueden no tener la misma fortaleza frente a crisis económicas, por lo que el gasto es menor. |
| y\_bonificaciones\_m | Ingreso monetario al mes | Ingreso monetario en el mes | Se tomaron varias variables de ingreso de la recolección de datos original con el fin de verificar cual era la indicada para realizar los ejercicios de estimación y predicción según la completitud de la información y las características de distribución. En este caso, se verificará como se comporta el ingreso monetario frente a los ingresos por salario. |
| y\_salarySec\_m | Salario nominal mensual (secundario) | Salario nominal mensual occ. secundario | Se tomaron varias variables de ingreso de la recolección de datos original con el fin de verificar cual era la indicada para realizar los ejercicios de estimación y predicción según la completitud de la información y las características de distribución. |
| y\_ingLab\_m\_ha | Ingreso por salario (Asalariados) | Ingresos laborales de asalariados – nominales por hora (Incluye propinas y comisiones) | Se tomaron varias variables de ingreso de la recolección de datos original con el fin de verificar cual era la indicada para realizar los ejercicios de estimación y predicción según la completitud de la información y las características de distribución. Dado que el objeto de análisis es obtener la predicción del salario por hora, se incluye esta variable que se presenta en dichas unidades (COP/hora). |
| y\_total\_m | Ingresos asalariados por mes | Ingresos asalariados + independientes total - nominal mensual | Se tomaron varias variables de ingreso de la recolección de datos original con el fin de verificar cual era la indicada para realizar los ejercicios de estimación y predicción según la completitud de la información y las características de distribución. Dado que es ingreso al mes, se verificará la información contra las variables de salario por hora. |
| y\_total\_m\_ha | Ingresos asalariados por hora | Ingresos asalariados + independientes total - nominal por hora | Se tomaron varias variables de ingreso de la recolección de datos original con el fin de verificar cual era la indicada para realizar los ejercicios de estimación y predicción según la completitud de la información y las características de distribución. Dado que el objeto de análisis es obtener la predicción del salario por hora, se incluye esta variable que se presenta en dichas unidades (COP/hora). Adicional a eso, está variable incluye los ingresos por salario de personas independientes, lo que la hace la más completa para ser seleccionada como variable regresada y a predecir. |

Una vez identificadas las variables de interés en la primera selección, se creó un data frame con las mismas, obteniendo una base de datos de 23 variables y 24.054 observaciones. Para iniciar el proceso de limpieza de la base de datos se identifican cuantos NAN tiene cada variable con el comando sapply. La cantidad de NAN por cada variable se puede ver en la Tabla 2.

Tabla Número de NAN por variable seleccionada

(Tabla exportada de R, con modificaciones de los autores)

| Variable | Número de NAN |
| --- | --- |
| age | 0 |
| clase | 0 |
| college | 0 |
| cuentaPropia | 0 |
| dsi | 0 |
| estrato1 | 0 |
| hoursWorkUsual | 7.657 |
| informal | 7.657 |
| ingtotob | 0 |
| maxEducLevel | 2 |
| microEmpresa | 7.657 |
| ocu | 0 |
| oficio | 7.657 |
| p6210 | 0 |
| p6620s1 | 13.613 |
| relab | 7.657 |
| sex | 0 |
| sizeFirm | 7.657 |
| y\_bonificaciones\_m | 23.688 |
| y\_salarySec\_m | 23.601 |
| y\_ingLab\_m\_ha | 14.269 |
| y\_total\_m | 9.422 |
| y\_total\_m\_ha | 9.422 |

Del análisis de la Tabla 2 se puede apreciar que las variables a*ge, clase, college, cuentaPropia, dsi, estrato1, ingtotob, ocu, P6210* y *sex* no tienen NAN. Sin embargo, algunas variables poseen más del 50% de sus observaciones como NAN, siendo *y\_bonificaciones\_m, y\_salarySec\_m, y\_ingLab\_m\_ha* y *p6620s1* las variables con mayor porcentaje de NAN. Dado lo anterior, se decidió eliminar estas variables de la base de datos. También se eliminan 7.657 observaciones de NAN para las variables *informal, microEmpresa, oficio, relab, sizeFirm* y *hoursWorkUsual*, ya que contaban con la misma información no disponible.

Al finalizar esta selección de observaciones, se decidió también eliminar una observación de NAN de la variable *maxEducLeve*l y no imputarla, ya que una observación no tiene un impacto significativo en las estimaciones finales. También se eliminaron los NAN de la variable de interés *y\_total\_m\_ha*, los cuales eran 1.765 después de eliminar los 7.657 de las variables categóricas mencionadas en el párrafo anterior. Con esto, se obtuvo un data frame de 14.631 observaciones y 17 variables sin NAN, como se puede ver en la Tabla 3.

Tabla Número de NAN por variable seleccionada

(Tabla exportada de R, con modificaciones de los autores)

| Variable | Número de NAN |
| --- | --- |
| age | 0 |
| clase | 0 |
| college | 0 |
| cuentaPropia | 0 |
| dsi | 0 |
| estrato1 | 0 |
| hoursWorkUsual | 0 |
| informal | 0 |
| maxEducLevel | 0 |
| microEmpresa | 0 |
| ocu | 0 |
| oficio | 0 |
| relab | 0 |
| sex | 0 |
| sizeFirm | 0 |
| y\_total\_m | 0 |
| y\_total\_m\_ha | 0 |

Una vez se realizó la limpieza por NAN, se procedió a verificar si había observaciones del salario en cero, dado que no aportarían información a la estimación de salario y presentarían errores a la hora de la estimación del logaritmo de la variable salario. Para hacerlo, se realiza una tabla con el comando summary. Los resultados se pueden ver en la tabla #4.

Tabla Resumen de la variable de salario por hora

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Min. | 1st Qu. | Median | Mean | 3rd Qu | Max |
| 0,5 | 3797,7 | 4,856,8 | 8.579,2 | 7.953,2 | 350.583,3 |

No se presentan ceros en las observaciones, sin embargo, se pueden evidenciar valores que pueden afectar la estimación de las regresiones a realizar. La tabla 4 se usa también para la determinación de la estrategia del tratamiento de outliers, pues el mínimo representa un salario al mes de 80 COP y el máximo un salario de 56.093.280 COP.

Para eliminar los outliers, se calculó la desviación estándar del salario, la cual es de 13.902,73 COP/hora. Es usual determinar los outliers como aquellas observaciones que están más allá de 3 desviaciones estándar de la media.

La limpieza de datos una vez se eliminaron los NAN correspondientes, se revisaron ceros y valores mínimos y máximos arroja un data frame de 14.286 observaciones y 17 variables. Con esto, se realizó la creación de la variable *log\_wageh*, la cual es el logaritmo natural de la variable de salario por hora seleccionada. Dadas las características de la variable sexo, donde 1 era hombre y 0 mujer, se realizó la creación de la nueva variable *female* para que la categoría base fuera mujer y así poder establecer las brechas por sexo. También se creó la variable age2, la cual expresa la edad al cuadrado, con la cual se estimará la regresión del punto #3. Finalmente, se ajusta el tipo de variable de *female* de *dbl* a *fct* por ser una dicótoma y no un entero con decimales, y se asegura que todas las variables categóricas queden como factores.

**Análisis descriptivo de los datos**

Para iniciar con el análisis descriptivo de la base de datos se genera un summary con las variables continuas de la base de datos como se puede ver en la tabla 5.

Tablan 5 Summary de variables continuas

(Tabla exportada de R)

======================================

age hoursWorkUsual y\_total\_m\_ha

--------------------------------------

1 age Min. :19.00

2 age 1st Qu.:28.00

3 age Median :37.00

4 age Mean :38.91

5 age 3rd Qu.:49.00

6 age Max. :91.00

7 hoursWorkUsual Min. : 1.00

8 hoursWorkUsual 1st Qu.: 40.00

9 hoursWorkUsual Median : 48.00

10 hoursWorkUsual Mean : 47.41

11 hoursWorkUsual 3rd Qu.: 50.00

12 hoursWorkUsual Max. :130.00

13 y\_total\_m\_ha Min. : 0.47

14 y\_total\_m\_ha 1st Qu.: 3750.00

15 y\_total\_m\_ha Median : 4783.44

16 y\_total\_m\_ha Mean : 6993.11

17 y\_total\_m\_ha 3rd Qu.: 7583.33

18 y\_total\_m\_ha Max. :41481.48

--------------------------------------

De la tabla 5 se puede observar que la media de la muestra es 38,9 años y que el primer cuartil es 28 años y el tercero 49, lo cual tiene sentido, pues en esta edad se concentra la fuerza laboral y los años productivos de las personas. Se destaca el valor máximo, pues representa una persona de 99 años que está recibiendo ingresos derivados del salario. En cuento a las horas trabajadas, los valores del del primer cuartil y el tercero son cercanos, pasando lo mismo con el salario por hora. En el caso de las horas trabajadas, tiene sentido que la mediana sea 48 horas, pues por la ley en Colombia la jornada laboral está compuesta de 48 horas. Los ingresos reflejan que, aunque se esté analizando Bogotá, hay personas que ganan poco (incluso menos que el mínimo).

**Estimación de perfil edad – salarios - Dani**

**Presentación perfiles edad-salarios y “edad-pico” - Dani**

**Estimación de la brecha salarial de género - Angie**

**Perfiles edad-salarios y “edad-pico” por género - Angie**

**Interpretación de las estimaciones de la brecha salarial de género - Angie**

**Construcción de Muestra para predicción**

La evaluación del desempeño predictivo de los modelos anteriormente especificados se realizó mediante la utilización de dos métodos de validación cruzada: el enfoque de conjunto de validación (*validation set approach*) y el LOOCV (por sus siglas en inglés).

Para el primer método, se dividió aleatoriamente la muestra en dos partes: conjunto de entrenamiento (*training set*) y conjunto de evaluación o de validación (*validation set*). La primera contiene el 70% de las observaciones y la última, el porcentaje restante. Con estos conjuntos, se entrenaron y evaluaron los siguientes diez (10) modelos:

Tabla 2. Especificaciones evaluadas en el método de validación cruzada: conjunto de validación

| Modelo | Especificación[[1]](#footnote-1)[[2]](#footnote-2) |
| --- | --- |
| Modelo 1 |  |
| Modelo 2 |  |
| Modelo 3 |  |
| Modelo 4 |  |
| Modelo 5 |  |
| Modelo 6 |  |
| Modelo 7 |  |
| Modelo 8 |  |
| Modelo 9 |  |
| Modelo 10 |  |

**Desempeño predictivo**

Como parámetro de evaluación, se utiliza el error cuadrático medio (RMSE, por sus siglas en ingles) de la prueba, que es estimado a partir de los valores predichos por el modelo ajustado en el conjunto de entrenamiento para las observaciones del conjunto de validación. En la siguiente gráfica, se muestran los resultados por modelo:

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Figura . RMSE de prueba de las especificaciones sujetas a evaluación

Como se puede ver en la gráfica anterior, los RMSE de los modelos tienen una magnitud significativa (entre 6.8% y 8.8%) respecto a la media del logaritmo del salario en el conjunto de prueba, que es 8.565. Cuando se incluyen los controles especificados (*informal, maxEducLevel, sizeFirm, relab*), el RMSE se reduce en aproximadamente 20%, variando este del modelo, y se mantiene relativamente estable con diferentes grados del polinomio de edad.

**Interpretación del desempeño predictivo**

Respecto al modelo con menor error de predicción, en el que se incluye un polinomio de grado 3 para la edad y los diferentes controles (*female, informal, maxEducLevel, sizeFirm, relab*), se tiene la siguiente distribución del error:

Gráfico, Histograma

Descripción generada automáticamente

Figura . Distribución del error de predicción en el training set – Validation Set Approach.

De acuerdo con la figura anterior, hay varios datos atípicos, especialmente en la cola de la izquierda de la distribución. Con el fin de identificar si esta distribución obedece a datos atípicos del salario por hora registrados en la base de datos o al desempeño del modelo, se presenta la siguiente figura:

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Figura . Diagrama de dispersión entre el error de predicción y el logaritmo del salario por hora observado en el training set – Validation Set Approach.

Los puntos con color rojo son aquellos que corresponden a datos atípicos en la distribución del logaritmo del salario por hora observado, mientras que los verdes son aquellos que están entre las líneas punteadas A y B, que delimitan el rango comprendido por tres desviaciones estándar de la media. De 4286 observaciones en el *training set*, 28 se identifican como datos atípicos (0.65% de la muestra); en términos de la participación en la suma de los residuos al cuadrado (RSS, por sus siglas en inglés), estos datos atípicos representan el 10.3%, por lo que se puede ver la influencia que tienen en el error de la predicción. Sin embargo, casi el 90% del RSS corresponde a datos del logaritmo del salario que no se consideran atípicos.

Por lo anterior, este modelo requiere un mejor ajuste para identificar si hay casos en los que haya un riesgo de fraude, dado que gran parte del error se concentra en observaciones que no son atípicas.

**LOOCV - Javier**

1. En algunos modelos se utilizan un conjunto de controles identificado con la variable , que contiene las siguientes variables: *informal,* *maxEducLevel, sizeFirm,* *relab* [↑](#footnote-ref-1)
2. Con el conjunto de variables de control se hicieron varias combinaciones, quitando una a la vez, y la mejor configuración es aquella con todos los controles. [↑](#footnote-ref-2)